مدیریت منابع اطلاعاتی وب

۲-۱

به کوشش
دکتر غلامعلی منتظر
فرزانه شاداندل پور


بسم الله الرحمن الرحیم

مدیریت منابع اطلاعاتی وب

نویسنده:

غلامعلی منتظر

ناشر چاپي:

سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

ناشر ديجيتالي:

مرکز تحقيقات رايانه‌اي قائميه اصفهان

فهرست

۶-بحث و نتیجه گیری افزایش روزافزون منابع اطلاعاتی وب و آشفتگی موجود در بازیابی اطلاعات جامع و مانع، لزوم سازماندهی منابع اطلاعاتی وب را پر رنگ تر می سازد از آن جایی که منابع وبی سازماندهی شده جزئی از نظام های ذخیره و بازیابی کتابخانه ای خواهند بود لازم

مدیریت منابع اطلاعاتی وب

مشخصات کتاب

سرشناسه: منتظر، غلامعلی، 1348 - ، گردآورنده

عنوان و نام پدیدآور:مدیریت منابع اطلاعاتی وب [کتاب]/ به کوشش غلامعلی منتظر و فرزانه شادان پور.

مشخصات نشر:تهران : سازمان اسناد و کتابخانه ملی ایران ، 1391.

مشخصات ظاهری:2ج.

شابک:دوره 978-964-446-344-0 : ؛ 150000 ریال: ج.1 978-964-446-343-3 : ؛ ج.2 978-964-446-345-7 : ؛ 200000ریال(ج.2، چاپ اول)

وضعیت فهرست نویسی:فاپا

یادداشت:ج.2 (چاپ اول: 1391).

مندرجات:ج.1.مبانی و تجربه های جهانی .-ج. 2. دیدگاه های فناورانه، اخلاقی و مدیریتی.

موضوع: وب-- سایت ها-- مدیریت

موضوع: منابع اطلاعاتی -- مدیریت

موضوع:آرشیوسازی وب

شناسه افزوده:شادان پور، فرزانه، 1344-، گردآورنده

شناسه افزوده:سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

رده بندی کنگره:TK5105/888/م8م4 1391

رده بندی دیویی:005/72

شماره کتابشناسی ملی:3077380

دسترسی و محل الکترونیکی:http://dl.nlai.ir/UI/2fb77759-f3eb-4f7f-ad9d-2cc0b917ed1d/Catalogue.aspx

خیراندیش دیجیتالی: انجمن مددکاری امام زمان (عج) اصفهان

ویراستار کتاب: خانم شهناز محققیان

ص: 1

جلد 1

اشاره

سازمان اسناد و کتابخانه ملی

جمهوری اسلامی ایران

مدیریت منابع اطلاعاتی وب

جلد اول

مبانی و تجربه های جهانی

به کوشش:

دکتر غلامعلی منتظر

و

فرزانه شادان پور

زمستان 1391

ص: 2

فهرست نویسی پیش از انتشار کتابخانه ملی جمهوری اسلامی ایران

سرشناسه: منتظر، غلامعلی 1348 - ، گردآورنده

عنوان و نام پدیدآور: مدیریت منابع اطلاعاتی وب / به کوشش غلامعلی منتظر و فرزانه شادان پور.

مشخصات نشر: تهران: سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران 1391

مشخصات ظاهری: 2 ج.

شابک: دوره: 0 - 344-446 - 964 - 978؛ ج. 1: 3 - 343-446-964-978؛

مندرجات: ج.1. مبانی و تجربه های جهانی - ج.2. دیدگاه های فناورانه، اخلاقی و مدیریتی

موضوع: وب--سایت ها -- مدیریت

موضوع: منابع اطلاعاتی-- مدیریت

موضوع: وب-- آرشیو سازی

موضوع: شادان پور، فرزانه، 1344- گردآورنده

شناسه افزوده: شادان ،پور فرزانه 1344 - ، گردآورنده

شناسه افزوده: سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

رده بندی کنگره: 72/ 5 00

رده بندی دیویی: 1391 TK5105/88884

شماره کتابشناسی ملی: 3077380

سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

عنوان: مدیریت منابع اطلاعاتی وب، جلد اول: مبانی و تجربه های جهانی

به کوشش: دکتر غلامعلی منتظر (دانشیار دانشگاه تربیت مدرس) و فرزانه شادان پور(مربی، سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران)

ویراستار ادبی: آرزو تجلی (کارشناس ارشد جامعه شناسی، سازمان اسناد کتابخانه ملی جمهوری اسلامی ایران)

تنظیم و تصحیح: مهشید برجیان (کار شناس ارشد کتابداری و اطلاع رسانی، سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران)

ویراستار استنادی (مقالات تألیفی): فروزان رضایی نیا کارشناس کتابداری و اطلاع رسانی، سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران)

نمونه خوانی و اصلاحات: مهشید برجیان، فاطمه رمضانپور، آمنه هزارخانی، زهرا زاهدی، محمد رضا میقانی، ملیحه حاجی زاده مقدم

طراحی جلد و صفحه آرایی: شهره خوری

ناظر فنی چاپ: نصرت الله امیرآبادی

ناشر: سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

شمارگان: 500 نسخه

بها: 15000 تومان

نشانی: تهران بزرگراه شهید حقانی (غرب به شرق)،

بعد از ایستگاه مترو بلوار کتابخانه ملی

تلفن فروشگاه 81623318- 81623315- 88941946 دورنگار: 88947496

وب سایت: www.nlai.ir

پست الکترونیک انتشارات: Publication @nlai.ir

سازمان استاد و کتابخانه ملی جمهوری اسلامی ایران

مدیریت منابع اطلاعاتی وب

جلد اول

مبانی و تجربه های جهانی

ص: 4

فهرست مطالب

سخن نخست

از ویژگی های قرون گذشته بی خبری بود و تمایز جدی عصر جدید نسبت به گذشته دسترسی آسان به اطلاعات است. بشر با از سر گذراندن سه موج و پارادایم ،کشاورزی صنعت و اطلاعات امروز در قرن بیست و یکم پا در عصر انفجار اطلاعات نهاده است این امر فی نفسه نه مطلوب است نه مذموم، بلکه به نحوه مدیریت ما نسبت به اطلاعات باز می گردد.

بشر امروزی به دلیل رشد روزافزون علم و فناوری در شرایط هشدارآمیز عدم قطعیت بسر می برد و همین مدیریت و تصمیم گیری را با چالش جدی روبرو ساخته است. اگر اطلاعات درست مدیریت شود و در تصمیم گیری ها به موقع به کار آید، و از دو ویژگی صحت و سرعت برخوردار باشد، می تواند منشأ تصمیم های تحول آفرین شود. ویژگی دیگر این عصر ظهور و حضور همه جانبهٔ اطلاعات دیجیتالی است. دورانی فرا رسیده است که در آن بناست دانش مدون و تفکر مضبوط بشر علاوه بر کاغذ، و حتی بیش از آن، بر محمل «بیت» ها مسیر تولید، نشر و اشاعه، و مصرف را بپیماید. هم اطلاعات تولید شده تحت وب و هم میزان استفادهٔ از این اطلاعات با سرعت فزاینده ای رو به رشد است. کشور ما بنابر اطلاعات وثیق از حیث تعداد کاربران و میزان حضور و فعالیت آن ها دروب جایگاه نخست را در منطقهٔ خاور میانه داراست. این روندِ رو به رشد، با نصب العین قرار دادن آرمان های بلند انقلاب اسلامی در ترویج تفکر رهایی بخش اسلامِ ولایت مدار، وظیفهٔ خطیری بر دوش نهادها و دستگاه های مسئول تولید، سیاستگذاری و نشر محتوا در محیط وب قرار می دهد و آن انجام بررسی های علمی و مستند به منظور ابتنای سیاستگذاری ها و عملکردها بر مبانی صحیح و کارآمد و متناسب با نیازهای گوناگون کاربران در این محیط است. اما وجه دیگر، صیانت از این محتوا و انتقال آن به نسل های آینده است که با توجه به ناپایداری محتوای قرار گرفته بر اینترنت و فناوری پیشرفته ای که برای چنین امر خطیری لازم است از اهمیت مضاعفی برخوردار می شود.

سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران بنابر مأموریت خویش دایر بر صیانت از میراث فکری کشور و اشاعهٔ آن، عزم راسخ داشته است که برای مدیریت منابع اطلاعاتی مهم و رو به رشد وبی نیز چاره اندیشی نماید؛ بنابر این در سال 1389 نخستین بار در کشور به تهیهٔ ساز و کار لازم برای ایجاد آرشیو ملی وب همت گماشته است.

از دیگر سو سازمان با علم به این که مدیریت در این حوزه مشارکت همهٔ صاحبان اندیشه در حوزه تولید، سازماندهی و اشاعه اطلاعات تحت وب را می طلبد، مصمم شد «نخستین کنفرانس ملی مدیریت منابع اطلاعاتی وب» را برگزار نماید تا اهل علم و فناوری در این مجمع با هم اندیشی و تضارب آراء همچون

گذشته این سازمان را یار و یاور باشند.

این اثر مجموعه ای است فراهم آمـده از تلاش پژوهشـگرانی که با وجود نو بودن مباحث مطرح شـده در محورهای موضوعی کنفرانس، به ارائهٔ ثمرهٔ پژوهش های خود همت نمودند؛ که با برگزیده ای از مقالات ترجمه ای در این عرصهٔ پژوهشی ادغام و به طبع رسیده است رجاء واثق دارم که با الطاف الهی از این پس مدیریت منابع اطلاعاتی وب، و آرشـیو وب به طور خاص، موضوع پژوهش و ابتکار عمل اهل دانش و فنـاوری در کشورمـان قرار خواهـد گرفت و در این عرصه نیز فرزنـدان این مرز و بـوم تجسم گفتـار نغز رسول اعظم صـلی الله علیه و آله خواهند بود که «علم اگر تا ثریا ،برود مردانی از فارس بدان دست خواهند یافت».

اسحق صلاحی

رئیس کنفرانس و رئیس سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

به جای مقدمه

گسترش روزافزون اطلاعات در شبکهٔ اینترنت و سادگی بارگذاری انواع داده ها بر وب جهان را با شکل جدیدی از تولید، انتشار و مصرف اطلاعات مواجه کرده است. تغییر جایگاه شهروندان جامعه از مصرف کنندهٔ صرف اطلاعات به مولّد و ناشر اطلاعات و فارغ از ساز و کارهای موسوم، سبب ساز روابطی جدید در عرصهٔ ارتباطات اجتماعی و فرهنگ شده است. از سویی حجم رو به تزاید داده ها و چرخهٔ عمر کوتاه اطلاعات موجود در وب، موجب شده که «گردآوری»، «پالایش»، «سازماندهی»، «ذخیره سازی» و «اشاعهٔ» آن ها در زمرهٔ مسائل پژوهشی در نهادهای علمی و نیز بخش های پژوهش و نوآوری شرکت ها قرار گیرد؛ ضمن اینکه حفظ و دسترسی پایدار به اطلاعات موجود در وب، که خود جزئی از میراث فکری ملّت ها محسوب می شود، به دغدغه ای جدّی برای سازمان ها متولی حفظ و اشاعهٔ میراث ،فکری به ویژه کتابخانه های ملّی، بدل شده است.

این حوزه در جهان موضوعی نسبتاً جدید است و پیشینهٔ آن به کمتر از پانزده سال می رسد، لیکن با سرعتی شتابان در حال رشد است و محققان مختلفی را از زوایای مختلف فنی، حقوقی، اقتصادی و حتی اخلاقی به سوی خود جذب کرده که گواه آن نیز طیف وسیعی از مقاله ها کتاب ها و گزارش های سازمانی است که در طی چند سال اخیر در سطح جهانی منتشر شده است. به رغم این نکات، در ایران همچنان این زمینه، حوزه ای بکر و کمتر مورد توجه محسوب می شود و در طی سالهای اخیر کمتر تحقیق بدان پرداخته، لیکن تزاید اطلاعات فارسی بر روی وب و برنامه های ملّی کشور مبنی بر توسعه کاربری های مختلف بر شبکه های اطلاعاتی (از جمله توسعهٔ دولت الکترونیکی، یادگیری الکترونیکی و کتابخانه های دیجیتالی) لزوم توجه به این موضوع را بیش از پیش نمایان می سازد. به همین دلیل سازمان اسناد و کتابخانهٔ ملی جمهوری اسلامی همزمان با برگزاری «نخستین کنفرانس ملّی مدیریت منابع اطلاعاتی وب» درصدد برآمد تا این حوزه را هرچه بیشتر به متخصصان و پژوهشگران بازشناساند. کتاب پیش رو حاصل همین نیّت متولیان این موضوع مهم است.

این کتاب مجموعه ای قریب به 30 مقالهٔ برگزیده از مهم ترین منابع علمی منتشر شده در جهان و نیز قریب به 15 مقالهٔ برگزیده از صاحب نظران ایران است که در قالب دو جلد تقدیم

حضور خوانندگان ارجمند می شود. این مقالات در چهار موضوع اصلی به شرح زیر تقسیم شده اند:

- مبانی مدیریت و آرشیو وب
- تجارب جهانی و مسائل بومی در مدیریت و آرشیو وب
- مسائل فنّاورانه
- مسائل اخلاقی و مدیریتی

بی گمان این مجموعه می توانست به افزودنی های دیگر (هم از منابع خارجی و هم از دیدگاه سایر متخصصان ایرانی) به اثری پربارتر بدل گردد لیک نخستین گامی است که در این حوزه برداشته شده و مطمئناً در مراحل بعدی با همت سایر اندیشمندان، ویراست هایی غنی تر از آن حاصل خواهد آمد. نگارنده امیدوار است این مجموعه به مثابه بذری باشد که در کشتزار ذهن پژوهشگران کاشته شده و ان شاء الله در آینده ای نه چندان دور به نهالی پر طراوت در عرصهٔ علم و عمل در جامعهٔ اسلامی مان مبدل گردد.

در پیدایی این اثر کسان بسیاری همراهی و همکاری داشته اند که مقدم بر همه اندیشمندانی است که متن هر مقاله به خامهٔ دانش افزای آنان امکان وجود یافته است. از این رو نگارنده سپاس فروتنانهٔ خود را نثار نگارندگان و مترجمان ارجمند این اثر می نماید. گردآوری، تنظیم و آماده سازی مطالب کتاب به همت خانم ها فرزانه شادان پور و مهشید برجیان بوده و ویراستاری آن را خانم آرزو تجلی بر عهده داشته اند. ویراستار استنادیِ مقالات تألیفی را سرکار خانم فروزان رضایی نیا به انجام رسانده اند و سرکار خانم دکتر میترا صمیعی زحمت چکیده نویسیِ شماری از مقالات را که فاقد چکیده بودند متقبل شدند. نمونه خوانی و اصلاحات اثر حاصل تلاش خانم ها مهشید برجیان فاطمه رمضانپور آهنگری، آمنه هزار خوانی، زهرا زاهدی، ملیحه حاجی زاده مقدم و آقای محمد رضا میقانی بوده است. ضمن اینکه زیبایی متن و صفحه آرایی آن مدیون حسن سلیقهٔ سرکار خانم شهرهٔ خوری است. زحمات لیتوگرافی، چاپ و صحافی کتاب نیز بر عهدهٔ جناب آقای امیر آبادی بوده که برخود فرض می داند از همهٔ این بزرگواران صمیمانه تشکر کند.

بی گمان پدیدآمدن این اثر به همت مسؤولان گرانمایه سازمان اسناد و کتابخانهٔ ملّی جمهوری اسلامی بوده است و نگارنده امیدوار است خداوند آنان را در مسیر خدمت به فرهنگ و دانش ایران اسلامی مورد تأیید قرار دهاد.

اللّهم وفقنا لما تحّب و ترضی

غلامعلی منتظر

تهران- بهمن ماه یک هزار و سیصد و نود و یک خورشیدی

فصل اول: مبانی مدیریت و آرشیو وب

اشاره

ص: 3

چکیده

بسیاری از جنبه های اجتماعی، اتفاقی هستند یا به طور کلی دربارهٔ اینترنت و به ویژه دربارهٔ وب بازتاب یافته اند. محافظت از وب، به این دلیل ضرورتی فرهنگی و تاریخی است. اما وب، از نظام های انتشاراتی قبل نیز برای ضرورت یک بازبینی ریشه ای از عملکردهای حفاظتی مرسوم، متفاوت می باشد. مفهوم میراث جمعی مشترک، شامل هر مصنوع محصول انسانی می شود از بناهای تاریخی معماری گرفته تا کتاب های جدید قرن بیستم ولو اینکه با فعالیت هایی حفاظتی مرتبط باشند (مثل آن ها که به طور اصولی و اختیاری سازماندهی شده اند) و قبلاً نمایان شده اند. در عصر حاضر علت بایگانی، دلیلی است که بر جزئیات تأکید میکند، به گونه ای پایدار ما را به اجتناب از عمومیت گرایی هدایت میکند و به طور خاص و منحصر به فرد که حوادث و رویدادها را مورد بررسی قرار می دهد. در حقیقت، امکاناتی که وب برای انتشار ایجاد می کند، منبعی منحصر به فرد از محتوا را ارائه می دهد که استدلال بایگانی وب، معقولانه انجام و تصدیق شده است. با توسعه شبکه های وب، به طور چشمگیری، حجم آن چه که می تواند منتشر شود و همچنین تعداد «ناشران» بالقوه یا خالقان محتوا با تقلیل هزینه های انتشار به تقریباً هیچ افزایش یافته است. دلگرم کننده است که ببینیم که بسیاری از مؤسسه های (حفظ) میراث، در بایگانی وب در حال به کارگیری هستند. بررسی اخیر توسط گروه پژوهش کتابخانه (RLG 2006) نشان داد که 60 درصد اعضای مورد بررسی اشان، بایگانی وب را قسمتی از مأموریت خود پنداشته اند .

بایگانی شبکه وب: مباحث و روش ها

اشاره

* بایگانی شبکه وب: مباحث و روش ها (1)

ژولین ماسانه (2) ترجمه: فهیمه باب الحوائجی (3)

1-مقدمه

محصولات فرهنگی گذشته همیشه نقش مهمی در اطلاع رسانی و خود ادراکی جامعه و ساختن آینده آن ایفا می نمایند. شبکه جهان گستر وب (به طور خلاصه وب) رسانه ای جامع و گذراست در جایی که با درکی عمیق، فرهنگ مدرن در جهت یافتن شکل طبیعی عبارات و اصطلاحات است. انتشار، مباحثه، ایجاد، کار و تبادل اجتماعی در یک درک عمیق: بسیاری از جنبه های اجتماعی، اتفاقی هستند یا به طور کلی دربارهٔ اینترنت و به ویژه دربارهٔ وب بازتاب یافته اند. محافظت از وب، به این دلیل ضرورتی فرهنگی و تاریخی است. اما وب، از نظام های انتشاراتی قبل نیز برای ضرورت یک بازبینی ریشه ای از عملکردهای حفاظتی مرسوم، متفاوت می باشد.

این فصل، بررسی موضوع های برخواسته از حفاظت از وب را ارائه می دهد و روش هایی که تا این تاریخ برای غلبه یافتن بر آن ها توسعه یافته است. ابتدا استدلال هایی را در برابر ضرورت و احتمالات بایگانی وب مطرح می کنیم. سپس، سعی می کنیم تفاوت های بسیار برجسته ای که وب را از دیگر

1- Web Archiving: Issues and Methods: in Masanes, Julien (ed.), Web Archiving. Berlin Heidelberg New .York: Springer.pp.1-46

2- Julien Masané s

3- دانشیار کتابداری و اطلاع رسانی و دانشیار گروه کتابداری و اطلاع رسانی دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران

محصولات فرهنگی متمایز می کند، ارائه دهیم و دلالت های شان را برای محافظت ترسیم می نماییم.

این اساس وب را در بر می گیرد و آن را به عنوان نوعی سیستم نشر فعال و یک ابررسانه ویرایش اجتماعی و با یک محصول فرهنگی سراسری مورد بررسی قرار می گیرد. احتمالات و محدودیت های حفاظت هر یک از این جنبه های وب را مطرح می کنیم. سپس، رویکردهای روش های بررسی مهم را برای اکتساب سازماندهی و ذخیره سازی محتویات وب ارائه می کنیم.

فصل 2 و 4 و 5، جزئیات بیشتری دربارهٔ روش های بررسی و ابزارهایی برای اکتساب مضامین تهیه کند و فصل های 8-6 بر روی دسترسی استخراج اطلاعات و حفاظت از محتوای وب تمرکز کرده است. دو فصل آخر این کتاب بررسی هایی موردی را ارائه می کند. بایگانی اینترنتی که بزرگ ترین بایگانی وب در جهان است (فصل 9) و DACHS یک پژوهش که در جهت گزینش بایگانی وب (فصل 10) می باشد. این فصل مقدمه ای کلی برای این کتاب مورد بررسی قرار گیرد:

در نهایت، موارد اولیه را در این حوزه فراهم می کند و طبقه بندی بایگانی های وب را برای ترسیم وضعیت جاری حفاظت وب را پیشنهاد می نماید.

- میراث جامعه و وب

2-محافظت از میراث

مفهوم میراث جمعی مشترک، شامل هر مصنوع محصول انسانی می شود، از بناهای تاریخی معماری گرفته تا کتاب های جدید قرن بیستم، ولو اینکه با فعالیت هایی حفاظتی مرتبط باشند (مثل آن ها که به طور اصولی و اختیاری سازماندهی شده اند) و قبلاً نمایان شده اند. شکل، اهداف، و کارآیی حفاظت از میراث، به طور قابل توجهی با زمان و رسانه های بررسی شده متفاوت اند و تلاش این مقاله برای خلاصه کردن این تکامل کافی نیست. اجازه دهید فقط به یاد آوریم که از آماده سازی روشنفکری مذهبی (طبق کتابخانه کاسیودوروس ویواریوم، ریچه 1998) (1) برای ایجاد مجموعه ای به عنوان نشانه های قدرت (ابتکارات موزه مدرن توسط مدیسیس (2) در فلورانس در اواخر قرن پانزدهم را ببینید) برای کنترل وضعیت نظام مند حفاظت از فرهنگ ملی (ابتکار سپرده قانونی فرانسیلر (3) را ببنید) استفاده شده است و از انگیزه های گوناگونی در جهت جمع آوری نظام مند و حفاظت از محصولات فرهنگی در تاریخ، ناشی شده است.

در عصر حاضر ، بایگانی ها به طور کلی، تمایل زیادی به فراگیر شدن دارند (آسبورن 1999) (4). همان طور که مایک فیدراستون (5) اظهار می کند:

علت بایگانی، دلیلی است که بر جزئیات تأکید می کند، به گونه ای پایدار ما را به اجتناب از عمومیت گرایی هدایت می کند و به طور خاص و منحصر به فرد که حوادث و رویدادها را مورد بررسی قرار می دهد. تمرکز بر بایگانی مؤثر کالاها به طور فزاینده ای تغییر یافته است به سوی جزئیاتی با واقعیتی

1- Vivarium library of Cassiodorus, Riché 1996

Medicis –2

Franceisler –3

osborn –4

Mike Featherstone –5

عظیم از زندگی روزمرهٔ دنیوی تمرکز یافته است (فیدرستون 2000) (1) .

در حقیقت، امکاناتی که وب برای انتشار ایجاد می کند، منبعی منحصر به فرد از محتوا را ارائه می دهد که استدلال بایگانی تمایل به ستایش آن دارد. از این رو، می توانیم فرض کنیم که قانونی بودن بایگانی وب، معقولانه انجام و تصدیق شده است. با وجود این، محافظت از وب، مورد سؤال قرار گرفته و تاکنون مورد قبول همه واقع نشده است. استدلال ها در برابر بایگانی وب می تواند در سه مقوله دسته بندی شود. آن ها که مبتنی بر محتوای یافت شده در وب هستند، آن ها که تصور می کنند که وب خود – محافظ است، و آن ها که فرض می کنند بایگانی وب امکان پذیر نمی باشد.

2-1-به اندازه کافی خوب نیست؟

نخستین مقوله، بحث هایی را درباره کیفیت محتوای وب در بر دارد که گفته می شود فرض بر عدم مطابقت معیارهای مورد نیاز برای محافظت می باشد. این موقعیت، مدت زیادی توسط برخی متخصصان نشر جهانی (ناشران و کتابداران) حفظ شده است و در جهت تهدید بزرگی که توسط این رسانهٔ جدید برای بقا اعمال می شود همراه می شود. معمولاً با تأکیداتی درباره میزان گسترده اطلاعات وب و فقدان دانش درباره روش های بایگانی وب و هزینه های آن همراه شده است.

منافع این موقعیت، از انتقال سیستم نشر پیوسته و آن هایی که بر حفاظت و بازدهی صنعت نشر پیوسته ادامه می دهند آگاهی دارد. اما آن ها از مرزهای آن چه که محافظت شده است را اجتناب می ورزند همان اندازه که شبکه وب، محدودیت آن چه را که انتشار یافته است را گسترش می دهد. معادلات اقتصادی تولیدات فیزیکی حامل دانش (ادواری ها، کتاب، و مانند آن) میراث انقلاب گوتنبرگ هستند که باید طبق این دیدگاه بر استقرار محدودیت ها برای کاری که باید حفاظت شود ادامه دهند. حتی زمانی که این معادلات عمیقاً تعدیل شده باشند. از نظر تاریخی، این حقیقت که چه چیزی می تواند منتشر شود، به واسطه هزینه های فیزیکی محدود می شود (که شامل تولید و نقل و انتقال، ذخیره سازی، و هزینه های اداره است) و فیلترسازی را ایجاب می کند، برای آن که نظام های نشر بیشتر از پنج قرن است که این کار را انجام می دهند. اما این مورد مناسبی نیست و نسبتاً میراث تعادل ثابتی از قرن پانزدهم است که نقض شده است. توسعه شبکه های وب، به طور چشمگیری، حجم آن چه که می تواند منتشر شود و همچنین تعداد «ناشران» بالقوه یا خالقان محتوا با تقلیل هزینه های انتشار به تقریباً هیچ افزایش یافته است. مباحث درباره ارزیابی کیفیت، به ناچار ذهنی، در واقع با پنهان نمودن مذاکرات واقعی دربارهٔ گسترش جو انتشار می یابد.

اگر چه رشد ادواری ها در پایان قرن نوزدهم، قابل مقایسه با وسعت آن در انقلاب جاری نمی باشد، برخی خصوصیات (مثل در همکرد نوع نشر با وضعیت جسمانی و فکری) را به اشتراک گذاشته و عکس العمل های مشابهی را به وجود آورده است. گاهی اوقات برای جامعه کتابخانه برای مثال برای پذیرش این نوع انتشار در قفسه هایشان و همچنین در قلب شان، در برگرفته شده است. همان طور که

1- Featherstone

فایت – شایب (2000) (1) برای موردی در فرانسه نشان داده است، رفتار توصیفی خاص که در سطح عنوان به آن نیاز داشت، توسط این جامعه مورد اغماض قرار گرفت و بخش کاملاً جدید مدیریت اطلاعات در کنار کتابخانه هارا، به همین دلیل، به وجود آورد (مستندسازی، نمایه سازی، نوشته های علمی). مباحث بر روی بایگانی وب، برخی شباهت هایی را بر حسب این رویدادهای فرعی، به اشتراک می گذارد که اگر در همان حالت باقی بماند، دیده خواهد شد.

فیلترسازی، اگر چه به مدت طولانی نیازی به تخصیص تولیدات فیزیکی منابع کار آمد حامل دانش را ندارد، به طور کامل از بین نرفته است. بلکه از نقشی مرکزی به یک نقشی پیرامونی تغییر می یابد و باز هم در برخی فضاها مورد نیاز است (برای مثال، اعتبار سنجی علمی) و به شکل های جدید تجربه می شود (مثل ویکی پدیا، اسلش دات، بوگس فر (2))

چنان چه اکسل برونز (3) توضیح می دهد:

عکس العمل ناگهانی این فعل و انفعالات و وسایل ارتباط جمعی مشارکتی بسیار محسوس است. اگر کسی بتواند یا حداقل توانایی بالقوه آن را، [مثل یک] یک ناشر دارد، چه اثراتی بر روی مؤسسه های انتشاراتی موجود خواهد داشت؟ اگر اطلاعات موجود در وب بتوانند با اطلاعات وسیع متنوع به آسانی ارتباط (4) داشته باشند، چه اثراتی بر روی چارچوب های (5) انتشاراتی سنتی می تواند داشته باشند؟ اگر توانایی بالقوه برای مخاطبان وب برای مشارکت در تولید و ارزیابی تعاملی محتوا وجود داشته باشد، برای ایجاد نقش های تولید کننده و مصرف کننده در وسایل ارتباط جمعی چه اتفاقی می افتد؟ (برانز ، 2005).

در زمینه حفاظت این مسئله به طور جدی مورد توجه قرار گرفته است. یک موضوع قطعی است و آن مدینه فاضله ای است که امیدوار باشیم که تعداد کمی از کتابداران فیلتر سازی ناشران را در مقیاس وب جهانی جایگزین نمایند حتی اگر آن ها سنت مدیدی در انتخاب محتوا داشته باشند، این کار را در محیطی ساختار یافته تر انجام می دهند که چندین مرتبه در ابعاد کوچک تر داشته باشد. اگر چه این کار هنوز احتمالی است و برای جامعه ای که خوب تعریف شده و اهداف کوچک مفید است (فصل دربارهٔ انتخاب روش شناختی ها و فصل 10 را دربارهٔ داچز (6) پژوهشی برگرفته درباره بایگانی وب، پژوهش انجام شده و هم چنین بروگر (7) 2005) را ببینید) به کارگیری این موارد به عنوان یک ساز و کار جهانی برای بایگانی ،وب واقع گرایانه نیست اما این حقیقت که گزینش دستی محتوا برای وسعت وب مقیاس گذاری نشده است دلیلی برای نپذیرفتن بایگانی وب نیست این فقط دلیل خوبی برای بررسی مجدد موضوع گزینش و کیفیت در این محیط است.

آیا می تواند بر اساس یک ارزیابی کیفی توزیعی در سطح بالا و جامع باشد؟ این ارزیابی، به طور

1- Fayet-schibe

2- impact bogosphere, Slashdo

3- Axel Bruns

4- link

5- format

6- DACHS

ضمنی در دو سطح ساخته می شود:

کاربران وب به وسیلهٔ دسترسی بر محتوا، خالقان (محتوا) با پیوند دادن شکل محتوای صفحه های شان (ما در اینجا به قضاوتی که توسط خود خالقان قبل از اینکه محتوای شان را به صورت برخط در وب بگذارند در نظر نمی گیریم، که اگر به عنوان یک معیار انتخاب استفاده شود، به معنای بایگانی هر چیز است) همچنین، می تواند به طور ضمنی توسط افزایش انتخاب کنندگان فعال ایجاد گردد.

اجازه دهید ابتدا دسترسی کاربران را بررسی کنیم. گسترش جوّ انتشار پیوسته تحت چیزی که ظرفیت اقتصادی برای چاپ فیزیکی مجاز دانسته است، نتایج دیگری را در بر می گیرد: افت مکانیکی در میانگین تعداد خوانندگان هر واحد از محتوای منتشر شده برخی صفحه ها حتی نه تنها توسط هیچ انسانی خوانده نمی شوند بلکه توسط هیچ روباتی هم نمایه نمی شود. بوفخاد و وایونات (2003) (1)، استفاده از سیاهه های مربوط و پرونده های سرور یک وبگاه بزرگ دانشگاهی را نشان داده اند که 5 درصد صفحه ها فقط توسط روبات ها قابل دسترسی بودند و به 25 درصد آن ها هرگز دستیابی نداشتند، بدان معنا که توزیع دسترسی به محتوای پیوسته بسیار طولانی مدت ارائه خواهد شد.

اما این تکامل در نشر مدرن، کاملاً جدید نیست. رشد و درجه بالایی از تخصصی شدن انتشارات ادواری تقریباً الگوی مشابه دسترسی را نشان می دهد. آیا این استدلالی برای عدم حفظ ادواری هاست؟ در بیش تر کشور ها، نظام های سپرده قانونی، به طور مستقل، از انتشارات آن چه را که مورد استفاده قرار می گیرند، را نگهداری می کنند. این بی تکلیفی علائق خوانندگان آینده پیش بینی می کند.

مطمئناً برای حفاظت کنندگان در ارزیابی مفید بودن محتوای پیوسته برای نمایش و تلاش برای پیش بینی برای آینده، تا زمانی که برای جوامع کاربران از پیش تعریف شده باشد، امکان پذیر است.

الگوهای دسترسی همچنین می توانند برای اداره نظام های بایگانی جهانی استفاده شوند: در مورد بایگانی وب اصلی، تاکنون، مجموعه بایگانی اینترنت توسط الکسا (2) اهدا شده است که از الگوی دستیابی برای تعیین عمق خزش برای هر سایت استفاده کرده است (فصل 9 کیمپتون) (3) و دیگران (2006) را ببینید). همچنین، می تواند به وسیله پرس و جو های فرستاده شده برای موتور جست و جو اجرا گردد (پاندی و اولتسون ، 2005) (4) . اما سؤال کلیدی برای بایگانی های وب این است: چگونه به این اطلاعات دست یابیم و کدام مرز مورد استفاده قرار می گیرد؟

ترافیک اطلاعات، معمولاً وجود ندارد و موتورهای جست و جو از ابتکار الکسا پیروی می کنند و آن را از میلیون ها نوار ابزار نصب شده در مرورگرهایی که از اطلاعات شناوری (5) کاربران به آن ها می گذرند به دست می آورد.

مؤسسه های بایگانی از کجا می توانند به آن [اطلاعات] دسترسی پیدا کنند چنان چه آن ها خودشان جست و جوی ویژه ای را ارائه ندهند؟ مرزهای آن چه باید باشند؟ آیا باید در صفحه یا در سطح سایت

Boufkhad and Viennot -1

Alexa -2

به کار برده شود؟ (الکسا در سطح سایت از آن استفاده می کند) آیا عمق خزش را فقط در سطح نخست هر سایت محدود می کند (که این بدان معناست که حداقل در سطح اول هر سایت در تمام موارد اشغال خواهد بود)؟

حتی اگر این معیار، مباحث اجرایی عملی زیادی را نمودار سازد، مزیت بردن به عنوان حمل کننده برای تمرکز بر بایگانی، ستانده های میلیون ها کاربر – نه اجتماعی کوچک – را دارد که به خوبی با مدل انتشار جامعه وب تطبیق یافته است.

معیار دیگر، سطح اهمیتی است که توسط درجه پیوند درونی یک صفحه (یا یک سایت) اندازه گیری می شود. این مسئله استدلال شده است (میسانز، 2002) (1) که این تعادل مناسب در محیطی فرامتن، با درجه ای از عمومیت است که ویژگی های انتشارات سنتی را مشخص می کند و عملاً مزیتش در قابلیت استفاده برای استخراج ماتریس های پیوندی وب است (پیج) (2) و همکاران،

1998؛ ابیتبول (3) و همکاران، 2003، 2002؛ پاستور – ساتوراس و وسپیگنانی (2004) (4). روش دیگری در انبوه کردن ارزیابی کیفی ایجاد شده است، البته نه توسط کاربران، بلکه به وسیله ایجاد کنندگان صفحه (و پیوندها) این مدلِ ارزیابیِ کیفی توزیع شده به خوبی با طبیعت انتشار بر روی اینترنت و به طور عملی با امکان اجرا مطابقت دارد.

سرانجام، ممکن است توسط مشارکت های بیشتر، وظیفه گزینش مطالب برای بایگانی کردن مقیاس را بالا برد. این کار می تواند با در بر گیری مؤسسه های بیشتری در انجام و تسهیل توسط ایجاد خدمات بایگانی انجام شود که با بخش فنی مسئله سروکار دارد. این کار به وسیله بایگانی خدمات در بایگانی اینترنتی پیشنهاد شده است که در سال 2006 شروع شده و عبارت است از توانمند سازی تنظیم و مدیریت آسان برای کتابخانه و آرشیو که نمی تواند در زیر ساختهای عملکردی مورد نیاز برای بایگانی وب سرمایه گذاری کند.

تحول ممکن دیگر با عمومیت بخشیدن به این توانایی برای هر کاربر وب در مشارکت فعال می باشد؛ البته چنان چه بخواهند در بایگانی وب شرکت کند. انگیزه اصلی کاربران در این مورد، سازماندهی حافظهٔ وب شخصی برای امکان بازگشت به مرجع بعدی برای محتوای با ثبات و، کاوش و سازماندهی آن به عنوان راهی برای مبارزه با علائم «گم شدن در فضای مجازی» است چندین بررسی در مورد کاربران نشان داده است که حفظ نشانه های محتوای مشاهده شده برای بسیاری از کاربران ضروری است (تیوان 2004) (5) البته آن ها از روش های مناسبی نیز استفاده می کنند (6) (جونز 2003، 2001). در بایگانی وب شخصی، دنبال کردن پیشینه کاربر بر روی وب می تواند برای یک سازمان دهی شخصی و محوریت زمانی

1- Masanes

2- Page

3- Abiteboul

4- Pastor-Satorras and Vespignani

5- Teevan

6- Jones

در حـافظه وب میسـر شود (رکیموتو 1999) (1) دامیس و همکـاران (2003) (2) ، رینگر و همکـاران، (2003) (3).خدمات پیوسته متعددی (Furl, My Yahoo) قبلاً در بایگانی وب شخصـی در سـطح صـفحه ارائه شـده که با قابلیت علامت زدن ترکیب شـده است. خدمات بایگانی های هانزو (4) ، حوزه گسترش یافته (مضمون، تمام سایت) و همچنین در هم و بر همی قابلیت های بایگانی، با ابزارها و خـدمات دیگر (مثـل بلاگ ها، مرورگرها، و ماننـد آن) از طریق باز کردن ای پی آی (5) را اجازه می دهـد که آن با یک بایگانی سـرویس گیرنـده با قابلیت های P2P توسـعه بیشتری خواهـد یـافت؛ و به طور چشـمگیری امکانات برای کاربران در ثبت تجربیات شان از وب را به عنوان بخشی از زندگی رقومی توسعه خواهد داد (فریمن و گلنتر 1996) (6) ، گمل و همکاران،2002). در مورد استفاده بالقوه از مخازن کاربران در یک بایگانی وب نظیر به نظیر نیز می توانید مانتراتزیس آرگون (2004) (7) را ببینید.

دیده شده است اگر این گسترش و دموکراسـی شدن نقش بایگانی بتواند مانند تفسـیر و سازماندهی اطلاعات توسـعه یابد موجب پیشـرفت علامت گذاری (8) گولدر و هامبرمن 2005 (9) و نظـام های بلاگ کردن می شود (هالاویس 2004) (10) ، ؛ برونز 2005 (11). که در این صورت کمـک ارزشـمندی خواهـد بود و برای محافظت سازمان هایی که می تواننـد نظارت طولانی مـدت بر این محتواها داشـته باشند، درونداد [خوبی] خواهد بود.

همان طور که دیـدیم اسـتدلال ها در برابر بایگانی وب بر اساس کیفیتی است که درباره فرضـیاتی بر پا شده است مانند (1) کیفیت محتوا تحت فضای سـنتی محتوای ویرایش شده به طور مرسوم، کافی و مناسب نیست؛ و 2) فقـط گزینش دستی و تک به تک که توسط حفاظت کنندگان ایجاد شده، می تواند جایگزین فقدان فیلترسازی ناشران شود (روشی که نمی تواند درست به اندازه مقیاس وب باشد، در حالی که همه با (فیلیپز 2005) (12) ، موافق اند. این دو استدلال، فقدان درک اساس توزیعی وب را نشان می دهند و اینکه چگونه می تواند وسیله نفوذ سازماندهی حافظه اش در مقیاس بزرگ باشد.

2-2-1-خود بقایی رسانه؟

دومین مقولـهٔ اسـتدلال ها اظهار می دارد که وب رسانه ای خود بقاست. در این دیدگاه، منابعی که برای حفاظت شدن مناسب هسـتند، بر روی سرورها نگهداری می شوند. بقیه به اراده به وجود آورندهٔ اصلی، ناپدید خواهد شد. از آن جا که نوع اول استدلال دربارهٔ کیفیت تقریباً در مجموعه برنامه های جهانی

1- Rekimoto

2- Dumais et al

3- Ringel et al

4- Hanzo

5- API

6- Freeman and Gelernter

7- Orgun,Mantratzis

8- Tagging

Golder and Huberman –9

Halavais –10

Bruns –11

Phillips –12

یافت شـده است، این ها بیشترین طرفـداران را در علوم کامپیوتر جهانی به دست آورده انـد. اگر چه به شدت در روزهای اول پشتیبانی شده است، بایـد بگوییم که همانطور که زمان می گـذرد و محتوا از وب محو می شود این مسـئله کمتر مورد [چالش] است. اسـناد مطالعاتی زیـادی بر طبیعت بی دوام منـابع وب، این ادعـا را که که وب رسـانه ای خود بقـاست دچار شـکست می کنـد. برای مثال، برای مروری بر ادبیات موضوع (کهلر 2004) (1) و (اسپاینلیس 2003) (2) را ببینیـد. این مطالعات بر دسترس پـذیری منابع با همان URL تأکید می کنند و نه تغییرات بالقوه ای که می تواند متحمل شود. مطالعات نشان می دهند که میانگین نیم عمر هر صـفحه وب (مدت زمانی که نیمی از صفحه ها ناپدید خواهد شد)، فقط دو سال است. این بررسی ها بر روی دسترس پذیری منابع در URL مشابه متمرکز هستند و تغییرات بالقوه ای وجود ندارد که آن ها متحملش شوند. همچنین برخی، مضـمون ها را مورد تحقیق و بررسی قرار داده و میزان تغییر را ارزیابی می کننـد. چو و گارسـیا – مولینا (2000) (3)، نیـم عمره 5 روزه ای را برای میـانگین صـفحه هـای وب کشف کردنـد. فترلی و همکـارانش (2003) (4) نشان دادند چگونه این میزان تغییر با اندازه و موقعیت محتوا مرتبط می باشد.

دلایـل بسـیار زیادی برای تمایل منابع به ناپدیـد شـدن از وب وجود دارد. نخست، محـدودیت زمانی مجاز شـدن دامنه نام (معمولاً 1-3 سال) است که به وسیله طراحی هر فضای وب در یک انتقال و شرایط غیر ایمن واقع می شود.

دوم، توان الکتریکی پایـدار، پهنای بانـد و سـرورهایی که به پشتیبانی انتشارات نیاز دارند – همان طور که در مقابل ماهیت خارجی انتشار صورت می گیرد. اما حتی وقتی که نامیدن فضا و منابع نشـر، مصون هسـتند، سازماندهی و طراحی اطلاعات می توانند نقش مهمی را در حالت ارتجاعی منابع روی سرورها ایفا کند (برنرز لی 1998) (5). همان طور که برنرز مخترع وب ادعا کرده است:

«اصلاً استدلالی در نظریه برای افراد وجود ندارد تا URL ها را تغییر دهند (یا اسناد نگهداری شده را متوقف کنند) اما میلیون ها استدلال در عمل وجود دارد (برنرز، 1998).

تغییر افراد، سازماندهی داخلی، طرح ها، فناوری های سـرور وب، عملیات نام دادن، و مانند آن می تواند ناشـی از بازسازی و گاهی فقدان اطلاعات باشد.

از این دیدگاه، سـبک رشد نظام مدیریت محتوا (6) در انتشـار، برداشت هـای گمراه کننـده در برقراری نظم هنگام بحران – چنان چه نظام مدیریت محتوا آورد چون معمولاً یک سبک ساختاری از اطلاعات یکپارچه و اغلب قابلیت های بایگانی کردن را دارا می باشد. مسئله این است که آن ها به لایه های دیگر وابسـتگی به نرم افزار اضافه می شونـد (نرم افزار نظام مدیریت محتوا) چون اسـتاندارد سازی در این حوزه وجود ندارد. معماری های اطلاعات بر اساس نظام مدیریت محتوا ثابت شده است که خنک است تا زمانی که نظام

1- Koehler

2- Spinellis

3- Cho and Garcia-Molina

4- Fetterly etal

5- Berners-Lee

6- CMS

مدیریت محتوا تغییر نکند، یعنی که خیلی طولانی نباشد.

اما آیا طراحی اطلاعات دستی است یا توسط سیستم انجام می شود. وب رسانه ای خود بقا نیست و نخواهد شد. مهم ترین دلیل آن مغایرت فعالیت های انتشار و حفاظت می باشد. انتشار، به معنای ایجاد تازگی است حتی زمانی که در مطالب کهنه هزینه شده باشد و (برای مثال در یک فضای نام گذاری مشابه یا کتاب های جدید و قدیمی باید در مخزن ناشران مشابه، با هم قرار گیرند).

تجربه ثابت می کند که انگیزه حفاظت، در میان تولید کنندگان محتوا کافی نیست و آن ها را برای حفاظت وابسته می سازد. در واقع، مرحله نخست حفاظت، اجبار در انجام آن توسط انواع مختلف سازماندهی، اجرای توسط اهداف متفاوت، انگیزه ها و حتی روش متفاوت است. وب، به عنوان زیر ساخت اطلاعاتی نمی تواند به طور اساسی مشکلات سازمانی را حل کند. از این رو، بایگانی کردن وب به عنوان فعالیتی مستقل از انتشار مورد نیاز است.

2-1-3-یک وظیفه غیر ممکن

سومین مقوله استدلال ها در مقابل بایگانی وب، از سوی افرادی مطرح می شود که نیاز به بایگانی وب را تصدیق می کنند، اما دربارهٔ امکان انجام آن شبهه دارند.

تردیدها، یا در مورد اندازه وب است یا در موارد دیگر (تأکید بر خصوصی سازی، خاصیت، روشن فکر گرایی و موانع حق مؤلف) که بایگانی وب را به چالش می کشاند.

نخستین جنبه، منتسب به بیکرانی وب است که باید در رابطه با هزینه های ذخیره و ظرفیت ابزارهای خودکار برای گردآوری حجم زیاد اطلاعات مورد بررسی قرار گیرد. خطوط DSL فعلی و ظرفیت پردازش کامپیوترهای شخصی، خزش روزانه میلیون ها صفحه را امکان پذیر می کند. مقیاس میانگین بایگانی وب، در تناسب با مقیاس خود وب می باشد. حتی اگر تخمین دقیق آن مشکل باشد (داهن (1) ، 2000؛ اگهه 2000 (2) ؛ دوبرا و فاینبرگ (3) ، 2004) از منابع مختلف در می یابیم (4) که اندازه وب سطحی، به طور متداول در دامنه ده ها بیلیون صفحه است و این اطلاعات به شکل سایر نظام های اطلاعاتی پیچیده وب که – نمی تواند خزش کند (وب پنهان) – قابل دستیابی می باشد، البته به اندازه یا دو مرتبه بزرگ تر است.

بایگانی وب سطحی ثابت شده است طی یک دهه کامل توسط بایگانی اینترنتی، سازمانی کوچک با سرمایه گذاری خصوصی کوچک شدنی است (کاهل (5) ،1997،2002).

دلیل این امر این است که این میزان مشابهی از محتوا، به وجود آورندگان، ارزش قابل توجهی را برای ایجاد، حفظ و نگهداری و دسترسی بالا پرداخت می کنند. ذخیره سازی فقط قسمت کمی از

Dahn -1

Egghe -2

3- Dobra and Fienberg

4- منابع به انـدازه نمایه موتورهای کاووش مسـتند شـده اند (یاهو ادعا می کند که 20 بیلیون صـفحه را نمایه می کنـد، گوگل می گوید که بیشتر نمایه می کنـد (بتل،2005) در یـک نگـاه کلی انـدازه بایگانی اینترنت 10 بیلیون صـفحه است)، مطالعـات اخیر بر اسـاس روش شناختی های نمونه گیری است (گلی و سیگنورینی، 2005).

5- Kahle

هزینه هـای انتشـار وب را امروز در بر می گیرد. بر عکس، بایگـانی اینترنتی فقـط برای ذخیره سازی، با اسـتفاده از فشـرده سازی برای مثال خزش گر توسط الکسا اهدا شده است) و دسترسـی ها هزینه را پرداخت می کند و مورد دوم پرداخت برای هر واحد محتواست، که بسیار کوچک تر از چیزی است که سرور اصلی می پردازد. این نتایج در میزبان کردن یک کپی برداری کاملاً گسترده از وب در مؤسسه ای واحد (کوچک) به طور محسوس ممکن است.

جنبهٔ دوم، نگرانی هـای خصوصـی سـازی، مالکیت معنوی و موانع حق مؤلف است فقط توجه داشـته باشـید که وب یک برنامه کاربردی انتشـار غیر تجـاری در اینترنت نیست. ارتباطـات پنهانی برای رخ دادن در وب تصور نشـده انـد، اما درباره برنامه های کاربردی ارتباطات (ماننـد ایمیل و انتقال پیام) زمانی که این کار انجام می شود (لیوگ و فیشـر (1)، 2003) همیشه احتمالاـت برای حفاظت از آن ها (که به طور وسـیع می شود) به وسـیله ورود به سیسـتم و اسم رمز وجود دارد. از این رو، فضاهای حفاظت شـده به عنوان بخشـی از وب عمومی مورد بررسـی قرار نگرفته اند و بنابراین نباید در بایگانی های عمومی حفاظت شوند. این طرح طبیعی از جوّ خصوصی / عمومی در اینترنت به وسـیله روشـی که خزش گرها اجرا می کنند، تقویت می شود (به وسیله دنبال کردن پیوندها) به این معناست که صفحه ها و سایت ها به داشتن درجه معینی از پیوند درونی برای کشف شدن و تصرف شدن نیاز دارند. بقیه، اجزای غیر متصل وب هستند (برادر و همکارانش (2) ، 2000) که به طور طبیعی از خزش گرها حذف می شوند. سایتی می تواند از این اسـتفاده کرده و مرزهای بیشتری را برای شامل شدن در مجموعه (بیش از یک لینک درونی) برای محدود کردن تصرف بخش های مرئی تر تنظیم نماید.

در رابطه با وضعیت قانونی بایگانی وب، به وضوح، موقعیت های گوناگونی در هر کشور وجود دارد و این یک فضای در حال نمو می باشد. کشف این جنبه هـا، فراتر از دامنه این کتـاب است که در کتاب چارلزورث (2003) (3) به آن ها اشاره شده است. توجه داشـته باشید که محتوای منتشر شده در وب غیرتجاری است چه توسط تبلیغات بر روی سایت ها یا بوسیله اشتراک پرداخت شود.

برای تمام موارد، بایگانی های وب، حتی با دسترسـی پیوسـته باید شـرایط غیررقابتی را با وبگاه های اصلی پیدا کنند و این کار می تواند در خصوص محـدودیت هـای دسترسـی به محتوا انجـام شود (برای مثـال همان طور که توسط تولیدکننـده در متن فایل روبات ها گفته شـده است). داشـتن یـک دوره ممنـوعیت، قـابلیت هـای کمـتری را نشـان می دهـد (جست و جـوی سـایت و تعاملاـت پیچیـده) و همچنین عملکردهای سـطح پایین (سـرعت دسترسـی به محتوا). بنابراین، اسـتفاده از بایگانی وب برای دسترسی به محتوا، زمانی انجام می شود که دسترسـی اصلی امکان پذیر نباشد و درآمد بازدهی داشته باشد. در این صورت، برای ناشر اصلی استفاده از بایگانی وب تهدیدی محسوب نمی شود (این موضوع را در لایمن 2002 (4) ببینید). در مقابل، بایگانی وب می تواند به طور قابل توجهی برای به وجود آورندگان سایت، حفظ بار محتوای منسوخ

1- Lueg and Fisher

2- Broder et al.

3- Charlesworth

4- Lyman

(قـدیمی) را کم کند امکان تمرکز بر روی محتوای فعلی را می دهد. حتی در این موقعیت، نویسـندگان و ناشـران ممکن است درخواسـت کنند که مطالب شان از بایگانی های قابل دسترس عموم برداشته شود. درخواست همچنین می تواند توسط شخص سوم به دلایل مختلف انجام شود. چگونه بایگانی های عمومی وب به این درخواست ها پاسخ خواهند داد؟

توصیه هایی در این زمینه در ایالات متحده پیشنهاد شده است. جدول 1-1 را ببینید، (آبویس 2002) (1).

عکس

(قدیمی) را کم کند امکان تمرکز بر روی محتوای فعلی را می‌دهد. حتی دراین موقعیت، نویسـندگان و ناشران ممکن اسـت درخواست کنند که مطالبشان از بایگانی‌های قابل دسترس عموم برداشته شود. درخواسـت، همچنین می‌تواند توسط شخص سـوم به دلایل مختلف انجام شود. چگونه بایگانی‌های عمومی وب به این درخواست‌ها پاسخ خواهند داد؟

توصیه‌هایی در این زمینه در ایالات متحده پیشنهاد شده است. جدول ۱-۱ را ببینید، (آبویس[1]،۲۰۰۲).

جدول ۱. ۱

توصیه	درخواست
۱- آرشیویست‌ها باید یک سایت به‌روش سلف سرویس ایجاد کنند و مالکان سایت می‌توانند مطالبشان را با استفاده از آن حذف کنند که براساس استفاده از معیار پروتکل روبوت متن می باشد (با این فایل متنی می‌توان، میزان دسترسی موتور جست‌وجوگر به محتوای یک سایت را کنترل کرد). ۲- درخواست‌کنندگان ممکن است بخواهند با دلیل مدرک مالکیت‌شان را با تغییر با افزودن یک پروتکل روبوت متن ، درسایتشان اثبات کنند. ۳- این کار به بایگان‌ها اجازه می‌دهد تا مطمئن شوند که مطالب گذشته بیش از این جمع‌آوری نخواهند شد یا قابل دسترسی نیستند. ٤- این درخواست، عمومیت نخواهد داشت از این‌رو آرشیویست‌ها باید کپی‌های تمام درخواست‌های حذف شده را نگه دارند.	درخواست مدیر وب توسط یک وبگاه خصوصی (غیردولتی) به دلایل خصوصی‌سازی، افترا یا خجلت
۱- آرشیویست‌ها ها باید در تلاش برای بررسی اعتبار شکایت‌ها با کنترل آنها باشند که آیا صفحه‌ها اصلی ثبت شده‌اند و اگر مناسب است بر روی درخواست نسبت به سایت اصلی از نظر قانونی اعمال حکم می شود. ۲- اگر شکایت معتبر باشد، آرشیویست ها باید موافقت نمایند. ۳- آرشیویست‌ها برای ایجاد درخواست‌های عمومی از بیانیه حق مؤلف هزاره رقومی، از طریق اثرات ناامیدکننده[1] واخطار به جست‌وجوکنندگان زمانی که صفحه‌ها درخواست شده حذف شده‌اند، تلاش خواهند کرد. ٤- بایگان‌ها به مدیران سایت‌های مذکوراز طریق ایمیل، اخطار خواهند داد.	درخواست‌های پاک شده شخص سوم براساس بیانیه حق مؤلف هزاره رقومی[2] سال ۱۹۹۸ می باشد.
۱-آرشیویست‌ها برای بررسی اعتبار شکایت با کنترل این که آیا صفحه‌ها اصلی ثبت شده‌اند، تلاش خواهند کرد، تا اگر مناسب است بر روی درخواست در رابطه با سایت اصلی از نظر قانونی اعمال حکم کنند. ۲- اگر صفحه‌ها اصلی پاک شده‌اند وآرشیویست‌ها تعیین کند که پاک کردن آنها از سرورهای عمومی، مناسب است، آرشیویست‌ها صفحه‌ها را از سرورهای عمومی‌شان بر می‌دارند. ۳-آرشیویست‌ها برای ایجاد این درخواست‌های عمومی به‌وسیله اثرات ناامیدکننده و اخطار به جست‌وجوکنندگان وقتی صفحه‌های درخواست شده برداشته شده باشد، تلاش می‌کنند. ٤- آرشیویست‌ها به مدیران سایت‌های مذکوراز طریق ایمیل، اخطار خواهند داد.	درخواست‌های پاک شده شخص سوم براساس شکایت‌های غیر از بیانیه حق مؤلف هزاره رقومی شخصی و (شامل علائم تجاری و رازهای تجاری در تولید).

1. Ubois

جدول 1. 1

ص: 15

Ubois -1

عکس

همان‌طور که در مجموعه قوانین کتابخانه بیل[۳] طبق ذکر شده کتابخانه‌ها باید مطالب و اطلاعات را تهیه کنند و تمام نقطه نظرات درباره مباحث جاری و تاریخی ارائه دهند. مطالب نباید ممنوعیت (انتشار) داشته باشد یا به علت عدم تصویب یا عدم رضایت طرفداران پاک شود. از این رو، آرشیویست‌ها نباید به‌طور کلی به این درخواست‌ها عکس عمل نشان دهند.	درخواست‌های پاک شده شخص سوم براساس اعتراض به محتوای بحث انگیز باشد (مثل مباحث سیاسی، مذهبی و عقاید دیگر)
گاهی اوقات، داده‌های محرمانهٔ افشاء شده به‌وسیله یک طرف به طرف دیگر ممکن است در نهایت توسط یک شخص سوم عمومیت پیدا کند. برای مثال، اطلاعات پزشکی که به‌طور محرمانه تهیه شده، گاهی اوقات عمومیت پیدا می‌کند وقتی عملکردهای شرکت‌های بیمه یا عملکردهای پزشکی متوقف می‌شود. این درخواست‌ها به‌طور کلی به معنای درخواست‌هایی تلقی می‌شوند که توسط نویسندگان یا ناشران داده‌های اصلی ایجاد می‌شود.	درخواست‌های پاک شده شخص سوم براساس اعتراض به افشاء داده‌های شخصی که به‌طور محرمانه تهیه شده است.
آرشیویست‌ها بهترین تاثیر مرتبط با قبول قابلیت کاربرد قرار صادره از دادگاه، به‌کار می‌برند. بعلاوه، همان‌طور که طبق قوانین کتابخانه بیل در ذکر شد کتابخانه ها باید سانسور عقاید در اجرای تعهداتشان برای تهیه اطلاعات و آگاهی حقیقی را به چالش در آورند.	درخواست توسط دولت
درخواست‌های دیگر و شکایت‌ها، مباحث حقوقی اصولی را در بر می‌گیرد. اینها براساس مورد به مورد توسط بایگانی و مشاوران آنها انجام می‌شوند. کنترل و ایجاد مجدد ویگاه‌ها براساس تغییر مالکیت می‌باشد. این توصیه‌ها می‌توانند با محیط‌های قانونی دیگر وفق داده شوند تا جایی که استفاده مجدد از ساز و کارهای عملکردی مهم ایجاد گردند (ارتباطات از مالک سایت از طریق استفاده گسترده از استاندارد متن رویوتس و همترازی روی آن چه که بر سایت اصلی در ادعای سوم انجام شده است).	درخواست‌های دیگر و شکایت‌ها، مباحث حقوقی اصولی را در بر می‌گیرد.

نیاز برای درک بهتر همزیستی بین به‌وجود آورندگان سایت و بایگانی‌های وب وجود دارد تا آنجا که بتوانند با عنایت به حقوق ایجاد کننده اطمینان پیدا کنند که از حافظه می‌تواند محافظت شود. اما این نیز بخشی از فرآیند تکامل رسانه وب است.

در مجموع، استدلال‌ها در برابر ضرورت و همچنین امکانات بایگانی وب است. جای تعجب نیست که، از نظر ما، در مغایرت با نقش مرکزی وب در خلق فرهنگ و انتشار آن و همچنین براساس طبیعت مطلقش، ایجاد شده است. فصل ۲ بینش بیشتری درباره چگونگی اهمیت بایگانی‌های وب برای پژوهش در بسیاری از حوزه‌ها را فراهم می‌کند.

در اینجا، سعی می‌کنیم نشان دهیم که در صورتی که چالش‌های جدی و مهمی را برای عملکردهای سنتی مطرح کردیم، بایگانی وب امکان‌پذیر می‌شود و یکی از موارد اصلی در برنامه حفاظت از میراث فرهنگی امروز است.

۳ـ ویژگی‌های وب برای حفاظت

وب خصوصیات مهمی دارد که هر تلاش حفاظتی باید درباره آن انجام شود. ما آنها را در این بخش در

نیاز برای درک بهتر همزیستی بین به وجود آورندگان سایت و بایگانی های وب وجود دارد تا آن جا که بتوانند با عنایت به حقوق ایجاد کننده اطمینان پیدا کنند که از حافظه می تواند محافظت شود. اما این نیز بخشی از فرآیند تکامل رسانه وب است.

در مجموع، استدلال ها در برابر ضرورت و همچنین امکانات بایگانی وب است. جای تعجب نیست که از نظر ما، در مغایرت با نقش مرکزی وب در خلق فرهنگ و انتشار آن و همچنین بر اساس طبیعت مطلقش، ایجاد شده است. فصل 2 بینش بیشتری درباره چگونگی

اهمیت بایگانی های وب برای پژوهش در بسیاری از حوزه ها را فراهم می کند.

در اینجا، سعی می کنیم نشان دهیم که در صورتی که چالش های جدی و مهمی را برای عملکردهای سنتی مطرح کردیم، بایگانی وب امکان پذیر می شود و یکی از موارد اصلی در برنامه حفاظت از میراث فرهنگی امروز است.

3-ویژگی های وب برای حفاظت

وب خصوصیات مهمی دارد که هر تلاش حفاظتی باید درباره آن انجام شود. ما آن ها را در این بخش در

زوایای مختلف مورد بررسی قرار می دهیم. ابتدا، کاردینالیتی وب است یعنی اینکه چه تعداد نمونه از هر قسمت محتوا موجود است. و دوم اینکه وب به عنوان یک نظام انتشاراتی فعال و آخرین مورد وب به عنوان یک محصول فرهنگی جهانی، با طبیعت مافوق رسانه ای و طبیعت نشر آزاد بودن آن مورد بررسی قرار گرفته است.

3-1-کاردینالیتی وب

نخستین سؤال که برای حفاظت از محصولات فرهنگی عنوان شده است، کاردینالیتی بودن آن است. تعداد مواردی که هر اثر در حال توزیع شدن است. بایگانی ها و موزه ها معمولاً، با محصولات منحصر به فردی سر و کار دارند و حتی اگر در برخی موارد چندین قالب وجود دارد که کپی یا نشانه هایی از یک پیکر تراشی واحد، نقاشی یا اثر عکاسی است.

بر عکس، کتابخانه ها تقریباً موارد غیر منحصر به فرد را در مجموعه چاپی نگهداری می کنند (حفاظت از نسخه خطی، از این دیدگاه نزدیک به عملکرد بایگانی می باشد) منحصر به فرد بودن دارای اهمیت اجتماعی و نمادین عمیق است (بنجامین 1963) (1). همچنین، اثر بسیار زیاد آشکاری بر عملکردهای حفاظتی دارد. کتابخانه ها همیشه یک فرصت ثانویه برای یافتن کتاب های چاپ شده بعد از انتشارشان دارند.

چنین تخمین زده می شود بیشتر از 20 میلیون کتاب برای 30/000 ویرایش، بین 1455 و 1501 به چاپ رسیده است (فبوره و مارتین، 1976) که به این معنا که به طور میانگین نخستین دوره کاردینالیتی، متجاوز از 650 بوده است. کاردینالیتی مستلزم این است که حفاظت با تأخیر معینی بعد از انتشار رخ دهد، همان طور که کپی های متعدد برای یک دوره زمانی حتی در غیاب حفاظت فعال باقی می مانند. همچنین، یک سطح طبیعی از افزونگی یک ویژگی در یک نظام را موجب می شود که کتابخانه متفقاً انجام می دهند. با استفاده از داده های یکی از بزرگ ترین پایگاه داده های کتابشناختی (worldcat) لاویه و شانفلد (2005) (2) سه ردیف توزیع کاردینالیتی اثر منتشر شده در کتابخانه ها را کشف نمودند که از Worldcat استفاده کرده است (تقریباً 20/000 در آمریکای شمالی): 37 درصد فقط یکبار نگهداشته شده اند، 30 درصد، 2- 0 بار و 33 درصد بیشتر از 5 بار نگاه داشته شده اند.

زمان و افزونگی (تکرار اطلاعات میان فایل های گوناگون) دو مزیت قابل توجه از یک چشم انداز حفاظتی هستند که یکدیگر را تقویت می کنند. آن ها همیشه وجود ندارند. تولید مجدد نسخه های خطی در رابطه نقایص آن برای قرن ها قبل از اختراع چاپ اقدام شده بود، از این رو، اکنون حتی زمانی که چندین نسخه (واقعاً تعدادی) گوناگون وجود دارد. کتابداران بزرگ ترین کتابخانه قدیمی اسکندریه (3) را تشکیل می دهند که از کپی برداری نسخه های خطی استفاده می کردند که به دورن شهرها انتقال یافته بودند اما آن ها

1- Benjamin

2- Lavoie and Schonfeld

3- Alexandria

نسخه اصلی را نگهداری می کردند (کن فورا، 1989).

ترجمه و جمع آوری، تفسیر و توضیحات و حاشیه نویسی، غالب اوقات بنیاد و پایهٔ اصلی برای تولید مجدد متن به جای حفاظت قابل اعتماد بوده اند که برای زیان و ضررهای اجتناب ناپذیر اضافه شده بود که مستلزم کپی برداری دستی بودند. بیشتر کپی برداری های اصولی از متون، اغلب به دلایل خارجی ایجاد می شدند مانند وقتی متون یونانی، اساساً در موقعیت اختراع یک نوشته جدید، (حروف کوچک) در دوران امپراطوری رم در قرن نهم کپی برداری شده اند، تثبیت و انتقال آن ها به شکلی که ما امروزه می شناسیم خواهد بود.

کپی برداری آینده به طور قابل توجهی، شرایط را در این رابطه تغییر می داد که آن محتوا را به حالت تثبیت کرد زمانی که توزیع گسترده تر آن را مجاز کرد (آیزنشتاین 1979 (1) ؛ فبوره و مارتین (1976) (2). همچنین، با افزایش قابل توجهی از کاردینالیتی آثار، راندمان حفاظت را بدون سابقه کرد. در جایی بر آورد شده است که یکی از 40 اثر شناخته شده از دوران قدیم، حفظ شده است (و کمتر اگر آثار ناشناخته را در نظر نگیریم). راندمان حفاظت به بیشتر از یکی از دو تا در قرن هفدهم در فرانسه و نزدیک 80 درصد یک قرن بعد از آن (استیوالز 1965) (3) بالا رفته است و برای یک مؤسسه واحد، کتابخانه سلطنتی نیز بعد از تقویت سپرده گذاری قانونی توسط فرانسس لر (4) در سال 1537 (استیوالز، 1961؛ بالیه، 1988) (5).

امروزه، حفاظت از کارهای چاپ شده، در بیشتر کشورها به کارآیی و رشد مؤثری دست یافته است؛ از نقطه نظر عملی و سازمانی که با ثبات مطالب چاپ شده و همچنین کار دینالیتی، مجاز شده اند.

هر آن چه که بود، کاردینالیتی در محصولات فرهنگی حداقل از ایجاد تا دسترسی یکپارچه می باشد. این تنها مورد در وب نیست. کاردینالیتی محتوایی وب ساده نیست، بلکه هر یک (چند جزئی) است. همان طور که منبع محتوا معمولاً یک سرور منحصر به فرد است شخص می تواند به طور محسوس نماید که کاردینالیتی اش مانند آثار هنری و نسخه های خطی، یکی می باشد. در واقع، همان آسیب پذیری را نشان می دهد، حتی توسط این حقیقت که محتوا به تولید کننده خود وابسته است، افزایش می یابد. از طرفی دیگر، دسترسی و همچنین کپی های محتوای وب می تواند از نظر مجازی، نامحدود باشد. این اختلاف میان دو کاردینالیتی های وب، ما را به سمت این مفهوم مهم از منبع وب هدایت می کند. هر منبع، یک منبع منحصر به فرد (سرور وب) و یک شناسهٔ منحصر به فرد دارد، اما می تواند از نظر مجازی به طور نامحدود و با درجه های گوناگون برای برنامه ریزی های گوناگون تولید شود.

از دیدگاه حفاظتی، هر منبع دو خصوصیت مهم دارد:

نخست اینکه به طور دائمی به منشأ انحصاری اش برای موجودیت وابسته است. این کار یک تفاوت قابل توجهی با انتشار ایجاد می کند، جایی که مدیران انتشار، فقط یک بار به آن نیاز پیدا می کنند و بعد

1- Eisenstein

2- Febvre and Martin

3- Estivals

4- François 1er

از آن، کتاب ها به وجود می آینـد. دوم اینکه سـرورهای وب می توانند محتوا را برای هر نوع منبع، مناسب سازند و آن را در هر زمانی برای یو.آر.ال مشـابه متفـاوت می سـازد. وب از این دیـدگاه، یـک ظرف محتوی فایل های ثابت نیست، اما یک جعبه سـیاه با منابعی است که کاربران فقط نمونه هایی را به دست می آورند.

همان طور که کریشنامرتی (1) و رکسفورد (2) درباره پروتکل وب توضیح داده اند:

یـک روش برای ادراک پروتکل، این تصور است که منبع سـرور حاوی جعبه های سـیاه با نمایش منابعی باشـد که توسط یو.آر.ال.ها معنا شده اند. منبع سرور اصلی، شیوه درخواست برای منبع مشخص شده را به وسیله یو.آر.ال درخواست می کند. دریافت مشترک از خواندن یک منبع از یک فایل و نوشـتن پاسخ برگشت به سرویس گیرنده، دور از دید جعبه سیاه، مجزا و مختصر شده است. این نگرش، مفهوم یک منبع را عمومیت می بخشد و آن را از پاسخ ارسال شده به سـرویس گیرنده تفکیک می نماید. درخواست های مختلف برای یو.آر.ال.های مشـابه، می توانـد ناشـی از پاسـخ های متفاوت باشـد و به عوامل مختلفی بسـتگی دارد. فیلـدهای بالایی درخواست، زمان درخواست با تغییرات برای منابعی که ممکن است رخ دهند (کریشنامرتی و رکسفورد، 2001) حفاظت وب، منابع را طبق کاردینالیتی دوگانه (به ظاهر مهمل و در واقع درست) مورد بررسـی قرار می دهـد و این کار مسـتلزم چنـدین اسـتنباط است. نخست اینکه چون از نظر مجازی تعداد نامحـدود کپی برداری می توانـد به آسـانی ایجـاد شود، شـخص می توانـد ادارک گمراه کننـده ای داشـته باشـد که بایگانی فعال وب برای حفاظت مورد نیاز نیست. از این رو، تعـدد نمونه ها، به طور گسترده مخفی انـد و به یک منبع تکی بسـتگی دارند. هر زمانی که لازم باشد. (سرور) کد می تواند برچیده و روزآمد شود از این رو برای یک بایگانی فعال مورد نیاز می باشد.

استنباط دوم این است که بایگانی های وب می تواند فقط برخی موارد در منابع را به طور بالقوه درجات گوناگونی از میان آن ها را به تصرف در آورد (3) . این مورد زمانی رخ می دهـد که محتوا برای یک مرور گر خاص یک زمان معین یا یک موقعیت جغرافیایی معین یازمانی که محتوا با هر کار بر وفق داده شده است، مناسب گردد. همچنین، در بخش بعد خواهیم دید که وب در حقیقت یک سیسـتم نشر فعال است و از این رو، تفاوت پاسخ ها در واقع جنبه ای مهم برای بررسی است، به خصوص وقتی که بایگانی انجام می شود.

3-2-وب به عنوان سیستم نشر فعال

وب، نوعی برنامه کاربردی نشر اصلی در اینترنت است همچنین، به طور اساسی شامل ترکیب سه

1- Krishnamurthy

2- Rexford

3- توسـعه پویای صـفحه ها برای ایجاد یگانگی در طراحی و معماری نیز در کل سایت استفاده شده است (دستگاه های شناوری و مانند آن). اسـتفاده از تمپلیت ها به طور یکسان نگاه کردن به صفحه ها را آسان کرده و تغییر طراحی توسط تمپلیت ها را آسان تر از صفحه های انفرادی می کنند برآورد شده است که تمپلیت های مبتنی بر صفحه ها 40 درصد تا 50 درصد از صفحه ها را نمایش می دهند (جیسون و دیگران 2005).

استاندارد می باشد: 1) URL (لی –برنرز، 1994) که فضای نام گذاری شده را برای یک شیء تعریف می کند (1) ؛ 2) HTTP (فیلدینگ و همکارانش (2) ، 1994) پروتکل تعامل سرویس دهنده – سرویس گیرنده را با استفاده از فراپیوندها در هسته اش تعریف می کند؛ و 3) HTML (برنرز–لی و کونولی (3) ، 1995)، نوعی SGML DTD (4) (تعریف نوع داده) که ارائه صفحه آرایی در مرورگرها را معین می نماید. اجرای این سه استاندارد، هر کامپیوتر متصل به اینترنت را برای وارد شدن به سیستم نشر، قادر می سازد. شبکه سرویس دهندگان وب، نوعی سیستم اطلاعاتی منحصر به فرد را تشکیل می دهد که می تواند در هر حالتی برای تولید، روزآمد شدن، و نشر محتوا در حالتی که کامپیوترهای جدید اجازه می دهند، به کار رود.

در مقایسه با وسایل انتشاراتی دیگر، انقلاب در نشر، گسترش امکانات در تمام جهات ممکن برای تولید، سازماندهی، دستیابی، و ارائه محتوا را نشان می دهد. برای مثال، پیوندها را مورد بررسی قرار دهید:

شخص می تواند استدلال کند این فقط شکل جدیدی از یک ارجاع است که پیش از زمانی که برای نخستین بار نوشته شده به وجود آمده است (5) . اما حقیقت این است که روشی که به وسیله قطعه قطعه کردن محتوا به تکه های نشانی پذیر کوچک تر و توجه کلی به برنامه ریزی خاصیت انتقال از طریق شناوری دسترسی به محتوایی که تغییرات عمیقی را در نوشتن و همچنین در خواندن پیدا کرده است، به دلیل تغییرات وب در روش ارجاع قابل تعقیب قانونی است (آرسث (6) ، 1997؛ لندو (7) ، 1997؛ بولتر (8) ، 2001).

این حقیقت، که محتوا تنها بر روی سیستم و با دقت بیشتری بر روی سرورهای ناشران موجود است، به انتشار دائم ایجاد کننده بستگی دارد. یک کتاب می تواند بعد از ترک چاپخانه به طور مستقل از ناشر باقی بماند، اما محتوای وب هیچ موجودیتی فراتر از سرور اصلی اش نخواهد داشت (به استثنای ساز و کارهای حافظه با سرعت بالای ناپایدار (9) (هافمن و بیومونت، 2005). انتشار دائم، کنترل واضح را گسترش می دهد که ایجادکنندگان بر روی محتوا دارند. آن ها می توانند، با وب، در هر زمانی تغییر کنند، به روز شوند، و در زمان واقعی مواردی را از انتشار پاک کنند. علاوه بر این، تولیدکنندگان وب، از نظام اطلاع رسانی وب (10) استفاده می کنند که بتواند اطلاعات را از هر نوع نظام اطلاعاتی موجود (پایگاه داده ها، مخزن اسناد، برنامه های کاربردی، و مانند آن) ترکیب، مجتمع و سازماندهی مجدد کند. از این رو، وب یک فضای اطلاعاتی ثابت نیست، بلکه یک فضای نشر فعال است که ناشی از اثرات یک مجموعه آمیخته شده از نظام های اطلاعاتی فعال می باشد.

1– این استاندارد مهم ترین در میان سه مبتکر وب است (لی – برنرز و فیشتی، 2000؛ گیلیز کیلیو، 2000؛ چنان چه وب را در موقعیت دسترسی جهانی قرار داده است که کلاً منبع سندی قابل دسترس در اینترنت است.

2– Fielding

3– Connolly

4– SGML DTD

5– برای مقایسه استنادهای علمی سنتی و این که چگونه می تواند برای ارزیابی علمی استفاده شود اینگورسن (1998) را ببینید بجورن بورن و اینگورسن (2001) تحلیل انتقادی آن در توال (2001)، توال و هریس (2004) و توال(2006).

6– Aarseth

7– Landow

Bolter –8

Hofmann and Beaumont –9

(Web information systems (WIS –10

از این رو، بایگانی وب نخست به جدا کردن محتوا از نشر ثابت ایجادکنندگان اصلی اش نیاز دارد، و دوم اینکه باید مطمئن شود که محتوا می تواند از عدم پذیرش و تکامل جاری وب، عدول کند.

قبلاً به کپی برداری و بایگانی محتوا دریک زیر ساخت مجزا نیاز بود (مطالب زیر، (1) فصل 3 و روشه، 2006 را ببینید). مورد آخر مستلزم حفاظت فعال از محتوای وب (فصل 8، دی (2) 2006 را ببینید) برای رفع وابستگی از اجزای سیستم های گوناگون (پروتکل ها، به فرمت های دیجیتال، برنامه های کاربردی، و نظیر آن) و اجتناب از منسوخ بودن اصول فنی آن هاست. حفاظت از وب، این نیازها را در کل برای حفاظت از اصول فنی فعال با اشیای رقومی به اشتراک گذاشته است، اما جداسازی از ایجاد کننده نشر دائم، برای حفاظت از وب مشخص است.

اما رفع هر گونه وابستگی از سرور اصلی، مستلزم این است که از قابلیت های گوناگون و شیوهٔ تعاملی وب، بایگانی وب بتواند فقط تعداد کمی را حفظ کند. هزینه هایی برای جداسازی از شبکه اصلی نظام های اطلاعاتی وب وجود دارد.

قابلیت هایی که بر بخش سرویس گیرنده ها اجرا می شوند، آن هایی هستند که شخص می تواند به طور معقولانه به حفظ آن امیدوار باشد. دامنه قابلیت هایی در کد صفحه و کد فایل مربوط جاسازی شده اند که به میل سرویس گیرنده اجرا می شوند و بیشتر اوقات بر روی نسخه های بایگانی قابل اجرا هستند؛ اما این قابلیت ها که توسط کد و یا با اطلاعات سرور تهیه شده اند جاسازی نمی شوند. جنبه های سندی مطالب اصلی هستند که گم شده اند (مانند انواع خاص تعاملات که شخص می تواند بر روی یک ویدئو ثبت نماید)، اما این کار فقط می تواند برای تعداد محدودی از صفحه ها و نقطه نظرات معین و شرایط خاص انجام شود (کریستنسن – دالسگاد (3) ، 2001 ؛ بروگر، 2005) (4).

3-3-وب به عنوان یک محصول فرهنگی

علاوه بر یک نظام نشر فعال، وب یک فضای اطلاعاتی با مشخصات خاص است. واژه وب در این مضمون، یک محصول فرهنگی رقومی گسترده را برگزیده است (لایمن و کاهله، 1998) که می تواند با حقایق زیر مشخص شود:

1- Roche

2- Day

3- Christensen-Dalsgaard

4- نقطه نظر طراح وبگاه در این مورد نیز جالب است. در دابرلی و دیگران (2002). جلیس هودج پیشنهاد می کند، سایت هایی که او در حال طراحی است بایگانی شوند: – درخواست برای پروپوزال؛ – بیان هدف و دلیل استفاده؛ – توصیف استفاده از مضمون (مثال های مورد نیاز)؛ – توصیف استفاده کنندگان واقعی و مورد هدف؛ – جایگزین های ثابتی که به اندازه کافی دیدگاه و احساس را احاطه می کنند؛ – مثال هایی از چند راه مهم در سایت؛ – توصیف فناوری هایی که به کار برده شده و یا حمایت می کنند؛ و – هر ماژول مرتبط مثل پویا نمایی های فلش، فیلم ها، PDF ها، و مانند آن.

– از هر محلی که متصل به اینترنت است قابل انتشار و دسترسی (معمولاً مجانی) می باشد؛

– به عنوان یک فوق رسانه با استفاده از پیوندهای مستقیم و قابل تعقیب قانونی بین قطعات محتوا ساخت

یافته است (1) ؛

– نه تنها شامل متن است، بلکه شامل ترکیبی از تصاویر، صداها، و محتوای متنی نیز می باشد؛ و

– ناشی از یک تألیف و تصنیف باز و توزیعی می باشد (2) .

اگر چه وب، این کارها را به طور بسیار گسترده انجام می دهد، از اشکال قبلی انتشار هم می نماید (3) (کراستون و ویلیامز 1997) (4)؛ اریکسون و آیهلستروم، 2000؛ شفرد و پولانی، 2000). همچنین، مواردی جدید را ابداع می کند. برای مثال، بلاگ ها با تلفیق ساده گسترده ای منتشر می شوند (حتی مهارت های فنی که برای سایت های عادی لازم است و بیش از اینها مورد نیاز نیستند). و یک مدیریت مرجع قدرتمند (شامل مراجع معکوس یا آگاه سازی از استنادها با استفاده از برگشت پینگ) و سهولت در به روزرسانی، افزودن توضیحات و حذف محتوا، همه اینها ناشی از یک نشر آزاد و توضیحات شخصی به وسیله ده ها میلیون نفر بوده است (5) .

این خصوصیات وب، به عنوان یک فوق رسانه توزیعی، به طور آشکار و ثابت در یک مقیاس سراسری، تألیف و فراهم شده است که بایگانی وب بتواند فقط، به حفاظت از جنبه های محدود شده محصولات فرهنگی موجود و بزرگ تر دست یابد.

اتصالات درونی محتوا، یک کیفیت مهم در وب است که زمانی که بایگانی می گردد، مباحثی از آن ها ایجاد می شود، اما به عنوان یک نتیجه عمومی، نشان داده است که بایگانی همیشه، بهترین نوع گزینش (به گزینی) را در برنمی گیرد، حتی اگر در موارد انتخاب دستی و سایت به سایت باشد. از این مباحث برای بایگانی های بزرگ و وسیع اجتناب می شود، تا جایی که ممکن است، قطع تداوم اطلاعاتی که وب ارائه می دهد (لایمن و دیگران 1998). یا برای تعریف یک هدف تحلیلی خاص برای زمینه یابی تصمیم های گزینش انجام می شود (بروگر، 2005).

اما در عمل، اجرای خزش، در ارتباط با موارد اولویت دار و سیاست گزینش دستی، اجزای بایگانی وب را در بر می گیرد که همیشه فقط یک برش در فضا و زمان وب اصلی خواهند بود.

چگونه این نمونه های معنی دار و جایگزین را در وب بزرگ تر ایجاد می کنید؟ چه استنباطی هایی

1– آیرون و مک کرلی(2003) نشان می دهند که یک سوم پیوندهای مستخرج از بیلیون ها صفحه از نقطه جهت مشابه می خزند، یک سوم از عرض، بالا یا پایین در سلسله مراتب جهت ها از همان سایت و یک سوم از سایت خارجی.

2– نویسندگی محدود به تعدادی از افراد نیست بلکه از طریق ده ها و صدها فرد توزیع شده است. مثلا تخمین زده شده است که در مورد فرانسه، گره های نشر، شخص یا ساختاری که منتشر می کند (ویراستاران نه نویسندگان) با انتشار در وب سه برابر اندازه توسعه یافته است: در حدود 5000 ناشر یا ساختار دهنده نشر به پنج میلیون سایت و سایت شخصی (منبع: انجمن فرانسوی دسترسی به اینترنت). این موارد

شامل وب نوشت ها نیست.

3- این در مورد چاپ نیز بود که برای مدت طولانی از دست نوشته ها و سازماندهی صفحه قبل از اختراع چاپ تبعیت می کرد (فبوره و مارتین، 1976).

4- Crowston and Williams

5- در مورد حفاظت از بلاگ ها انتلیج (2004) را ببینید.

در درک از آینـده کاری که وب انجام خواهـد داد داریـد؟ تمام این سؤال ها بایـد زمانی که بایگانی و را به کار گرفته ایـد مورد بررسـی قرار بگیرنـد. حتی تعریف کـاری که «وب اصـلی» انجـام می دهـد مجموعه تجربیـات اسـتفاده کننـدگان از وب یـا مجموع برنامه ریزی های محتواست که باید از قبل در نظر بگیریم که بر مجموعه ای از محتواهای ثابت بیشتر در نظر گرفته می شوند.

خصوصـیات دیگری که برای انـدیشه و سازمانـدهی مجـدد عملکردهای حفاظت سـنتی باید انجام طبیعت نویسـندگی باز وب است. در حقیقت، این موضوع فیلتر کردن و حفاظت ساختار را بسـیار مشکل می سازد که بر اساس ناشران و نویسندگان است. آن ها به اندازه بسیار قابل توجهی بر روی وب می باشـند و برای شـناسایی و ثبت مشـکل اند. گاهی اوقات، اطلاعات تألیف و تصـنیف، بر روی سایت موجود است، گـاهی اوقـات هم نیست، و گاهی اوقات در روش قابل اعتمادی نیز قابل دسترسـی نیست. تنها اطلاعات ثبت شـده (در یک روش کاملاً کنترل نشـده و آزاد)، اطلاعات درباره کسانی که نام دامنه را برای مدیرتی DNS مجاز می کنند. اگر چه این اطلاعات مقادیر بزرگی برای کامل کردن مطالب بایگانی شده در وب هستند، و به طور قطع، برای تفسیر و استفاده مستقیم آسان نیست.

به عنوان یـک محصول فرهنگی، وب، سـبک متفاوتی از سازمانـدهی اطلاعات و الگوهای ساختاری متفاوت را برای اسـتفاده در حفاظت ارائه می دهد. نشانه های پیوند محتواها و شـناوری کاربران از ساختارهای طبیعی که بایگانی ها باید بیشتر وقت ها برای سازماندهی موارد نمونه های جمع آوری شـده اسـتفاده کنـد. از این رو، خصوصـیات وب به تغییر شـکل و روش های حفاظتی عمیق نیاز دارد. رویکرد کل نگری برای بایگـانی وب، آمادگی بیشتر برای انطباق با خصوصـیات وب است، اما هر نوع بایگانی وب بایـد آن ها را در هسـته روش هایش مشارکت دهد.

4-روش های جدید برای رسانه جدید

کتابخانه ها آرشـیوها (بایگانی)، و موزه ها، روش های بسـیار کارآمـدی را با موضوعات مورد علاقه شان تطبیق داده اند که نقش مهمی در ساخت حافظه اجتماعی ایفا می کنـد. اگر چه بایـد بیشتر فراگرفته شوند و بتواند برای حفاظت از وب مجدداً اسـتفاده گردند. ماهیت وب و کیفیت های مورد نیاز، همان طور که دیده شد برای بررسی مجدد و تطبیق عملکردهای حفاظتی به ارث برده شده از این سنت طولانی در حفاظت از محصولات فرهنگی، فیزیکی می باشد. این بخش، یک بررسـی عمومی از روش های جدید و رویکردهایی را نشان خواهد داد که باید برای حفاظت از وب مورد استفاده قرار گیرند (فصل 3 تا 8 مباحث را با جزئیات تهیه می کنند).

4-1-حفاظت از وب و زیر ساخت های اطلاعات

قبل از شروع مبحث روش شناختی، باید درباره استقرار بایگانی وب در زیرساخت های اطلاعات (1) به طور کلی، و به ویژه در اینترنت پرسش هایی مطرح شود.

بورگمن (2000)، تعریف کتابخانه رقومی جهانی را مطرح کرد (pp47sq) و تفاوت میان نگرش تکاملی و انقلابی در فناوری اطلاعات را توضیح داد:

«نگرش انقلابی، کتابخانه هایی است که رقومی هستند و با پایگاه های داده ها به وسیله شبکه های کامپیوتری پیوند یافته اند و در کل، می توانند آرایه ای از خدمات را فراهم نمایند که کتابخانه ها را از ریشه برخواهند کند. نگرش تکاملی این است که کتابخانه های رقومی سازمان هایی هستند که به تهیهٔ محتواها و خدمات در شکل های گوناگون و فقط به عنوان پیش نیازهای مؤسسه ها ادامه خواهند یافت و سرانجام کتابخانه های مکملی خواهند بود که امروزه وجود دارند» (همان، ص 48).

او، تعریف میان گذر از تکامل – انقلابی را پیشنهاد کرده است: که «کتابخانه های رقومی یک گسترش، افزایش و یکپارچه سازی در نظام های بازیابی اطلاعات و مؤسسه های با اطلاعات چندگانه و کتابخانه ها که فقط یکی باشد را بیان می کند. محدوده قابلیت های کتابخانه های رقومی نه تنها شامل بازیابی اطلاعات است، بلکه ایجاد و استفاده از این اطلاعات می باشد».

موقعیت برای بایگانی های وب در جهتی که آن ها برای موجود شدن مورد بررسی قرار می گیرند و قبلاً فضای اطلاعات را ساخته اند، متفاوت است. همچنین، به طور آشکار قابل دستیابی اند. از روی سنجش و اندیشه، در این فضا، فقط به دروازه بان ها نیاز نداریم. همان طور که محدودیت های دسترسی فیزیکی وجود ندارند. در این زمینه، نقش بایگانی وب، در سازماندهی اطلاعات زیاد است.

کتابخانه های فیزیکی باید در هر دو سازماندهی فیزیکی و فکری اشیا ایجاد شوند و این طیف وسیعی از امکانات و انتخاب ها را میسر می سازد. همچنین، در حالی که دسترسی فیزیکی به محتوا را مدیریت می کنند، نقش میانجی گری اجتناب ناپذیری دارند. کتابخانه های رقومی، این نقش میانجی گری را به وسیله ایجاد محیط های همکاری و دانش وابسته به متن تحت تابع بنیانی جست و جو و دستیابی را گسترش می دهد (لاگوز و دیگران (2)، 2005).

بایگانی های وب در قسمت های مربوط به خود محتوایی بارگذاری شده با روابط تعبیه شده و قابل

1- مفهوم زیر ساخت به طور کل در استار و روهتلر (1994) با ابعاد مختلف تعریف شده است: – جاسازی (در هم جا دادن)، – شفافیت؛ – به دست آوردن یا دامنه (ساختار به یک حادثه عواحد یا عمل یک جانبه رسیده است؛ – به عنوان بخشی از عضویت فراگرفته شده است (اعضای جدید برای عضو شدن نیاز به یک آشنایی دارند)؛ – پیوندهایی با کنوانسیون ها؛ – بر پایه پیاده سازی ساخته شوند؛ و – در حالت تفکیک قابل رؤیت باشند. این مفهوم در مضمون ساختارهای اطلاعاتی در بورگمن،(2003،2000) بحث شده است

2- Lagoze et al.

تعقیب قانونی و ساختارهای اطلاعاتی غنی ایجاد شده که توسط میلیون ها نفر در سراسر جهان ویرایش می شود وقتی بایگانی های سنتی و کتابخانه ها نگرش سازمانی ،شخصی و ابزارها را در این محتوا ایجاد می کنند (مدخل های شخصی و وب سازان و مانند آن)، فقط در این ویرایش جهانی وب شرکت می کنند. این کیفیت را به عنوان متخصصان حوزه تقلیل نخواهد داد اما آن ها را در یک تلاش سازمانی بزرگ تر قرار خواهد داد.

به عنوان بایگانی ،وب آن ها مسئولیت های بیشتری دارند چون محتوا و مفهوم را تصرف کرده و تحت سیطره خواهند داشت و می توانند آزمایش هایی را در بهبود نقش منحصر به فرد قدیمی خود در سازمان دهندگان اطلاعات داشته باشند همچنین می توانند فقط برای تثبیت و محافظت از نمونه های شخصی از یک محصول فرهنگی جاودانی بزرگ تر به دست آیند.

این کار می تواند قانونی باشد زمانی که طبق خط مشی گزینش مناسب با نیاز جامعه کاربران برپا یا توسط اهداف پژوهشی معین شده اداره شود لچر (1) ، 2006؛ میزانس، 2006 ب). اما هزینه ها و محدودیت ها و همچنین امکانات فنی برای بایگانی هر دو در یک مقیاس بزرگ تر و در یک مسیر بی طرف نیز نیازمند بررسی روش های جایگزین در بایگانی وب می باشند این جایگزین در نقش خود معتدل تر ولی در دامنه جاه طلب تر است. نقش سازمان دهندگان اطلاعات برای به تصرف در آوردن محدود است و برای ساختار اصلی ایجاد شده توسط ویرایش میلیون ها نفر در سراسر جهان درست است.

در مشکل رسیدن به جامعیت، همچنین در بخش قبلی دیدیم که شخص می تواند حداقل تلاش در بی طرفی را برای تصرف محتوا با پیروی از ماهیت جمع آوری و توزیعی وب برای راهنمایی در تصرف و گسترش آن، تا جایی که ممکن است داشته باشد از این رو تلاش بر روی کمیت و موضوع مقیاس گذاری بر روی منابع فنی می باشد. این رویکرد، چندین کتابخانه ملی برای حوزه ملی و بایگانی اینترنتی در یک مقیاس جهانی داشته است.

هیچ یک از این ابتکارهای عملیاتی نمی توانند از نظر عمق و کیفیت محتوای بایگانی شده به تنهایی توسعه یابد. تلاش های گوناگونی به عنوان بخشی از یک بایگانی جهانی مورد بررسی قرار خواهد گرفت البته وقتی که اتصالات درونی میان بایگانی وب به عنوان اتصالات درونی بین سرورهای نشر از طریق وب سازماندهی شوند.

تنها با این کار ،کاربران قادر به نفوذ به تمام این تلاش ها خواهند بود و به بهترین حافظه ممکن وب منتج خواهد شد. در این جهت، هر چه شراکت بیشتر مؤسسه ها و افراد مختلف وجود داشته باشد، بهتر می توانند مکمل یکدیگر باشند و زوایا ،عمق و کیفیت های مختلف بایگانی های متفاوتی را ارائه دهند. اما این کار نیازمند این است که آن ها در برخی نقاط از یک شبکه بایگانی وب بزرگ تر، همکاری کنند. چنین شبکه ای باید بایگانی وب را پیوند دهد، به طوری که با یکدیگر نوعی فضای شناوری جهانی مانند خود وب را شکل دهند این امر فقط در صورتی ممکن است که آن ها در یک مسیر نزدیک به وب اصلی ساخته

Lecher -1

شوند و آزادانه قابل دسترسی باشند کنسرسیوم (شرکت) حفاظت از اینترنت بین المللی (1) بر روی ایجاد و تنظیم زمینه هایی برای ایجاد کننده به وسیله توسعه استانداردها و ابزارهایی که ساخت این نوع بایگانی را تسهیل می کند عملکردهایی داشته است (برخی از آن ها در پایان این بخش توصیف خواهند شد).

دسترسی باز در خصوص مقررات و خط مشی است و در این برهه از زمان به عنوان یک بحث آزاد باقی می ماند.

بایگانی وب، به طور اختصاصی یا به عنوان کل می تواند در ایجاد زیر ساخت های اینترنتی مناسب باشد. آن ها از پروتکل ها و استانداردهای مشابه برای سازماندهی اطلاعات و ایجاد دسترسی به آن استفاده کنند. وب می تواند به طور طبیعی آن ها را در برگیرد چون آن ها کاملاً با آن سازگار هستند (2) . از نقطه نظر زیربنایی بایگانی وب می تواند به آسانی موقعیتی را به عنوان مکمل ایجاد زیر ساخت های اینترنتی پیدا نماید. آن ها در حال فراهم کردن حافظه وب هستند که خود بخشی از وب است واثر شدید منفی ماهیت ضروری ناپایدار انتشار وب را محدود می نماید.

شخص می تواند با چشم پوشی از این نقش ناراضی باشد شرایط این کار در نظر نگرفتن ارزش طبیعت توزیع گراو جامع بودن این رسانه است که آن را توجیه می کند.

4-2-فراهم آوری

اصطلاح فراهم آوری برای معانی فنی گوناگونی به کار می رود، مانند رسیدن محتوا به درون بایگانی. این اصطلاح شامل تصرف پیوسته و غیر پیوسته تحویل محتوا می شود که فرآیند انتخاب را پوشش می دهد و نه فرآیند درج و توسعه فرداده را.

از دیدگاه فنی این مرحله از تعامل با تولید کنندگان و در صورت سنتی برای مؤسسه میراث حافظه، هر چیزی می تواند با شرایط در بایگانی وب جزئی می باشد. زیرا هیچ روش واحدی برای فنون انتشار گسترده وب کافی نیست گسترده سازی دامنه سازندگان و افزایش اندازه محتوا به درجه ای معین با خودکارسازی که در محیط وب ممکن می شود متعادل می شود. از این رو، موانع اصلی اکتساب ابزارهایی است که باید بر آن غلبه یابند و عدم توانایی پروتکل HTTP در ایجاد کپی دسته ای از محتوای سرور می باشد. سرورهای HTTP فقط می توانند تا زمانی که URI درخواست نماید، فایل به فایل تحویل دهند. این کار موجب کشف گذرگاه فردی برای هر فایل را در یکی از مباحث کلیدی در بایگانی وب

می شود.

در این بخش سه نوع روش فراهم آوری را بررسی می کنیم. چرا سه روش؟ چون فرآیند جمع آوری می تواند یا به عنوان یک سرویس گیرنده دور افتاده، در نزدیکی به خروجی سرور انجام شود یا به وسیله دسترسی مستقیم به فایل های سرور صورت گیرد (تصویر 1). گزینهٔ نخست با خزنده بایگانی یا ماشین

1- International Internet Preservation Consortium (IIPC)

2- فرد می توانـد بحث کنـد که یک بازگشت قهقرایی بالقوه در اینجا وجود دارد در مورد آن چه که با آن مخالفت می شود که آرشـیوهای وب باید از بایگانی کردن سایر آرشیوهای وب اجتناب ورزند و خود را با وب زنده (موجود) محدود کنند.

کپی کننده وبگاه انجام می شود، مشتق شده و در فناوری موتور جست و جوی سازگار و استنتاج می شود یک ابزار قدرتمند برای تصرف موقعیت سرویس گیرنده را فراهم می کند. روشه (2006)، توصیفی با جزئیات این ابزارها و کاربردشان برای بایگانی وب ارائه می دهد. در این مقاله فقط یک بررسی عمومی از این فناوری را ارائه خواهیم داد که برای ارزیابی در هر یک از موارد می تواند به کار برده شود. چون خزش گر، برای سرور وب یک سرویس گیرنده مانند دیگران است از اصطلاح «بایگانی جانبی سرویس گیرنده» برای این روش فراهم آوری استفاده می کنیم. بسته به ساختار پسین وب و سطح تعامل با سرویس گیرنده خزش گر ها می توانند یا یک وبگاه کامل و یا بخشی از اجزای آن را به تنهایی تصرف کنند. جزء باقی مانده برای خزش گرها غیر قابل دسترسی است در اصطلاح شناسی موتور جست و جو «وب عمیق» یا وب پنهان نامیده شده است.

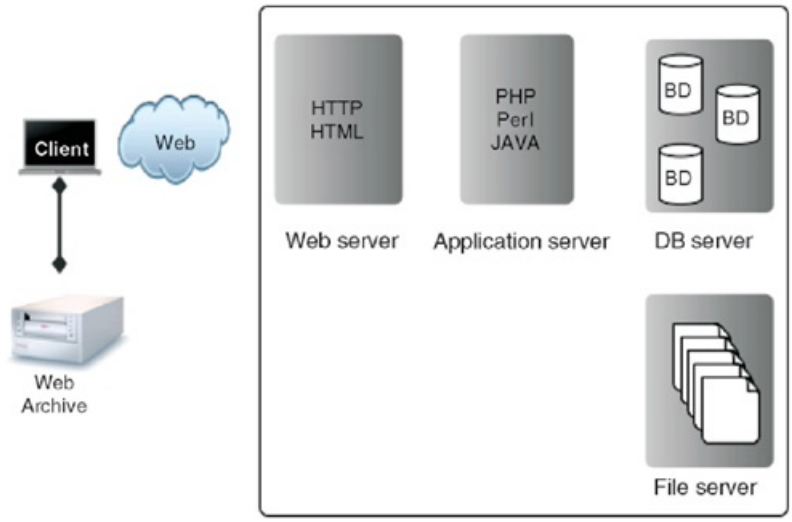
این اصطلاح شناسی را تصدیق خواهیم کرد تا زمانی که مشخص شود که تعیین حدود وب مخفی، به طور محض فنی بوده و به طور دائم به عنوان خزنده ها توانایی شان را در یافتن راهی برای اسناد، بهبود بخشد. دو روش متناوب برای گردآوری محتوا باقی می ماند، حتی اگر بسیار کم به کار برده شوند و حتی هم چنان تحت تحقیق و بررسی مانده باشند هر دو نیاز به فعالیت از جانب سرور دارند که نه تنها به یک مجوز نیاز دارد بلکه همچنین به یک مشارکت فعال انتشار دهنده سایت برای استفاده شدن نیز نیاز دارد. اولی، بر اساس کاربران سایت است که مسیر هدایت را توسعه می دهد و شناسایی محتوای سایت را برای بایگانی آن انجام می دهد و چون بر اساس ثبت تراکنش های ایجاد شده بین کاربران سایت و سرور می باشد آن را بایگانی تراکنشی می نمامیم.

دومی، شامل بایگانی به طور مستقیم از ناشر اجزای گوناگون سیستم اطلاعات وب و انتقال آن ها به یک شکل بایگانی شده می باشد بنابراین بایگانی جانبی سرور نامیده می شود. این روش های متناوب، سخت تر از بایگانی جانبی سرویس گیرنده می باشند چون همانطور که در بالا ذکر شد، نه تنها به یک مشارکت فعال از جانب سازندگان نیاز دارند بلکه باید بر مبنای مورد به مورد اجرا گردند اما حتی اگر افزایش مقیاس نداشته باشند می توانند در مواردی به کار برده شوند؛ برای مثال در جایی که خزنده برای تصرف دقیق موفق نمی شود و زمانی که محتوا در آن کمتر به کار رفته باشند.

4-2-1-بایگانی جانبی سرویس گیرنده

این روش مهمی در فراهم آوری است در هر دو، به علت سادگی، قابلیت مقیاس پذیری، و تطبیق با یک محیط سرور – سرویس گیرنده می باشد (شکل 1-1 را ببینید). خزنده ها با آن چه که روش معمول در دستیابی به وب است سازگار شده اند و این کار بایگانی هر سایت را مجاز می سازد که آزادانه به وب باز شبکه داخلی یا خارجی دسترسی داشته باشند تا زمانی که خزنده به یک اجازه مناسب دست یابد.

عکس



تصویر ۱- آرشیو سمت کاربر (Client-side): آرشیو وب در تقابل با آرشیو کاربر است تا محتوا را از سرور وب جمع‌آوری کند. سرور وب می‌تواند محتوا را از سرورهای متعدد و متنوعی فراهم کند (درخواست‌ها، پایگاه داده‌ها، فایل‌های سرور).

این روش، نه تنها موقعیت مشابهی را که کاربران عادی وب نیاز دارند، اتخاذ می‌کند؛ بلکه شکل تعامل‌هایش با سرورها را تقلید می‌کند. خزش‌گرها از صفحه‌ها هسته شروع می‌شوند، آنها را تجزیه می‌کنند، پیوندها را برداشت می‌کنند، و سند پیوندی را واکشی می‌کنند؛ سپس آنها این پردازه با سند واکشی شده را تکرار می‌نمایند و تا زمانی که پیوندهایی را کشف کنند[1] و سند را درون حوزه تعین شده پیدا کنند. این پردازه مورد نیاز است، چون HTTP فرمانی را که باید فهرست کامل سند قابل دسترس بر روی سرور را باید بازگرداند که برای مثال بر خلاف FTP است. از این رو، هر صفحه بایدبه وسیله پیوندی از صفحه‌ها دیگر «کشف» شود.

فناوری خزش، برای اهداف نمایه‌سازی توسعه یافته است[2]. به‌کار بردن آن برای بایگانی وب، برخلاف اینکه جنبه‌های بیشتری از این فناوری را دوباره استفاده می‌کند، تغییراتی را برای آن ایجاد می‌کند.

نخست اینکه، خزش‌گرهای بایگانی باید برای واکشی تمام فایل‌ها تلاش کنند هرچه فرمتشان برای بایگانی یک مدل کامل از سایت‌ها باشد بر خلاف خزش‌گرهای موتور کاوش که معمولاً فقط فایل‌هایی را واکشی می‌کند که آنها بتوانند فهرست نمایند. خزش‌گرهای موتور کاوش، برای مثال اغلب از انتقال از

۱. برای نظرات اخیر درباره فناوری خزشگر پلنت و دیگران (۲۰۰۴) و چاکرابارتی (۲۰۰۲) را ببینید.
۲. برای نظر اخیر درباره توسعه موتور کاوش تجاری سوننریچ (۱۹۹۷) را ببینید.

تصویر 1- آرشیوسمت کاربر (client-side): آرشیو وب در تقابل با آرشیو کاربر است تا محتوا را از سرور وب جمع آوری کند. سرور وب می تواند محتوا را از سرورهای متعدد و متنوعی فراهم کند (درخواست ها، پایگاه داده ها، فایل های سرور).

این روش، نه تنها موقعیت مشابهی را که کاربران عادی وب نیاز دارند، اتخاذ می کند؛ بلکه شکل تعامل هایش با سرورها را تقلید می کند. خزش گرها از صفحه ها هسته شروع می شوند، آن ها را تجزیه می کنند، پیوندها را برداشت می کنند، و سند پیوندی را واکشی می کنند؛

سپس آن ها این پردازه با سـند واکشـی شده را تکرار می نمایند و تا زمانی که پیوندهایی را کشف کنند (1) و سند را درون حوزه تعین شده پیدا کنند این پردازه مورد نیاز است چون HTTP فرمانی را که باید فهرست کامل سند قابل دسترس بر روی سرور را باید بازگرداند که برای مثال بر خلاف FTP است. از این رو هر صفحه بایدبه وسیله پیوندی از صفحه ها دیگر «کشف» شود.

فناوری خزش، برای اهداف نمایه سازی توسعه یافته است (2) . به کار بردن آن برای بایگانی وب، بر خلاف اینکه جنبه های بیشتری از این فناوری را دوباره استفاده می کند تغییراتی را برای آن ایجاد می کند.

نخست این که خزش گرهای بایگانی بایـد برای واکشـی تمام فایل ها تلاش کنند هر چه فرمت شان برای بایگانی یک مدل کامل از سایت ها باشـد بر خلاف خزش گر های موتور کاوش که معمولاً فقط فایل هایی را واکشـی می کنـد که آن ها بتواننـد فهرست نماینـد. خزش گر های موتور کاوش برای مثال اغلب از انتقال از

1- برای نظرات اخیر درباره فناوری خزشگر پلنت و دیگران (2004) و چاکرابارتی (2002) را ببینید.

2- برای نظر اخیر درباره توسعه موتور کاوش تجاری سوننریج (1997) را ببینید.

فایل های کاربردی و ویـدئویی بزرگ چشم پوشـی می کنـد. بارگذاری این نوع فایل ها می تواند تفاوت قابل توجهی را در رابطه با زمان و پهنای باند مورد نیاز برای خزش در کل سایت ایجاد نماید.

تفاوت دوم، با مدیریت موقتی خزش ها مرتبط می باشد. برای اجتناب از اضافه بار سرورهای وب، خزش گر ها از قوانین مطلوب و معتبری استفاده می کنند (روشه، 2006) . این مستلزم این است که تصرف یک وب می تواند در طی چندین دقیقه منتهای مراتب چندین ساعت و گاهی اوقات چنـدین روز طول بکشد. یک محاسـبه ساده نشان می دهد زمانی را که در خصوص یک تأخیر 3 ثانیه ای بین دو درخواست، بیشتر از 3 روز برای بایگـانی یک سایت با 100/000 صـفحه طول خواهد کشـید. این تأخیر، مبحث ثبات موقت تصرف را که دستخوش تغییراتی در طول زمان می شود، افزایش می دهـد. اگر صـفحه نمایه برای مثال در طی تصـرف تغییر یابد روش بایگانیاش با آخرین بایگانی که با صفحه ها بایگانی پیوند داده شده است سازگار نمی باشد.

این یـک مورد برای خزش گر هـای بایگانی است چون خزشـگر برای تهیه محتوا فرض شـده است البته نه فقط در جهت هـدایت محتواها خزش گر های موتور کاوش عادت دارند که به صفحه ها زنده اشاره کنند. به این معنی که مضمون ابرمتن برای آن ها، یکی است که توسط سرور اصـلی تهیه شـده است (که البته در سراسـر صـفحه هـا دارای ثبـات است و به هنگام سازی شـده است). بر عکس، خزش گر های بایگانی باید محتوا را به طور کلی تصـرف کنند که انجام خواهند داد با وابستگی یا بدون وابستگی داخلی، به عنوان فقط مضمون هایی برای شناوری و تفسیر.

این امر به مراتب نتایج قابل دسترسـی نسبت به خط مشی خزش گرها دارد. چون مطلوبیت برای سرورها حکم می کند که همیشه عملیات خیلی محدودی برای خزش داشـته است، خزش گر های موتور های کاوش از اولویت خزش در سـطح پهنای نخست اسـتفاده کرده اند با برخی تغییرات اساسـی که با خزش در بهترین صفحه های اول، هدفمند شده است (چو و همکاران، 1998؛ ناجون و هیدان 2001) (1) ؛ ناجورک و واینر 2001 (2)؛ کاستیلو و همکاران، 2004 ؛ باازا – یتس و کاستیلو (2005) (3) .

اتخاذ این خط مشـی، همچنین، روش برای به حداقل رسانیدن اثرات شدید تله های روباتیک (4) بر روی کل خزش با قرار گیری بیرون از خزش بر روی تعداد زیادی از سایت های مختلف می باشد.

اما این راهبرد زمان بندی خزش، دردسر افزایش تفاوت موقتی خزش ها در سطح سایت را دارد. ،بنابراین برای اتخاذ خزش های بایگانی در اولویت نخست یک سایت پیشـنهاد شده است. (5) اما برای خزش های مقیاس بزرگ، هنوز برای بهینه سازی بازده خزنـده با اطمینان از منابعی که در حداکثر ظرفیت خود استفاده شده اند، ضروری است تأخیر آشکار بین درخواست ها و منابع قابل دسترس خزش

1- Najork and Heydon

2- Wiener

3- Baeza-Yates and Castillo

4- فاصله ای عمدی در یک سیستم پردازش اطلاعات که به منظور جمع آوری تغییر یا خراب کردن آتی اطلاعات به وجود آمده است

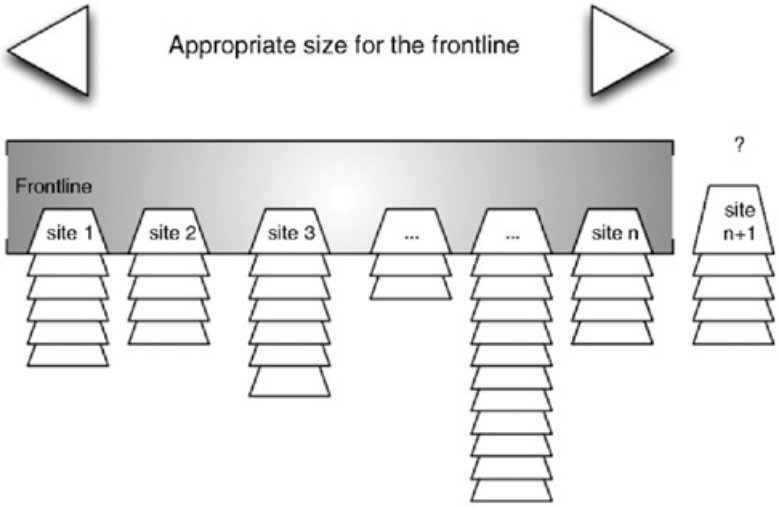
5- به طور موردی، برای ملزومات درون بحث شـده است (میسانس، 2004)؛ در مورد سـیاست های الگوریتم زمان بندی خزش که در سایت به عنوان بعدی عمودی آمیخته می شود منابع زیر را ببینید (Castillo et al. 2004, Baeza-Yates and Castillo (2005).

موجود است، شخص باید تعداد مطلوبی از سایت ها را برای شروع در زمان مشابه با ایجاد فرکانس های درخواست های مطمئن جست و جو کند که به وسیله قوانین مطلوب بدون هیچ تأخیر غیر ضروری بین درخواست ها واقع شده اند. شکل 2-1 خط مقدم یک خزش گر را نشان می دهد که اندازه ای مناسب با

تخصیص بهینه منابع خزش دارد.

عکس

موجود است، شخص باید تعداد مطلوبی از سایت‌ها را برای شروع در زمان مشابه با ایجاد فرکانس‌های درخواست‌های مطمئن جست‌وجو کند که به وسیله قوانین مطلوب، بدون هیچ تأخیر غیرضروری بین درخواســـت‌ها واقع شده‌اند. شکل ۲-۱ خط مقدم یک خزش‌گر را نشان می‌دهد که اندازه‌ای مناسب با تخصیص بهینه منابع خزش دارد.



تصویر ۱-۲. Frontline شامل سایت‌هایی است که باید با یک خزشگر یکسان و به‌طور همزمان خزش شوند. اندازه آن (n) در حد اُپتیمم است و اگر بین درخواست‌ها تأخیری رخ دهد. فقط به‌وسیله قوانین ساده‌ای محدود می‌شود و منابع خزش همچنان درگیر و مشغول خواهند بود. اگر n+1 سایت خزش شوند، محدودیت منابع خزش نوع تأخیر اضافی و عدم ربط موقت را به کار خواهند گرفت.اگر n-1 سایت خزش شوند، منابع بدون استفاده خواهند ماند.

محدودیت‌هایی برای هر آنچه که بتوان با استفاده ازاین روش بایگانی کرد وجود دارد. بیش از همه طی برداشـــت پیوند و برخی در طی بازیابی از طریق واســـط HTTP رخ می‌دهد. دلیل مورد قبلی این حقیقت است که URL استخراج شده، به‌طور نامساعد شکل گرفته است یا از پارامترهای پیچیده استفاده کرده‌اند، یا به‌سختی برای تجزیه کردن URL از فایل آغازگر یا فایل اجرا یا حتی کد HTLM استفاده کرده است. دومی می‌تواند به‌علت تجدید مسیرها، مذاکره محتوا، اجازه، پاسخ‌های تدریجی(کند)، اندازه نهایی، اتصالات TCP استثناهایی، پاسخ‌های سرور نامعتبر، و مانند آن باشد.

با اســـتفاده از این نوع ابزارها، فراهم‌آوری مقیاس بزرگی از متوا در یک مسیر کل نگ، که البته این طور نیست. مجاز می‌شود.

تصویر 1-2 Frontline شامل سایت هایی است که باید با یک خزشگر یکسان و به طور همزمان خزش شوند. اندازه آن (n) در حد اپتیمم است و اگر بین درخواست ها تأخیری رخ دهد فقط به وسیله قوانین ساده ای محدود می شود و منابع خزش همچنان در گیر و مشغول خواهند بود. اگر n+1 سایت خزش شوند، محدودیت منابع خزش نوع تأخیر اضافی و عدم ربط موقت را به کار خواهند گرفت. اگر n-1 سایت خزش شوند، منابع بدون استفاده خواهند ماند.

محدودیت هایی برای هر آن چه که بتوان با استفاده از این روش بایگانی کرد وجود دارد. بیش از همه طی برداشت پیوند و برخی در طی بازیابی از طریق واسط HTTP رخ می دهد. دلیل مورد قبلی این حقیقت است که URL استخراج شده به طور نامساعد شکل گرفته است یا از پارامترهای پیچیده استفاده کرده اند یا به سختی برای تجزیه کردن URL از فایل آغازگر یا فایل اجرا یا حتی کد HTLM استفاده کرده است. دومی می تواند به علت تجدید مسیرها، مذاکره محتوا، اجازه، پاسخ های تدریجی (کند)، اندازه نهایی، اتصالات TCP استثناهایی، پاسخ های سرور نامعتبر، و مانند آن باشد.

با استفاده از این نوع ابزارها، فراهمآوری مقیاس بزرگی از متوا در یک مسیر کل نگ، که البته این طور نیست. مجاز می شود.

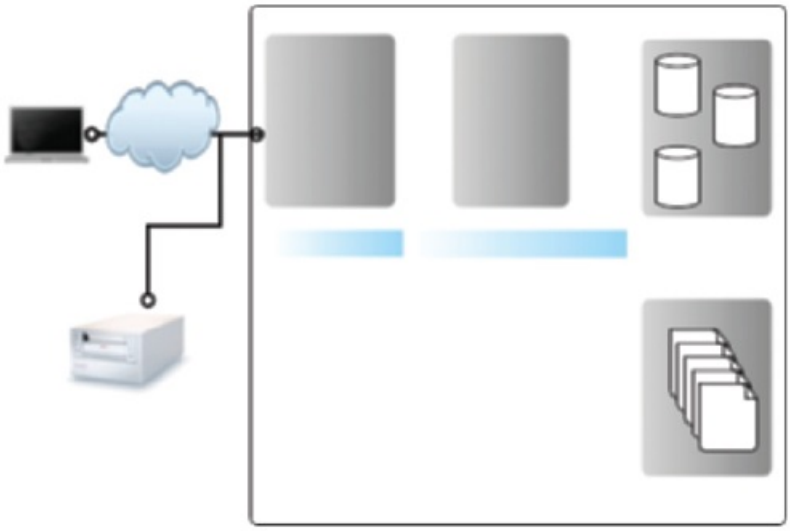
2-2-4-بایگانی تراکنشی

بایگانی تراکنشی (شکل 3-1 را ببینید) به وسیله فیچ (2003) (1) پیشنهاد شده است که شامل تصرف و بایگانی «تمام پاسخ های متمایز اساسی که توسط یک وبگاه تولید شده است و در رابطه با محتوایشان و چگونگی تولید می باشد». این کار در سیستم پرش صفحه (2) با استفاده از یک فیلتر درون ورودی سرور وب (درخواست) جریان و جریان خروجی (پاسخ) اجرا شده است. این توابع عملیاتی اکنون بر روی برخی از نظام های مدیریت محتوای وب مانند Vignette TM قابل دسترس هستند.

عکس

٤-۲-۲- بایگانی تراکنشی

بایگانی تراکنشـــی (شکل ۳-۱را ببینید) به وســـیله فیچ[1](۲۰۰۳) پیشنهاد شده است که شامل تصرف و بایگانی «تمام پاسخ‌های متمایز اساسی که توسط یک وبگاه تولید شده است و در رابطه با محتوایشان و چگونگی تولید می‌باشد». این کار در سیستم پرش صفحه[2] با استفاده از یک فیلتر درون ورودی سرور وب (درخواست) جریان و جریان خروجی (پاسخ) اجرا شده است.این توابع عملیاتی اکنون بر روی برخی از نظام‌های مدیریت محتوای وب مانند Vignette TM قابل دسترس هستند.



شکل ۳-۱بایگانی تراکنشی

جفت‌های درخواســـت/ پاسخ منحصر به‌فرد ذخیره و بایگانی شده‌اند از این رو، ایجاد یک بایگانی کامل از تمام محتوا برای یک سایت مشخص پیش‌بینی شده است. درخواست‌ها فقط با اندکی تفاوت(«غیر مادی») به‌صورت منحصر به‌فرد مورد بررسی قرار می گیرند، به استثنای محاسبه مجموع مقابله‌ای قسمتی از کد، که آنها را به‌صورت رمز در آورده است،

کیفیت دقیـــق اینها چگونه می‌تواند با تعداد زیادی از روش‌های شخصی‌ســـازی محتوا، که واضح نمی‌باشد، منطبق گردد.

ایـــن نوع بایگانی وب می‌توانـــد فواید پیگردی و ثبت هر برنامه‌ریزی ممکن در محتوا را ثابت کند.

1. Fitch
2. http://www.projectComputing.com/products/pageVault

شکل 3-1 بایگانی تراکنشی

جفت های درخواست /پاسخ منحصر به فرد ذخیره و بایگانی شده اند از این رو، ایجاد یک بایگانی کامل از تمام محتوا برای یک سایت مشخص پیش بینی شده است. درخواست ها فقط با اندکی تفاوت(«غیر مادی») به صورت منحصر به فرد مورد بررسی قرار می گیرند، به استثنای محاسبه مجموع مقابله ای قسمتی از کد، که آن ها را به صورت رمز در آورده است،

کیفیت دقیق اینها چگونه می تواند با تعداد زیادی از روش های شخصی سازی محتوا، که واضح نمی باشد، منطبق گردد.

این نوع بایگانی وب می تواند فواید پیگردی و ثبت هر برنامه ریزی ممکن در محتوا را ثابت کند.

ص: 31

Fitch -1

http://www.projectComputing.com/products/page Vault -2

محتوایی که هرگز دیـده نشـده، بایگانی نخواهـد شد (همان طور که ذکر شد، بوفخواد و ویناد (2003) (1) ، برآورد کردنـد که 25درصد صفحه ها از یک وبگاه بزرگ علمی، هرگز قابل دستیابی نیسـتند). اما محتوای وب پنهان تا زمانی که به دستیابی برسـد ثبت خواهـد شد و این یک مزیت مهم است.

محـدودیت مهم این روش این است که باید با موافقت و همکاری مالک ،سـرور، اجرا شود. از این رو، برای بایگانی داخلی وب نشان داده شده است. و این مسئله مزیت توانا بودن ثبت دقیق از چیزی را که و زمانی را که دیده است دارا می باشد. برای بایگانی شرکت یا سازمانی، اغلب به وسـیله جواب گویی قانونی، برانگیخته می شود. این کار، می توانـد یک مزیت باشد. حتی برای ترکیب با اطلاعات از سـرور ثبت وقایع درباره کسـی که محتوا را دیـده است، امکان پـذیر باشـد. به طور واضح، هر چه را که می تواند به عنوان یک مزیت برای بایگانی وب داخلی دیده شود ممکن است برای یک بایگانی عمومی مشـکل باشد، چون می تواند تأکید بر محرمانگی جدی را بالا ببرد. اما به هر حال محتواقابل استفاده نمی باشد.

3-2-4-بایگانی سرور- جانبی

نوع آخر از روش فراهم آوری بایگانی وب، کپی مسـتقیم فایل ها از سـرور، بدون اسـتفاده از واسـطه های HTTP است. این روش مانند روشی قبلی می تواند فقط با همکاری مالکان سایت استفاده گردد (شکل 4-1).

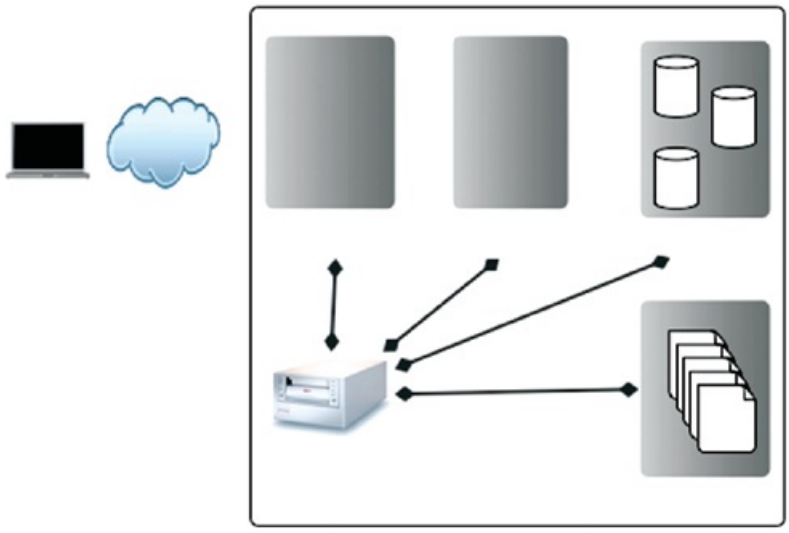
عکس

محتوایی که هرگز دیده نشـــده، بایگانی نخواهد شـــد (همان‌طور که ذکر شد، بوفخواد و ویناند[1]، ۲۰۰۳ برآورد کردند که ۲۵درصد صفحه ها از یک وبگاه بزرگ علمی، هرگز قابل دستیابی نیستند). اما محتوای وب پنهان تا زمانی که به دستیابی برسد، ثبت خواهد شد و این یک مزیت مهم است.

محدودیت مهم این روش، این است که باید با موافقت وهمکاری مالک سرور، اجرا شود. از این رو، برای بایگانی داخلی وب نشان داده شده است. و این مسئله مزیت توانا بودن ثبت دقیق از چیزی را که و زمانی را که دیده است دارا می‌باشد. برای بایگانی شرکت یا سازمانی، اغلب به‌وسیله جوابگویی قانونی، برانگیخته می‌شود. این کار، می‌تواند یک مزیت باشد. حتی برای ترکیب با اطلاعات از سرور ثبت وقایع درباره کسی که محتوا را دیده است، امکان‌پذیر باشد. به‌طور واضح، هر چه را که می‌تواند به‌عنوان یک مزیت برای بایگانی وب داخلی دیده شـــود، ممکن است برای یک بایگانی عمومی مشکل باشد، چون می‌تواند تأکید بر محرمانگی جدی را بالا ببرد. اما به هر حال محتواقابل استفاده نمی باشد.

٤-٢-٣-بایگانی سرور- جانبی

نوع آخر از روش فراهم آوری بایگانی وب، کپی مستقیم فایل‌ها از سرور، بدون استفاده از واسطه‌های HTTP است. این روش مانند روشی قبلی می‌تواند فقط با همکاری مالکان سایت استفاده گردد (شکل ٤-١).



شکل ٤-١.آرشیو Server – side: قطعه‌های مختلف اطلاعات به‌طور مستقیم از سرورها آرشیو می‌شوند. تولید نمونه کار محتوای آرشیو شده و یک نمونه پشتیبان (backup) از فایل‌ها چالش اصلی این روش است.

1. Boufkhad and Viennot

ص: 32

Boufkhad and Viennot -1

اگر چه بسیار ساده به نظر می رسد، در واقع مشکل زیادی را برای ایجاد محتوای کپی شده قابل استفاده افزایش می دهد و حتی در مورد فایل های ایستای HTML، شخص می تواند به زحمت در محتوا از طریق پیوندهای مطلق به عنوان نام دامنه هدایت شود که در بایگانی متفاوت خواهند بود. اما بیشترین مشکل ناشی از محتوای توسعه یافته پویاست که تکه های به هم پیوسته محتوا از منابع گوناگون (قالب ها (1) و پایگاه های داده) است که توسط درخواست های کاربرد در حالت پرواز در ارتفاع کم توسعه یافته است. کپی برداری فایل های پایگاه داه ها، قالب ها، و فایل های آغاز گر به این معنی نیست که آن برای تولید مجدد محتوا از بایگانی آسان خواهد بود. بر عکس یک وظیفه چالش برانگیز خواهد بود چون نیازمند توانا شدن برای اجرای در محیط مشابه، با پارامترهای مشابه در بایگانی است. در واقع، وقتی امکان پذیر شد، محتوای توسعه یافته پویا در شکل نهایی اش بهتر حفظ می شود، معمولاً در فایل های HTML مسطح بهتر حفظ می شود (برای مثال این موردی برای بیشتر CMSها، بلاگ ها و ویکی ها (2) صورت می گیرد).

اما گاهی اوقات این کار مشکل است، حتی برای خزش گرها یافتن مسیر برای برخی اسناد یک وبگاه و فایل هایی که می تواند از طریق یک تعامل پیچیده به دستیابی برسند، غیر ممکن است (مانند ارسال یک پرس وجو و یک فرم یا پنجره) که به سختی توسط خزش گرها تصرف خواهد شد این بخش وب، وب پنهان یا عمیق نامیده شده است (برگمان، 2001؛ چانگ و همکاران 2004) که بزرگ تر از وب سطحی است و (همچنین به طور عمومی وب قابل نامیده شده است). (3)

در این مورد، بایگانی جانبی سرور می تواند یک راه حل باشد، همان طور که در بالا ذکر شد، به مشارکت فعال مجری سایت نیاز دارد. بیشتر از یک پشتیبانی ساده است که دسترسی به محتوا را در نمونه های اصلی اش تضمین نمی کند، آن بر توانا بودن برای «نمایش» مجدد سایت در محیط بایگانی را دلالت می کند. این کاهش وابستگی به پایگاه داده ها و اجرای فایل های آغاز گر جانبی سرور به قدری که امکان پذیر است را نشان می دهد. این کار می تواند به وسیله استخراج اطلاعات ساخت یافته محتوی در پایگاه داده ها و انتقال آن درون XML انجام شود. نوعی معماری اطلاعات دروازه اسناد نامیده شده است که شامل اسناد غیر وب با آن هایی که به وسیله کاتالوگی می تواند مانند اینها بایگانی شود، قابل دستیابی می باشد. این کار، برای چندین سایت انجام شده است که به مقوله سایت های پنهان به وسیله کتابشناسی ملی فرانسه (4) وابسته می باشد (فصل 5 را ببینید).

این کار فقط در چارچوب واسپاری قانونی امکان پذیر می باشد که در فرانسه مانند بسیاری از کشورهای دیگر به کار گرفته می شود. حقیقت این است که وب پنهان همچنین غالب اوقات دارای محتوای بسیار غنی با این نوع معماری اطلاعات است که انبوه زیادی از محتوای از قبل موجود بر روی وب منتشر کرده است. عمومیت این نوع معماری اطلاعات، بایگانی سایت سرور را می سازد ،روشی که به جایی که می تواند به کار رود توجه دارد.

---

templatestemplates -1

wikis -2

3- این اصطلاح برای قسمتی از وب مشخص شده است که می تواند توسط خزش گرها نمایه شود (لورنس و گیلز، 1998،1999)

Bibliothè que nationale de France -4

3-4-سازماندهی و ذخیره سازی

همان طور که قبلاً دیده شد کپی برداری از یک وبگاه، یک وظیفه غیر پیش پا افتاده است. آن در واقع بر ایجاد مجدد یک سیستم اطلاعاتی اشاره می کند که برای کاربران قابل دسترسی خواهد بود. همان طور که آنتونیول و همکارانش (1) آن را در وبگاه قرار دادند «وبگاه ممکن است به سادگی یک فایل واحد یا یکی از پیچیده ترین مجموعه های محصولات مصنوعی نرم افزاری مشترک باشد که تا به حال درک شده است». به طور مطلوب، بایگانی می توانست در اصل متناظر (هم ریخت) (ساختار سلسله مراتبی مشابه، نام گذاری فایل ها، سازوکار پیونددهی، چارچوب) باشد، اما به دلایل عملی اینگونه نیست. همان طور که در بخش قبلی دیده شده اکتساب سایت ها در حقیقت موارد یک تغییر شکل فایل را به طور مؤثر کاهش می دهد.

چالش بیشتر، در ایجاد مجدد سیستم های اطلاعاتی مشابه می باشد. نظام اطلاعاتی وب معماری اطلاعاتی پیچیده را نشان می دهد که به سیستم های عملیاتی خاص پیکربندهای سرور و محیط کاربردی وابسته هستند که در بیشتر موارد حتی برای ایجاد مجدد استفاده از (فایل) چرکنویس برای طراحان و مدیران مشکل است به همین دلیل است آرشیویست ها مجبور به اتخاذ راهبرد تبدیل یا تغییر شکل می شوند. این تبدیل ها می توانند اثر شدیدی بر ساز و کارهای آدرس دهی و پیوندی، فرمت ها، و همچنین تغییر خود شیء تأثیر داشته باشند.

تاکنون سه راهبرد برای ساخت بایگانی وب اتخاذ شده است. راهبرد نخست برای ایجاد کپی محلی از فایل های سایت و هدایت از طریق این کپی در یک مشابه برای مثال بر روی وب می باشد.

راهبرد دوم، برای اجرای یک سرور وب با به کار بردن مضمون در یک محیط برای مرورگرهای کاربران است راهبرد دوم راه اندازی سرور وب و به کارگیری محتوا در این محیط برای مرورگرهای کاربران است. سومین راهبرد، سازماندهی مجدد اسناد منطبق بر نام گذاری متفاوت (غیر وب)، آدرس دهی منتقل کردن است. بخش های زیر موافقان و مخالفان این رهبردهای متفاوت و همچنین موارد استفادهٔ را نشان می دهد.

1-3-4-نظام فایلهای محلی به خدمت گرفته شده بایگانی ها

توصیف

این نوع بایگانی(شکل 5-1) بر اساس احتمالاتی است که مشخصه های URL از پیشوند سیستم فایل محلی، «فایل» (2) استفاده کند در یک شمای URL برای کپی و دسترسی به فایل های محلی از وبگاه اصلی استفاده کند مانند:

Http://www.example.org/example.HTML

File://Users/archire2005/eample.org/example.HTML

این، استفاده از سیستم فایل محلی را برای شناوری از طریق مواد بایگانی شده در وب، قادر می سازد.

Antoniol et al -1

file -2

همچنین به استفاده از شکل جزئی (وابسته) از URL که نه تنها از پیشوند دوری می کند بلکه همچنین از نام سرور و مسیر هدف نیز اجتناب می کند.

مرورگرهای استاندارد می توانند به طور مستقیم باز شوند (به طور مثال بدون سرور وب) و چنین فایل های ذخیره شده را محلی نمایند و تا زمانی که پیوندها اسناد وابسته هستند، شناوری در بایگانی مشابه مانند همانی است که در سایت اصلی خواهد بود، به جز فقط در قسمت آدرس دهی (1) مرورگر وقتی در جست و جوی پیشوند URL هستند. (در اینجا «فایل» به جای HTTP می باشد).

توضیح

فایده اصلی این راهبرد، تسهیل دستیابی به بایگانی با مسیردهی ساختار وبگاه اصلی به سوی نظام بایگانی فایل است. با استفاده از مرورگر استاندارد و سیستم فایل مجاز به اجتناب از هزینه های بالا که همراه است با اجرای دسترسی به بایگانی مبتنی بر سرور. بنابراین، حتی گروهی با مهارت های فنی فناورانه بسیار اساسی می توانند این نوع بایگانی را بر پا و اجرا کند. ولی محدودیت هایی در این رویکرد وجود دارد. از یک نگرش محافظه کارانه نقص اصلی این است که تبدیل های گوناگون فایل های اصلی، مورد نیاز می باشند. از این رو، صداقت محض برای فایل های اصلی، نمی تواند مورد ملاحظه قرار گیرد مگر به وسیله سندسازی با تغییرات دقیق که در فایل اصلی به کار می رود یا به وسیله نگهداری یک کپی از اصل. انتقال محتوا در دو سطح در رویکرد بایگانی «FS» محلی مورد نیاز می باشد.

نخست، به علت تفاوت در نام قرار داد بین URL و نظام فایل (تصویب و ذخیره کاراکترها، رهایی از قوانین، حساسیت موردی)، نام اشیاء ممکن است به تغییراتی نیاز داشته باشد (فصل B – 1 را برای نشان دادن جزئیات بیشتر این تغییرات ببینید). در مورد جایی که صفحه طبق پارامترها پرس وجو شده و به طور چشمگیری تولید شده است، یک نام حتی برای ایجاد شدن در صفحه نتیجه، پارامترهایی را برای اطمینان از منحصر به فرد بودن صفحه بایگانی شده تأمین می کند.

دوم اینکه، پیوندهایی مطلق باید در پیوندهای مرتبط در خودکد صفحه منتقل شوند تا شناوری مبتنی بر نظام فایل را امکان پذیر کنند. حتی اگر این کار بتواند به سادگی توسط تغییر شکل URL اصلی درون یک توضیح در کد، سندیت داده شود این دستکاری منبع را نشان می دهد.

از نظر عملی، اشکالات اصلی ناشی از خود نظام فایل می باشد که یک معماری اطلاعات متفاوت قابل توجه نسبت به وب دارد. نخست اینکه، سازماندهی بایگانی باید در سازماندهی سلسله مراتبی نظام های فایل مناسب باشد، با این حال، یک بایگانی نه تنها از ترکیب سایت ها، بلکه همچنین گروه هایی در سایت ها (مجموعه ها) و نسخه هایی از سایت ها تشکیل شده است. مسیردهی این سازماندهی برای یک ساختار سلسله مراتبی، بدون تغییر و انتخاب روی نمی دهد. چگونه سایت ها در یک حالتی که زمان پایدار، یک مبحث مهم برای بررسی می باشد، با یکدیگر در یک گروه قرار می گیرند. نام های مجموعه

1- Address bar

باید دائمی باشند و زمان گروه بندی نیز باید برای ضبط فرکانس ها، تطبیق یافته باشد. در مسیر تمام این مباحث، تصمیم گیری کامل باید پیشاپیش انجام شده باشد. آن ها بر روی چگونگی انتخاب ساختار که در برابر توسعه مجموعه مقاومت می کنند، تحت فشار قرار می گیرند. سازماندهی زمان در اجرای هر جمله از یک برنامه برای اهداف اشکال زدایی (از یک مدل به مدل دیگر) سایت نیز یک مبحث مهم برای هر یک از لایه های نرم افزار است که باید به سیستم فایل با استاندارد بالا افزوده شود. این لایه باید قادر باشد حداقل مدل های متفاوت سایت ها را که به تاریخ شان بستگی دارد با یکدیگر متصل نماید (نسخه پردازی) و این کار را با یک واسط کاربر مناسب برای شناوری از طریق زمان به سادگی ارائه می دهد. این کار اغلب با استفاده از مدیریت خارجی یک پایگاه داده در سایت ها و ضبط اطلاعات و ابزارهایی برای ایجاد صفحه های نمایش میانجی با فهرستی از تاریخ اجرا می شود که در آن سند بایگانی شده است.

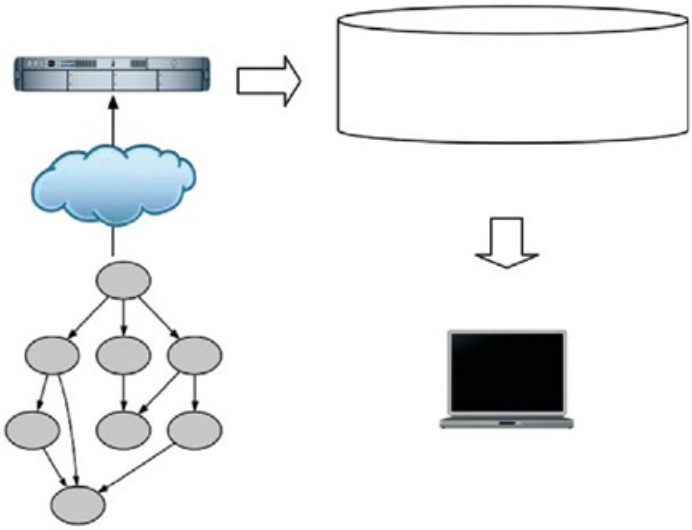
محدودیت دیگر این رویکرد، ناشی تعداد عظیم فایل های بایگانی وب است که باید جا به جا شوند. این مسئله عادی است که بایگانی هایی با بیلیون ها صفحه را ببینیم. این شکل به محدودیت های ظرفیت نظام های فایل جاری می رسد و دسترسی می یابد حتی وقتی یک نظام فایل بتواند این مقدار فایل را جابه جا کند، عملکرد می تواند تحت تأثیر قرار گیرد. برای کم کردن باری که بر روی نظام فایل قرار گرفته، بایگانی با مقیاس بزرگ برای فایل های مخزن استفاده شده است. البته این کار، ارتباط مستقیم در نام گذاری و پیوند دادن را که رویکرد بایگانی نظام فایل محلی برای اتخاذ رویکرد دوم ارائه می دهد را می شکند و بایگانی مبتنی بر سرور به کار رفته در وب برای تحویل محتوا از این فایل مخزن است.

شکل 1-5. بایگانی سیستم فایل محلی. فایل اصلی مورد خزش قرار می گیرد و صفحه ها و سایر فایل ها به صورت انفرادی روی سیستم فایل بایگانی ذخیره می شوند. دست یابی بوسیله شناوری مستقیم در نظام فایل انجام می شود.

عکس

باید دائمی باشند و زمان گروه‌بندی نیز باید برای ضبط فرکانس‌ها، تطبیق یافته باشد. در مسیر تمام این مباحث، تصمیم‌گیری کامل باید پیشاپیش انجام شده باشد. آنها بر روی چگونگی انتخاب ساختار که در برابر توسعه مجموعه مقاومت می‌کنند، تحت فشار قرار می‌گیرند. سازماندهی زمان در اجرای هر جمله از یک برنامه برای اهداف اشـــکال‌زدایی(از یک مدل به مدل دیگر سایت) نیز یک مبحث مهم برای هر یک از لایه‌های نرم‌افزار است که باید به سیستم فایل با استاندارد بالا افزوده شود. این لایه باید قادر باشد حداقل مدل‌های متفاوت سایت‌ها را که به تاریخ شان بستگی دارد با یکدیگر متصل نماید (نسخه‌پردازی) و این کار را با یک واســـط کاربر مناسب برای شـــناوری از طریق زمان به سادگی ارائه می‌دهد. این کار اغلب با استفاده از مدیریت خارجی یک پایگاه داده در سایت‌ها و ضبط اطلاعات و ابزارهایی برای ایجاد صفحه‌های نمایش میانجی با فهرستی از تاریخ اجرا می‌شود که در آن سند بایگانی شده است.

محدودیت دیگر این رویکرد، ناشی تعداد عظیم فایل‌های بایگانی وب است که باید جابه‌جا شوند. این مسئله عادی است که بایگانی‌هایی با بیلیون‌ها صفحه را ببینیم. این شکل به محدودیت‌های ظرفیت نظام‌های فایل جاری می‌رســـد و دسترســـی می‌یابد حتی وقتی یک نظام فایل بتواند این مقدار فایل را جابه‌جـــا کند، عملکرد می‌تواند تحت تأثیر قرار گیرد. برای کم کـــردن باری که بر روی نظام فایل قرار گرفته، بایگانی با مقیاس بزرگ برای فایل‌های مخزن استفاده شده است. البته این کار، ارتباط مستقیم در نام‌گـــذاری و پیوند دادن را که رویکرد بایگانی نظام فایل محلی برای اتخاذ رویکرد دوم ارائه می‌دهد را می‌شکند و بایگانی مبتنی بر سرور به‌کار رفته در وب برای تحویل محتوا ازاین فایل مخزن است.



شکل۱-۵. بایگانی سیستم فایل محلی. فایل اصلی مورد خزش قرار می‌گیرد و صفحه‌ها و سایر فایل‌ها به صورت انفرادی روی سیستم فایل بایگا نی ذخیره می‌شوند. دستیابی بوسیله شناوری مستقیم در نظام فایل انجام می‌شود.

استفاده ترجیحی

این روش برای بایگانی سایت شـرکت ها یا سازمانی و بایگانی که در مقیاس کوچکی توسعه یافته، توصیه شده است. بسته به استفاده از این بایگـانی، صـحت مبحث بایـد به دقت مورد بررسـی قرار گیرد مخصوصاً برای بایگانی سازمانی برای بایگانی با مقیاس کوچک، تعادل بین سختی برای سازمانـدهی مـداوم مجموع فایل ها و سادگی دسترسـی ایجاد شـده توسط این رویکرد باید بر اساس «مورد به مورد» ارزیابی گردد. برای بایگان های وب با مقیاس بزرگ و متوسط از این روش باید اجتناب شود.

ابزارها

این راهبرد برای بایگانی وب در مقیاس کوچک و متوسط ساده ترین رویکرد است که با بسـیاری از ابزار های موجود مانند HTTrack قابل دسترس است.

2-3-4-بایگانی های مبتنی بر خدمت وب

اشاره

*بایگانی های مبتنی بر خدمت وب (1)

بـا وجود تقاضاهای بیشتر، این گزینه تطبیق بهتری را برای نام گـذاری منبع و ساختار اسـناد ایجاد می کنـد (شـکل 6-1). همچنین، برای اجتناب از محدودیت های اندازه نظام فایل که برای بایگانی وب در مقیاس بزرگ بحرانی است، مجاز می کند.

عکس

استفاده ترجیحی

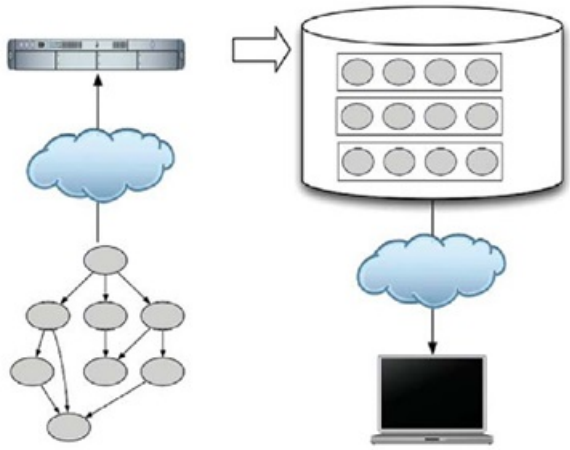
این روش برای بایگانی سایت شرکت ها یا سازمانی و بایگانی که در مقیاس کوچکی توسعه یافته، توصیه شده است. بسته به استفاده از این بایگانی، صحت مبحث باید به دقت مورد بررسی قرار گیرد مخصوصاً برای بایگانی سازمانی. برای بایگانی با مقیاس کوچک، تعادل بین سختی برای سازماندهی مداوم مجموع فایل‌ها وسادگی دسترسی ایجاد شده توسط این رویکرد باید براساس «مورد به مورد» ارزیابی گردد. برای بایگان‌های وب با مقیاس بزرگ و متوسط از این روش باید اجتناب شود.

ابزارها

این راهبرد برای بایگانی وب در مقیاس کوچک ومتوسط ساده ترین رویکرد است که با بسیاری از ابزار های موجود مانند HTTrack قابل دسترس است.

٤-٣-٢-بایگانی‌های مبتنی برخدمت وب[1]

با وجود تقاضاهای بیشتر، این گزینه، تطبیق بهتری را برای نام گذاری منبع وساختار اسناد ایجاد می‌کند (شکل ٦-١). همچنین، برای اجتناب از محدودیت‌های اندازه نظام فایل که برای بایگانی وب در مقیاس بزرگ بحرانی است، مجاز می‌کند.



شکل ۱-٦. مدل مبتنی بر خدمت وب.سایت اصلی مورد خزش قرار می گیرد و پاسخ هادر مخزن بدون تغییر ذخیره می شوند(فایل های)که اجتناب از پیمان نامه مسیریابی نام گذاری فایل های نظام و تغییر ساختار پیوند را اجازه می دهند. دستیابی نیازمند یک سرور وب است که محتوا را در مخازن واکشی کند و آن را به عنوان یک پاسخ به کاربر بفرستد.

1. Web-served archieve

شکل 1- 6. مدل مبتنی بر خدمت وب. سایت اصلی مورد خزش قرار می گیرد و پاسخ ها در مخزن بدون تغییر ذخیره می شوند (فایل های) که اجتناب از پیمان نامه مسیریابی نام گذاری فایل های نظام و تغییر ساختار پیوند را اجازه می دهند. دست یابی نیازمند یک سرور وب است که محتوا را در مخازن واکشی کند و آن را به عنوان یک پاسخ به کاربر بفرستد.

ص: 37

1- Web-served archieve

توصیف

این روش، بر اساس بایگانی پاسخ است (در مقایسه با اولی که بر اساس بایگانی فایل است). پاسخ ها از سرور اصلی (منبع) بدون تغییر در فایل های مخزن WARC (1) ذخیره می شوند که اجازه می دهد تا پشت سر آخرین تا کاربر بایگانی با یک سرور HTTP خدمت رسانی را انجام دهد.

پیشینه های یک فایل WARC (فایل آرشیوی مبتنی بر خدمت) رشته ای از فایل های وب را ثبت می کند هر صفحه به وسیله یک (برچسب) سرآمد که به طور مختصر محتوای حاصل شده و طولش توصیف کند، پیشی می گیرد. به علاوه، محتوای اولیه ثبت شده و WARC (فایل آرشیوی مبتنی بر خدمت) محتوای ثانویه را نیز در بر می گیرد مانند ابر داده و نقل و انتقالات فایل اصلی (منبع). اندازه یک فایل WARC (فایل آرشیوی مبتنی بر خدمت) می تواند تا حدود صدها مگابایت فرق داشته باشد. هر پیشینه یک نقطه شروع دارد که دسترسی مستقیم به پیشینه های انفرادی را بدون بارگیری و تجزیه تمام فایل های WARC (فایل آرشیوی مبتنی بر خدمت) انجام می دهد. نقاط شروع پیشینه های انفرادی در یک نمایه مرتب شده توسط URL ذخیره می شوند. از این رو، به سرعت برای پیشینه های انفرادی که بر اساس URL شان خارج از یک مجموع فایل WARC (فایل آرشیوی مبتنی بر خدمت) هستند، استخراج می شوند؛ که برای دسترسی در حالت شناوری تطبیق یافته اند. سپس ثبت هایی برای سرور وب تصویب می شوند که آن ها را سرویس گیرنده تهیه کرده است.

نگهداری و حفاظت از طرح نام گذاری شمای اصلی منبع (شامل پارامترهایی در صفحه های پویا)، شناوری در سایت را مجاز می کند همان طور که خزش شده است کاربر بایگانی می تواند از تمام سیرهای پیروی شده توسط خزشگر، بار دیگر پیمایش نماید (حرکت کند).

توضیح

مزیت عمده استفاده از مخزن های WARC امکانات غلبه یافتن بر محدودیت سیستم فایل ذخیره سازی در رابطه با اندازه (تعداد کمی از فایل های انفرادی در نهایت در نظام فایل بایگانی ذخیره می شوند) و فضای نام (نام گذاری انفرادی فایل های وب می تواند حفظ و نگهداری شود) می باشد. دستیابی به بایگانی اینترنت از طریق Wayback Machin (که دسترسی به tb500 از مجموعه وب را می دهد) نشان می دهد که این رویکرد به نسبت ثابت افزایش داشته است نه مانند دیگران .اشکال این رویکرد این است که دسترسی مستقیم به فایل های ذخیره شده غیر ممکن است. دو لایه اضافی استفاده شده برای دسترسی به محتوا ضروری هستند:

نظام نمایه فایل WARC و سرور وب. این دو لایه، پیچیدگی برجسته ای ندارند، اما برای دستیابی به محیط به اجرا نیاز دارد که می تواند برای راه اندازی و نگهداری در سازمان های کوچک سخت باشد. این میانجی گر می تواند همچنین مشکلاتی را برای انتقال محتوا افزایش دهد چون نیاز دارد که ساز و کارهای

1- نسخه جدیدتر این نوع چارچوب در مدت هاست توسط بایگانی اینترنت استفاده شده است و بوسیله کنسرسیوم محافظت جهانی از اینترنت International Internet Preservation Consortium (IIPC) استاندارد شده و در ایزو ثبت نام گردیده است.

پیونـد به طور مناسب از محیط وب زنـده به محیط بایگانی، مسـیردهی نماینـد (فرض می کنیم که پیوندهای اصـلی بدون تغییر در بایگانی نگهداری می شوند که این مزیت مهم این روش است). این کار می تواند در سطح نمایش صفحه و در سطح نماینده انجام شود.

گزینه نخست در افزودن به صـفحه فرسـتاده شـده به مرورگر کاربر بایگانی را در بر می گیرد که متن سـند این کار را خواهـد کرد، در خزش تفسیر مجدد پیوندها در صفحه ای به نقطه ای در بایگانی انجام خواهد شد (یا آن ها را در یک شکل مرتبط تفسیر می کند) بایگانی اینترنت برای مثال این کار را با پیروی از کد جاوا-اسکریپت (1) هر صفحه فرستاده شده برای کاربران پیوست شده انجام می دهد.

">>SCRIPT language="Javascript<

<!-

FILE ARCHIVED ON 20050308085053 AND RETRIEVED FROM THE//

.INTERNET ARCHIVE ON 20060514055212//

JAVASCRIPT APPENDED BY WAYBACK MACHINE, COPYRIGHT//

.INTERNET ARCHIVE

ALL OTHER CONTENT MAY ALSO BE PROTECTED BY COPYRIGHT//

.U.S.C(17

)).SECTION 108(a)(3//

/";var sWayBackCGI ="http://web.archive.org/web/20050308085053

)function xLateUrl(aCollection, sProp

var i = 0( )

++)for(i = 0; i < aCollection.length; i

if (aCollection[i][sProp].indexOf(``mailto:") == -1

)aCollection[i][sProp].indexOf(``javascript:") == -1

];aCollection[i][sProp] = sWayBackCGI + aCollection[i][sProp

var i = 0 }

++)for(i = 0; i < aCollection.length; i

if (aCollection[i][sProp].indexOf(‘‘mailto:") == -1

)aCollection[i][sProp].indexOf(‘javascript:") == -1

];aCollection[i][sProp] = sWayBackCGI + aCollection[i][sProp

var i = 0 }

++)for(i = 0; i < aCollection.length; i

if (aCollection[i][sProp].indexOf(mailto:") == -1

)aCollection[i][sProp].indexOf(‘javascript:") == -1

];aCollection[i][sProp] = sWayBackCGI + aCollection[i][sProp

}

“);if (document.links) xLateUrl(document.links, “href

Web Archiving: Issues and Methods 35 1

“);if (document.images) xLateUr (document.images, “src

“);if (document.embeds) xLateUrl(document.embeds, "src

+if (document.body document.body.background(document.body.background = sWayBackCGI

سند. بدنه. زمینه

//->

>SCRIPT</

>HTML</

مشکل این روش این است که برخی پیوندها (جاسازی شده در متن سند) (1) تفسیر نخواهد شد و از این رو، در نقطه ای در وبگاه منبع می مانند. در برخی موارد، تفسیر کد صفحه، برخی رفتارها را فعال می کند، مانند تغییر مسیر حتی قبل از اینکه کد پیوست شده به عنوان مرورگر جدید تفسیر گردد. منتظر نباشید تا سند کامل برای تفسیر به دست آید و آن را نمایش دهد.

با استفاده از نماینده ای که تمام درخواست ها از مرورگر کاربر به بایگانی تغییر مسیر می دهد، بسیار مؤثرتر و کار آمد می باشد همانطور که مسیر دهی بعد از تفسیر پیوند رخ می دهد توسط تعامل کاربر (با کلیک کردن) و مرور گر انجام می شود و کد را برای تولید درخواست مناسب تفسیر می کند (HTML سند متن سرویس گیرنده جانبی فرمت های دیگر) این کارآمدترین به عنوان ظرفیت مرورگر اصلی برای تفسیر مجموعه کدهای استاندارد برای کاری که معمولاً بر روی وب انجام می شود. این رویکرد، نیاز به ایجاد یک نماینده دارد که به یگانی تغییر مسیر دهد و به یک پارامتر که یک مرورگر از آن استفاده کند تا بتواند تقاضاهای زیادی برای یک محیط بایگانی پیوسته داشته باشد. استفاده از اتصال مرورگر برای مدیریت انتقال شکل باز به محیط نماینده می تواند این کار را برای کاربران نهایی آسان سازد.

استفادهٔ ترجیحی

این روش برای بایگانی مقیاس بزرگ و متوسط و همچنین بایگانی های کوچک تر مناسب است که در زمینهٔ حفاظت از صحت محتوا

هستند. چون این روش ها ، پاسخ ها را از سرور منبع همانطور که از سرویس گیرنده دریافت شده است ذخیره می کنند بدون تغییر و تبدیل. در واقع، این روش پایداری

1- ساختار الگو گونه برای نمایش ترتیب حوادث :Scipt

بیشتری را از روش های دیگر فراهم می کند. چون بر سازماندهی فایل محلی بستگی دارد، همچنین مناسب برای انتقال و همچنین تحویل محتواست.

ابزارها

این روش، به یک زیر ساخت دسترسی و به یک خزنده بایگانی (مانند HERITRIX) و یک نظام نمایه سازی برای فایل های WARC نیاز دارد. IIPC زنجیره کاملی از ابزارهایی دارد که برای ایجاد این توابع عملیاتی توسعه یافته است.

3-3-4-بایگانی غیروب

توصیف

در این رویکرد که در شکل 7-1 ترسیم شده است، اسنادی که بر روی وب بوده و از متن ابر متن استخراج شده و در یک سبک متفاوت برای دسترسی منطقی و /یا فرمت سازمان دهی مجدد می شوند.

این می تواند موردی باشد که وقتی یک مجموعه اسناد از وب برداشته می شوند، از دسترسی منطقی مبتنی بر پیوند به یک پیوند مبتنی بر فهرست دوباره سازماندهی می شوند.

عکس

بیشـــتری را از روش‌های دیگر فراهم می‌کند. چون بر ســـازماندهی فایل محلی بستگی دارد، همچنین مناسب برای انتقال و همچنین تحویل محتواست.

ابزارها

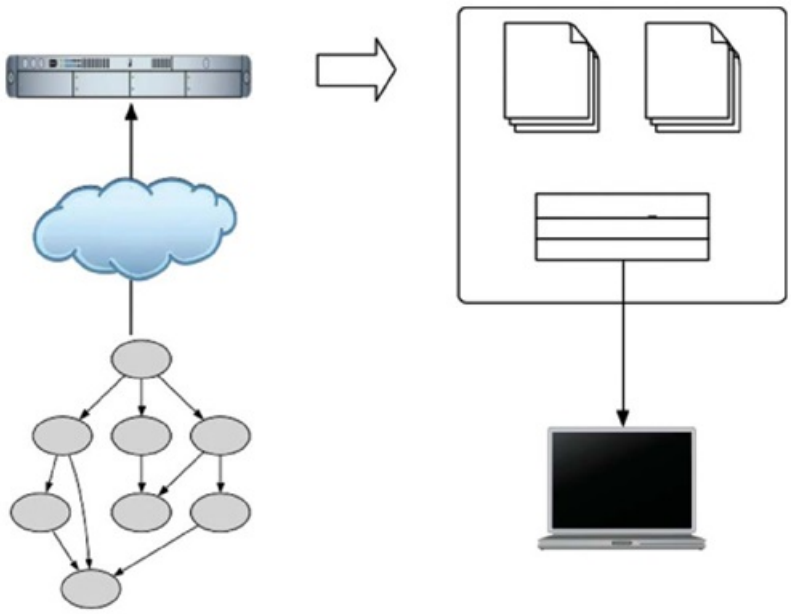
این روش، به یک زیر ساخت دسترسی و به یک خزنده بایگانی (مانند HERITRIX) و یک نظام نمایه ســـازی برای فایل‌های WARC نیاز دارد. IIPC زنجیره کاملی از ابزارهایی دارد که برای ایجاد این توابع عملیاتی توسعه یافته است.

٤-٣-٣- بایگانی غیروب

توصیف

در این رویکرد که در شکل ۷-۱ ترسیم شده است، اسنادی که بر روی وب بوده واز متن ابر متن استخراج شده و در یک سبک متفاوت برای دسترسی منطقی و/یا فرمت سازمان‌دهی مجدد می‌شوند.

این می‌تواند موردی باشد که وقتی یک مجموعه اسناد از وب برداشته می‌شوند، از دسترسی منطقی مبتنی بر پیوند به یک پیوند مبتنی بر فهرست دوباره سازماندهی می‌شوند.



شکل۱-۷. اسناد از سایت اصلی در بایگانی سازماندهی مجدد می‌شوند، ساختار غیر وبی را پیروی می‌کنند، به‌طور موردی از فهرستی که دسترسی به اسناد انفرادی را فراهم می‌کند، استفاده می کند.

شکل 1- 7. اسـناد از سایت اصلی در بایگاه سازماندهی مجدد می شوند، ساختار غیر وبی را پیروی می کنند، به طور موردی از فهرستی که دسترسی به اسناد انفرادی را فراهم می کند، استفاده می کند.

ص: 41

همچنین، این موردی است که وقتی یک صـفحه یا حتی یک وبگاه کامل به فرمت PDF تغییر شکل یافته است. (Adobe Acrobat (1 این توابع عملیاتی را دارد و می تواند تمام یک وبگاه را در یک سـند واحد PDF تغییر شـکل دهد. در این مورد سند به طور مجازی چاپ شده است که یک انتقال بی حرکت و یک صفحه یاد داشت مانند سازماندهی را شامل می شود، حتی اگر پیوند دهی بتواند هنوز با استفاده از شمای داخلی نام گذاری مناسب کار کند.

توضیح

این رویکرد اساساً برای دریافت موضوع هایی که در اصل ایجاد شده اند، به طور مستقل از وب سازماندهی شده اند، ایجاد شده است. این مورد برای مجموعه هـای بزرگی از کتـاب های رقومی به عنوان نمونه، مقاله ها، موزیک، و ویـدئوهای ساخته شـده که بر روی وب موجود هستند، ولی سازماندهی اصلی شان ابرمتنی نمی باشد چون مبتنی بر کاتالوگ هستند.

این رویکرد می توانـد در این مورد برای چسبیـدن به معماری اطلاعات اصـلی و بایگانی این مجموعه ها با یکـدیگر با کاتالوگ های ادغام شده در کاتالوگ بایگانی، مرجع باشد. فرض شده است که محتوای ابر متن بدون ربط پنداشته شده است و می تواند منتشر شود.

برای مثـال این مورد در طرح انبارهای الکترونیکی kb در نیوزیلنـد جایی که انتشارات علمی الزویر (2) اجرا شـده است که در یک نظام مبتنی بر فهرست بایگانی شده است. حقیقت این است که الزویر به این مطالب دسترسـی داشته است و به عنوان یک مجموعه کمکی برای محتوای خودش مورد بررسی قرار گرفته و به عنوان انتشار علمی سنتی ساخت یافته است.

مرجع

اشاره

این روش برای مجموعه های محتوا و نه ساخت یافته در حالت وب دلالت می کند.

4-3-4خلاصه

جدول 2-1- انواع گونـاگونی از بایگـانی هـای وب، موارد مرجـح هـای مورد اسـتفاده، ابزارهای و مزیت های و عـدم مزیت ها را به طور خلاصه شرح داده است.

1- نرم افزاری از محصول شرکت Adobe برای ساخت و تهیه فایل های (PDF)

2- Elsevier

عکس

جدول ۲-۱. خلاصه ای از انواع بایگانی وب

نوع‌بایگانی	نظام فایل محلی	مبتنی بر خدمت	غیروب
توصیف	تمام پیوند به پیوندهای مرتبط(ربطی) تبدیل شده‌اند و شناوری ابر متن، به‌طور قطع بر روی نظامفایل محلی انجام شده است.	سرور یک وب، برای دسترسی از طریق هر سندی که به‌کار رفته است نصب شده است و شناوری ابر متن به بایگانی اصلی بسته شده است.	اسناد از محتوای ابرمتن اصلی استخراج شده و مجددا برای یک منطق مختلف سازماندگی شده‌اند.
استفاده‌مرجع	بایگانی سایت منفردو بایگانی در مقیاس متوسط وکوچک	بایگانی در مقیاس کوچک و متوسط	بایگانی مجموعه‌های خاص (غیروب)
ابزارها	کپی کننده وبگاه (مانند HTTrack)	خزنده بایگانی (مانند Heritrix) و نظام نمایه‌سازی برای فایل‌های WARC	بستگی به ساختار محتوای نهایی دارد.
مزیت‌ها	برای اجراساده است	صحت، قابلیت مقیاس پذیری	قادر به ایجاد یکپارچگی در فهرست‌های سنتی یاسایر معماری‌های اطلاعاتی محلی
معایب	افزایش مقیاس ندارد. نیاز به نامگذاری مجدد و محدود شدن سازماندهی مجدد محتوا برای شناوری ابر متن‌ها است. نیاز به مدیریت در سطح نظام فایل در مجموعه بایگانی و نگارش موارد دارد.	اجرا در غیاب نرم افزار یکپارچه‌سازی سخت می‌باشد (این کار ممکن است در آینده تغییر کند).	فقدان ساختار ابر متن. فقط می‌تواند برای اسناد مجزا و غیروب به‌کار برده شود.

٤-٤-کیفیت و تمامیت (کامل بودن)

به‌طورکلی، کیفیت می‌تواند دریک حالت وظیفه‌ای (متناسب برای استفاده خاص) یا در یک حالت هدفمند (منطبق با ویژگی‌های اندازه‌گیری) تعریف شود. اصطلاح کیفیت، برای مجموعه فرهنگی در مضمون‌های مختلف به‌کار برده شده است. شخص می‌تواند از آن برای کنترل وضعیت حفاظت، کامل نمودن موارد یا مجموعه‌ها، سطح محتواهای هوشمند و علمی و غیره استفاده شود. در هر مورد، آن با مقیاس مطلوب کامل در یک ناحیه خاص و (حفاظت فیزیکی، پوشش یک قلمرو، صحت انتخاب) مرتبط می‌باشد.

برای بایگانی‌های وب همان‌طور که دیده شده، بیشترین نقایص ناشی از سختی برای جمع‌آوری محتوا از طریق واسط HTTP (بخش قبلی در مورد بایگانی سرویس گیرنده جانبی را ببینید) و سختی تحویل در یک حالت منسجم از نتایج محتوا (بخش «سازماندهی و ذخیره‌سازی» را ببینید) می‌باشد.

از این‌رو، کیفیت بایگانی وب در این فصل به‌عنوان ۱) تکامل مطالب (فایل‌های پیوندی) بایگانی شده درون فضای پیرامون هدف، و ۲) توانایی برای تحویل شکل اصلی سایت، به‌ویژه در رابطه با هدایت و تعامل با کاربران (میزانس، ۲۰۰۵) مورد بررسی قرار گرفته است.

جدول 2-1. خلاصه ای از انواع بایگانی وب

4-4-کیفیت و تمامیت (کامل بودن)

به طور کلی، کیفیت می تواند در یک حالت وظیفه ای (متناسب برای استفاده خاص) یا در یک حالت هدفمند (منطبق با ویژگی های

اندازه گیری) تعریف شود. اصـطلاح کیفیت، برای مجموعه فرهنگی در مضـمون های مختلف به کار برده شده است. شـخص می تواند از آن برای کنترل وضـعیت حفاظت، کامل نمودن موارد یا مجموعه ها، سـطح محتواهای هوشمند و علمی و غیره استفاده شود. در هر مورد، آن با مقیاس مطلوب کامل در یک ناحیه خاص و (حفاظت فیزیکی پوشش یک قلمرو، صحت انتخاب) مرتبط می باشد.

برای بایگانی های وب همانطور که دیده شده، بیشترین نقایص ناشی از سختی برای جمع آوری محتوا از طریق واسط HTTP (بخش قبلی در مورد بایگانی سـرویس گیرنـده جانبی را ببینیـد) و سـختی تحویل در یک حالت منسـجم از نتایـج محتوا (بخش «سازماندهی و ذخیره سازی» را ببینید) می باشد.

از این رو، کیفیت بایگـانی وب در این فصل به عنوان 1) تکامل مطالب (فایل های پیونـدی) بایگانی شده درون فضای پیرامون هدف و 2) توانایی برای تحویل شکل اصلی سایت به ویژه در رابطه با هدایت و تعامل با کاربران (میزانس، 2005) مورد بررسی قرار گرفته است.

ص: 43

تکامل می توانـد به طور افقی با تعدادی از نقاط ورودی مربوط که در فضای پیرامون طراحی شده، با نمایش هندسـی ارزیابی شود و از نظر برنامه کاربردی عمودی با تعـدادی از گره های پیونـدی می توان که از این نقطه ورودی دریافت شده اند اندازه گیری می شود. معمولا نقاط ورودی صـفحه های اصـلی سایت هستند و پیوندها می توانند کاربران را به هر یک از نقاط ورودی هدایت کنند (سایت دیگر) یا به عناصر همان سایت بفرستند. این مورد برای بایگانی مبتنی بر سایت است.

به هر حال، در برخی موارد، گنجایش برنامه کاربردی عمودی برای جا سازی عناصـر(برای مثال تصاویر همراه با یک صـفحه) محدود شده است و مجموعه فقـط به طور افقی بـا نادیـده گرفتن سـطح سایت، سازمانـدهی شـده است. برای مثال این مورد برای خزش موضوع های خالص در جایی که صفحه ها را در بر نمی گیرد بر اساس تعلق شان به سایت می باشد، اما فقط بر روی ارتباط شان با موضوع است.

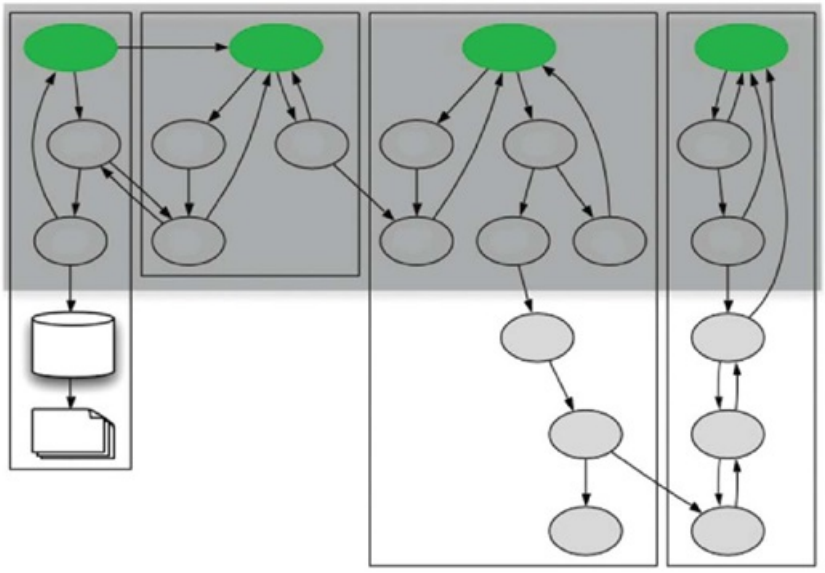
به طور مطلوب، بایگانی وب باید به طور عمودی همچنین افقی کامل شود. اما در عمل، برای دسترسـی سـخت است و اولویت هایی باید قرار داده شوند. بایگانی زمانی که کامل شدن افقی به کامل شدن عمودی ترجیح داده شود، «بسیط» نامیده می شود.

عکس

تکامل می‌تواند به‌طور افقی با تعدادی از نقاط ورودی مربوط که در فضای پیرامون طراحی شده، با نمایش هندسی، ارزیابی شود واز نظر برنامه کاربردی عمودی، با تعدادی از گره‌های پیوندی می‌توان که از این نقطه ورودی دریافت شده‌اند، اندازه‌گیری می‌شود. معمولاً، نقاط ورودی، صفحه‌های اصلی سایت هستند و پیوندها می‌توانند کاربران را به هریک از نقاط ورودی هدایت کنند (سایت دیگر) یا به عناصر همان سایت بفرستند. این مورد برای بایگانی مبتنی بر سایت است.

بهر حال، در برخی موارد، گنجایش برنامه کاربردی عمودی برای جا سازی عناصر(برای مثال تصاویر همراه با یک صفحه) محدود شده است ومجموعه فقط به‌طورافقی با نادیده گرفتن سطح سایت، سازماندهی شده است. برای مثال، این مورد برای خزش موضوع‌های خالص در جایی که صفحه‌ها را در بر نمی‌گیرد براساس تعلقشان به‌سایت می‌باشد، اما فقط بر روی ارتباطشان با موضوع است.

به‌طور مطلوب، بایگانی وب باید به‌طور عمودی همچنین افقی کامل شود. اما در عمل، برای دسترسی سخت است و اولویت‌هایی باید قرار داده شوند. بایگانی زمانی که کامل شدن افقی به کامل شدن عمودی ترجیح داده شود، «بسیط» نامیده می‌شود.



شکل ۱-۸ مجموعه های بسیط شامل بیشتر سایت‌هاست ولی فقط در سطح سطحی بایگانی می‌شود. فقط محتوا در ناحیه سایه‌دار بایگانی می‌شود. صفحه‌ها عمیق در سلسله مراتب (سی ٦، سی۷، سی۸، دی۳، دی۴، دی۵) و نیز محتوایی که در پایگاه داده پنهان شده (وب پنهان) تصرف خواهند شد.

شکل 1-8. مجموعه های بسیط شامل بیشتر سایت هاست ولی فقط در سطحی بایگانی می شود. فقط محتوا در ناحیه سایه دار بایگانی می شود. صفحه ها عمیق در سلسله مراتب (سی 6، سی 7، سی 8، دی 3، دی 3، دی 5) و نیز محتوایی که پایگاه داده پنهان شده (وب پنهان) تصرف خواهند شد.

ص: 44

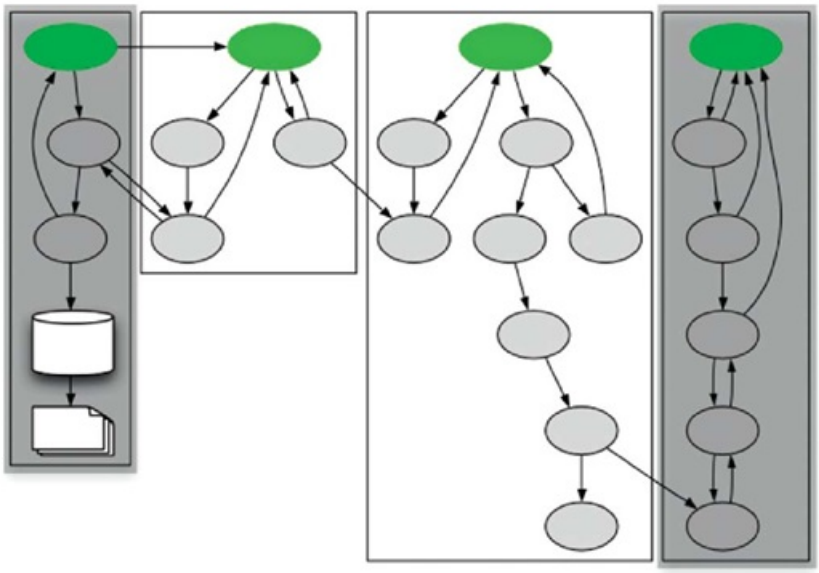
برای مثال، این مورد برای مجموعه بایگانی اینترنت است که به وسیله الکسا (همچنین برنر، 1997؛کیمپتون و همکاران،2006) ارائه شده است، خزشگر الکسا یک رویکرد خزش سطح اول را استفاده می کند و عمق خزش را برای یک سایت بر طبق ترافیک اندازه گیری شده برای این سایت وفق می دهد.

بر عکس، بایگانی «متمرکز» نامیده می شود زمانی است که تکامل عمودی به کامل شدن افقی ترجیح داده می شود (شکل 9- 1 را ببینید).

عکس

برای مثال، این مورد برای مجموعه بایگانی اینترنت است که به وسیله الکسا (همچنین برنر، ۱۹۹۷؛ کیمپتون و همکاران، ۲۰۰۶) ارائه شده است، خزشگر الکسا یک رویکرد خزش سطح اول را استفاده می‌کند و عمق خزش را برای یک سایت بر طبق ترافیک اندازه‌گیری شده برای این سایت وفق می‌دهد.

برعکس، بایگانی «متمرکز» نامیده می‌شود زمانی است که تکامل عمودی به کامل شدن افقی ترجیح داده می‌شود (شکل ۹-۱ را ببینید).



شکل ۱-۹. بایگانی بسیط سایت‌های کمتری مورد خزش قرار می‌گیرند ولی خزش در عمق انجام می‌شود. ولی سایت الف و دی مورد بایگانی قرار می‌گیرند ولی در حالت کلی شامل بخش وب پنهان سایت الف.

برای مثال، در این مورد، وقتی اولویت نخست ـ سایت برای خزش گرها استفاده شده یا وقتی که یک بازبینی دستی، در جایی که مورد نیاز است، بر روی آن انجام می‌شود، بایگانی تکمیلی انجام می‌شود. بایگانی متمرکز حتی متقاضی بیشتری برای سایت‌های وب پنهان دارد (همچنین، سایت‌های وب عمیق نامیده شده است) جایی که دسترسی به تمام محتوا با خزش گرها ممکن نیست.

۵ ـ بررسی عمومی مراحل اولیه جاری

بایگانی وب می‌تواند در چندین روش طبقه‌بندی شود. در این بخش، روش‌های مهم را بررسی خواهیم کرد و این فرصت را برای ارائه برخی مراحل اولیه بایگانی وب و مقایسه با رویکردهای گوناگون نشان می‌دهیم.

شکل 1-9. بایگانی بسیط سایت های کمتری مورد خزش قرار می گیرند ولی خزش در عمق انجام می شود. ولی سایت الف و دی مورد بایگانی قرار می گیرند ولی در حالت کلی شامل بخش وب پنهان سایت الف.

برای مثال، در این مورد، وقتی اولویت نخست – سایت برای خزش گرها استفاده شده یا وقتی که یک بازبینی دستی، در جایی که مورد نیاز است، بر روی آن انجام می شود، بایگانی تکمیلی انجام می شود. بایگانی متمرکز حتی متقاضی بیشتری برای سایت های وب پنهان دارد (همچنین، سایت های وب عمیق نامیده شده است) جایی که دسترسی به تمام محتوا با خزش گرها ممکن نیست.

5-بررسی عمومی مراحل اولیه جاری

بایگـانی وب می توانـد در چنـدین روش طبقه بنـدی شود. در این بخش روش های مهم را بررسـی خواهیم کرد و این فرصت را برای ارائه برخی مراحل اولیه بایگانی وب و مقایسه با رویکردهای گوناگون نشان می دهیم.

ص: 45

1-5-بازیگران بایگانی

نوع ایجاد سازماندهی و میزبانی از بایگانی، نخستین معیار برای طبقه بندی بایگانی های وب می باشد. برخی دسترسی عمومی به مجموعه های شان را فراهم می کنند (WA عمومی) برخی این کار را نمی کنند WA خصوصی (یا مخفی).

در میان بایگانی های عمومی وب، برخی دسترسی پیوسته را فراهم می کنند برخی دسترسی به سایت را در اتاق های مطالعه فراهم می کنند (بایگانی وب عمومی پیوسته و بایگانی وب عمومی غیر پیوسته). همچنین برخی و در بیشتر موارد در ابتدا مجموعه غیر رقومی را مدیریت می کنند (بایگانی وب هیبرید) و بالاخره برخی در حالت سرمایه گذاری یا بدون منفعت (بایگانی وب غیر تجاری) با در نظر گرفتن اینکه برخی شرکت های تجاری هستند (بایگانی وب تجاری).

مؤسسه های میراث سنتی (کتابخانه ها، بایگانی ها، موزه ها) که مجموعه هایشان را برای وب گسترش داده اند، با یکدیگر بیشترین قسمت مقوله بایگانی وب هیبرید عمومی را تشکیل داده اند. کتابخانه های ملی چندین کشور متعلق به این مقوله هستند (سوئد و استرالیا که نخستین بار به 1996 برمی گردد و اکنون کشورهای زیادی هستند) (1) .

بایگانی های ملی، منطقه ای و شهری همچنین شروع به بایگانی وبگاه های مجاز دولتی و محلی نموده اند (2) . سازمانی بر روی شکل های جدید صنعتی مانند v2 مستقر در روتردام هلند کار می کند که در حال یکپارچه سازی شبکه در یک انعکاس و عملکردی عمومی برای حفاظت از رسانه ناپایدار می باشد (فوکونیر و فرومه 2004) (3).

تمام این بایگانی ها می توانند به عنوان آرشیو وب عمومی هیبرید رده بندی شوند. همان طور که محتوای وب را در یک مضمون بزرگ تر از مجموعه ها مجتمع می کنند. بیشتر آن ها فقط دسترسی غیر پیوسته برای اسناد را در حال حاضر فراهم می کنند.

از میان آن ها، کتابخانه اسکندریه در مصر یکی از معدود دسترسی های پیوسته برای مجموعه بایگانی وب است (معکوس نمودن بایگانی اینترنتی) و یک مثال از بایگانی وب عمومی هیبرید غیر تجاری پیوسته می باشد.

نفوذپذیری اینترنت همچنین ضرورت برخی از انواع جدید سازمان های بایگانی را مجاز می داند که فقط مجموعه رقومی و تهیه دسترسی پیوسته را نگه می دارد که به عنوان بایگانی وب پیوسته غیرتجاری عمومی، رده بندی خواهند شد.

بایگانی اینترنت، مثال مهمی در این مقوله می باشد (فصل 9 را ببینید) (کیمپتون و همکارانش، 2006).

1- با پیروی از آن ها، چندین کتابخانه ملی بایگانی وب را شروع کرده و برنامه هایی را اجرا می کنند (این سیاهه جامع نیست): در اروپا: فنلاند، دانمارک، نروژ، ایسلند فرانسه، جمهوری چک، اسلوونی، ایتالیا و یونان؛ در آسیا ژاپن چین و سنگاپورو کتابخانه کنگره در آمریکا با پیروی از آن ها، چندین کتابخانه ملی بایگانی وب را شروع کرده و برنامه هایی را اجرا می کنند (این سیاهه جامع نیست): در اروپا: فنلاند، دانمارک، نروژ، ایسلند فرانسه، جمهوری چک، اسلوونی، ایتالیا و یونان؛ در آسیا ژاپن چین و سنگاپورو کتابخانه کنگره در آمریکا.

2- آرشیوهای ملی استرالیا (آرشیوهای ملی استرالیا، 2001) بریتانیا (براون، 2006)؛ کانادا ایالات متحده آمریکا (کارلین، 2006) بایگانی وب نظام مند را شروع کرده اند. هم چنین طرح شهر Antwerp DAVID را ببینید (بودرس و آینده، 2002).

برخی شـرکت های تجاری مجموعه های بزرگی از محتوای وب عمومی را بایگانی می کننـد مانند گوگل با نهانگاهش (1) و بایگانی هانزو (2) . مثال هایی از بایگانی وب تجاری عمومی پیوسته هستند.

سرانجام، بسـیاری از سازمان ها، بایگانی وب داخلی را برای اهداف شخصـی توسعه می دهند که به عنوان بایگانی وب خصوصی (مخفی) طبقه بندی خواهند شد. منحصـر به فرد بودن نوع دسترسی (پیوسته یا غیر پیوسته) و همچنین وضعیت تجاری در اینجا ارتباط کمتری دارد چون این بایگانی ها فقط برای استفاده خصوصی هستند.

2-5-حوزه (دامنه)

روش مفید دیگر برای دسـته بندی بایگانی های وب، بررسی حوزه ای است که آن ها اتخاذ می کنند. بایگانی های وب می توانند هر یک، سایت، سر فصل ها یا متمرکز بر حوزه باشند.

5-2-1-بایگانی مرکزی سایت

این نوع بایگانی، بر روی یک سایت مشـخص متمرکز شـده است که تقریباً به وسـیله و برای ایجاد کننده سایت اجرا شده است. این حوزه بندی ، از این رو، تقریباً برای بایگانی وب خصوصـی اسـتفاده شده است. مثلاً بسـیاری از شـرکت ها، مسـئول تمام محتواهایی هستند که منتشـر می کنند و باید مطمئن باشند که می توانند به روش های قدیمی تر سایت وبنوشت ها مراجعه کنند. این نوع بایگانی ترجیحاً از کپی کنندگان سایت و برخی تهیه کنندگان خدمات اینترنتی استفاده می کند که برای این نوع بایگانی داخلی مناسب، به وجود آمده اند (3)

5-2-2-بایگانی مرکزی عنوان

بایگانی های وب، عمومی و عمومی تر شده اند، اغلب به وسیله نیازهای پژوهشی مستقیم اجرا می شوند تا موقعی که کار بر روی یک فیلد مشـخص و انعکاس برروی وب انجام می شود، دانش پژوهان زیادی با طبیعت بی دوام انتشار وب روبه رو می شونـد، جایی که طول عمر وبگاه برای بازبینی علمی (تکذیب نیازمند دستیابی به همان داده است) و همچنین برای ارجاع با دوام، نامناسب می باشد.

به این دلیل است که طرح های مختلف اغلب در کتابخانه های دانشگاه میزبانی می شوند و تحت حفاظت مطالب اولیه برای پژوهش قرار می گیرند، مانند بایگانی رقومی برای مطالعات چینی در دانشـگاه هایدلبرگ (4) در آلمان یا آرچیپل (5) برای تحلیل سایت های سیاسی هلندی در دانشگاه گرونینگن (6) در هلند

1- Cashe

2- Hanzo

3- برای نمونه hanzoarchives.com را ببینید.

4- Digital Archive for Chinese Studies (DACHS) at Heidelberg University

5- Archipol

6- Groningen

(ورمن (1) و همکاران، 2002) این پروژه ها نه تنها جهت دهی یک عنوان را به اشتراک می گذارند، بلکه از

یک شبکه آگاهی دهنده نیز استفاده می کنند.

از طرفی دیگر، پژوهشگران که تغذیه های صحیح و به روز

طرح هـای متمرکز دیگری در کتابخانه ها از طریق جست و جوی فعال و بایگانی وبگاه گزینشـی (2) انجام شده است و مانند طرح مینروا (3) از کتابخـانه کنگره (اشـنایدر و همکاران 2003) یـا بایگانی وب انتخابات فرانسه که توسط کتابشناسـی ملی فرانسه ایجاد شـده اسـت (میزانس ، 2005). در مقـایسه بـا رویکرد قبلی مبتنی بر عنوان متمرکز، کشف سـایت ها به طور طبیعی به عنوان محصول فرعی از فعالیت پژوهشی ایجاد نشده است. و به عنوان یک فعالیت خاص نیاز به بررسی دارد.

بالاخره، برخی طرح ها که با این مقوله مرتبط هسـتند. از خزش عنوان برای کشف و ضـبط محتوا مرتبط با مطلب مشابه استفاده. می کنند (چاکرابارتی (4) و همکاران، 1999؛ برگمارک 2002 (5) ،؛ برگمارک و همکاران، 2002؛ کیواین (6) و همکاران، 2004). کشف و فیلتر کردن خودکـار با اسـتفاده از فن سـنتی خزش در ترکیب ارزیابی سـطح صـفحه در محتوای متنی انجام شـده است که گاهی اوقات با کاویدن ساختارهای پیوندی آمیخته می شود. مجاورت با عنوان، می تواند از مجموعه ای از نوشتجات یا از باز خورد کاربر، فرا گرفته شود. اگر چه محتمل است، این ناحیه باز هم نیاز به پژوهش برای به کار برده شدن در بایگانی دارد.

3-2-5-بایگانی مرکزی حوزه ای

ساخت بایگانی می توانـد بر اساس محل محتوا نیز انجام شود و به این ترتیب، نوع سوم در بایگانی وب حوزه ای را مشـخص می کنـد. در اینجا کلمه «حوزه» (7) در جهت واژه شـبکه یا به وسـیله پسونـد در جهت واژه ملی اسـتفاده شده است که یک معیار ترکیبی برای سایت های هدف از یک کشور خاص می باشد (8) .

نظـام ملی حوزه ای یـک گزینش سـاده و قابل تعقیب قانونی در محتوا را بر اساس نام های دامنه اجازه می دهـد. این حقیقتی است که نام های دامنه حتی برای سطوح بالایی حوزه ای توسط نمایندگی رسمی آی.سی.آن (9) مدیریت شده است، که واقعاً از قوانین در رابطه با نام دادن مشـخصه عملیاتی و سازمان ها پیروی نمی کننـد، ولی بیشتر سـنت ها را پیروی می کننـد (لیو و آلبیتز 1999) (10). همچنین می توانید در مورد تکامل بر روی نام های اینترنتی به (کوهلر، 1999) نگاه کنید. از این رو، شخص می تواند انواع پسوندهای

1- Voerman

2- Electoral Web sites

3- Minerva

4- Chakrabarti

5- Bergmark

6- Qin

7- Domain

8- برای بحث در مورد روش ممکن مرزبندی کردن فضای اینترنت ملی آرویدسون و دیگران (2000)؛ ابایت بول و دیگران (2002)؛ لامپوس و دیگران (2004) را ببینید، برای مطالعاتی در مورد ویژگی های فضای اینترنت ملی بیزایتس و دیگران (2005 الف، 2005ب) و گومس و سیلوا (2003) را ببینید.

9- ICANN

10- Liu and Albitz

عملیـاتی یا عمومی ماننـد Com و (edu و انواع سیسـتم اطلاعـاتی جغرافیایی ماننـد (1) (ch.jp) و انـواعی را در دامنه سـطح نخسـت تشـخیص دهد و که (اغلب دامنه سـطح بالا نامیده می شود). دامنه های سـطح بالای جغرافیایی اغلب بخش های فرعی عملیاتی دارند و (مانند asso.fr gob.mex). که به این معنـاست که دامنه سـطح دوم (2) همچنین، در همان روش مـدیریت خواهد شد. اسـتثناها برای عملکردهایی مانند TLDهـا وجود دارد که بخش هـای فرعی جغرافیـایی دیگر را دارا می باشـند. (به وسـیله ایالت ها). به هر حال، توجه داشـته باشـید که تمام این قسمت ها فضای حوزه اینترنتی است که توسط نمایندگی مدیریت شده اند (3) . هر هویت تحت فرمـان آن ها می تواند یک سـیاست خاص را در رابطه با تخصیص و کنترل فضاهای شان به کار ببرد بنابراین رسیدن به بهره برداری از SLD,TLD برای گزینش بایگانی وب، در هر مورد طبق ارزیابی این سیاست بستگی دارد (مثلاً org,.com.gTLD) به وسلیه تمام انواع سازمان ها استفاده شده اند نه تنها توسط سازمان های تجاری برای com. و سازمان های بدون فایده برای org. چون محدودیتی برای ثبت نام وجود ندارد) به علاوه، برخی هویت ها تحت فرمان مدیریت TLD، سیاستشان را طبق زمان تغییر می دهند.

(org.net با داشـتن محـدودیت ها قبل از سال 1996 اسـتفاده شده اند و frTLD به طور قابـل توجهی محـدودیت ها را در سال 2005 کاهش داده است).

مزیت بزرگ دیگری که در اینجا بایـد ذکر کنیم، آوردن معیارهایی است که بتواند به طور خودکار توسط خزشـگرها آشـکار شوند مانند نام هـای دامنـه پروژه هـای دیگری در واقـع، رویکرد مرکزی دامنـه را اجرا می نماینـد. برخی، بر روی یـک دامنـه عمـومی ماننـد (کروس و و همکاران، 2003؛ کارلین، (2004 و یا edu (لایل، 2004) متمرکز شده اند.

برخی از دامنه هـای ملی ماننـد ایـالت کـالتیورارو (4) در سـال 1997 که توسط کتابخانه سـلطنتی سوئـد ایجاد شـد، اسـتفاده می کنند (آرویدسون 2000) که SeTLD و همچنین صفحه ها سوئدی پیوند یافته از آن و در مانند com. واقع شده است را پوشش می دهد.

3-5-روش های استفاده شده

طرح ها می تواننـد همچنین به طور قابل توجهی نسـبت به رویکرد روش های بررسـی که برای کشف اکتساب و توصـیف محتوا در بر می گیرند و متفاوت باشند. یکی از تفاوت های مهم این است که محدوده ها در سراسر این مراحل، از دست به جای پردازش خودکار استفاده می کنند. اگر چه سادگی ظاهری این تضاد باید به عنوان پردازش خودکار متعادل می شود و در چندین سطح رخ می دهد (ضبط، استفاده از موتور جست و جو برای کشف دسـتی و غیره (میزانس، 2006 ب)، این مسـئله، همچنین می تواند بایگانی وب را بر طبق این تضاد رده بندی نماید که به طور مستقیم اثر شدیدی بر روی قابلیت مقیاس پذیری و

1- ایزو 3166 دو حرف اول نام کشورها را اجازه می دهد به جز uk که باید gb باشد، و همچنین به جز این که اخیراً به سه حرف برای هر منطقه گسترش یافته است مثل حوزه cat در کاتولینا در اسپانیا.

2- SLD

3- در مورد حوزه دولتی و مفاهیم آن میولر(2002) را ببینید.

4- Kulturarw

کیفیت بایگانی ها دارد.

همـان طور که توانستیم پیش بینی کنیم، خود کـار سازی این وظایف، در پایین آوردن فاحش هزینه های هر دستیابی به سایت توانا هسـتند (1) . به طور مطلوب یک اپراتور تنها با اجرای خزش می تواند میلیون ها صـفحه را کشف و بارگذاری کند. نظر به این که نمایه سازی تمام متن کمکی برای یافتن قدرتمنـد قابل مقایسه ای را فراهم میکند که اگر در برخی موارد برای فهرست کردن عالی نباشد، بنابراین می بینیم که بار دیگر اینجا چقـدر خود کارسازی به طور چشـمگیری هزینه ها را کاهش می دهـد، چون می توانـد بر روی یک مقیاس بزرگ برده شود (استک، 2005) (هال گریمسون، 2006).

متأسفانه، خودکار سازی، محدودیت هایی دارد و بررسی دستی باید در برخی موارد انجام شود. برای مثال کشف می تواند به طور دستی با اتوماتیـک انجـام شود. وقتی کـار به طور دستی انجام شود می توانـد یک فعالیت خاص یا یک محصول فرعی از فعالیت های دیگر باشـد مانند DACHS (لچر، 2006) و نمـایش بایگـانی های وب (وئرمان (2) و دیگران، 2002) Achipol این نـوع رویکرد معمولاً ـ برای بایگـانی مرکزی عنوان در انجـام می شود. اگر چه خزش عنوان به طور مؤثر برای کشف مطلب سایت های یا صـفحه های مربوط آزمایش می گردد. ابزارهـای خودکـار می تواننـد به طور مطمئن (نه در این زمان) در مقایسه با یک شـبکه متخصـصان، مرجع هایی را برای بهترین مطالب که از آن مطلع هستند، فراهم نمایند.

به هر حال، فقدان حوزه تخصصی و عدم درک تنها معایب خزش گرها نیستند.این مسئله مورد توجه است که تأخیر برای یافتن سایت های جدیـد مورد نیاز است. زمان بسـیار زیادی برای یافتن سایت های کلی گرفته می شود. وقتی وارد سایت های زودگـذر (بی دوام) می شود، برای مثال مرتبط با یک رویـداد، تاخیری می تواند بسـیار طولانی باشد و آن ها را بایگانی کند این تفاوت به وسـیله (میزانس،2005) با یک مقایسه سایت های کشف شده توسط خزشـگر الکسا و قابلیت دسترسی و روزانه به بایگانی اینترنت و سایت های مرتبط با انتخابات فرانسه در سال 2002 که توسـط گروهی از کتابـداران مرجع انجام شـد و به وسـیله کتابخانه ملی فرانسه بایگانی گردیـد، مورد بررسـی قرار گرفته است.

این بررسی، مزیت های آشکاری را برای گزینش فعالیت دستی در مجموعه های مبتنی بر ربط، برای کشف به موقع و در تمرکز عمیق نشان می دهد.

رده بندی بایگانی وب بر طبق روش بررسـی اش همچنین می توانست در یک شاخه مالی انجام شود. بر تراز پردازش اتوماتیک و دستی دو حالته، شخص می تواند برای مثال نوع منبع استفاده شده برای کشف، تناوب جست و جو، و ضبط، سطح کیفیت بازبینی ایجاد شده و دانه دانه بودن آیتم های بایگانی شده (سایت ها، صفحه ها)، و مانند آن را مورد بررسی قرار می دهد.

1- فیلیپز (2005) تخمین های بسـیار مفیـد و دارای جزئیات در خصوص زمان و هزینه های فرآینـد دستی خودکارسازی سایت ها برای یکی از قدیمی ترین آرشیوهای وب موجود را فراهم کرده است. تخمین های زمانی به صورت ذیل هستند: - تشخیص و انتخاب 30 دقیقه - گردآوری، اطمینان از کیفیت، و موارد آرشیوی: 210 دقیقه - فهرست نویسی 81 دقیقه

2- Voerman

به هر حال، این امر که بیشتر بایگانی های وب به دو مدل عمده تمایل دارند، تمایز اصلی است که آیا انتخاب به طور دستی یا غیر آن انجام شـده است. یکی مـدل های خزش گر های کلی است که معمولاً مبتنی بر مرکز (دامنه های ملی یا دامنه های عمومی) است یا به صورت آزاد (بایگـانی اینترنت)، و دیگری مـدل گزینش فردی تعـداد محـدودی از جست و جوها یا نقاط ورودی است که به طور دستی انجام می شود (معمولاًـ سایت ها). تمایز در بیشتر در رویکرد روش شـناختی آن ها به نـدرت دیـده شـده یا هیـچ یک برای طبقه بنـدی آن ها مورد استفاده قرار نمی گیرد.

6-نتیجه گیری

وب فقط 15 سـال است که وجود دارد و می توان گفت که نگهـداری و حفاظت از حافظه اش، به طور نسبی در مقایسه با رسانه های دیگر از ابتدا آغاز شده بود. (1) اما فقط مراحل لازم اولیه برای حفاظت آن ایجاد شده است. وضـعیت حفاظت جاری، به تعداد بسـیار کمی از مؤسسه ها وابسته است و پوشش زیادی را حاصل نشده است.

نقش ها و مسـئولیت ها برای بیشتر سـهامداران بسـیار واضح و آشـکار نیست و توانایی در حمایت از بسـیاری از مجموعه های شاخص به وجود نیامده است. و ما هنوز در دوره ای هستیم که هیچ گسیختگی فناورانه از زمان آغاز وب، رخ نداده است.

مرورگرهای جاری با یکـدیگر با تعداد محدودی که برنامه های کامپیوتری متصل می توانند فرمت های زیادی را جابه جا کنند که می تواند بر روی وب یافت شود.

اما این موقعیت، تا ابد طول نخواهد کشید و حفاظت از وب با یک چالش جدی روبه رو خواهد شد وقتی تغییرات مهم فناوری در وب رخ می دهد (که ممکن نیست مانند اینها بعداً دیده شود).

بنابراین، دلگرم کننده است که ببینیم که بسـیاری از مؤسـسه های (حفظ) میراث، در بایگانی وب در حال به کارگیری هستند. بررسی اخیر توسط گروه پژوهش کتابخانه (RLG2006) نشان داد که 60 درصد اعضای مورد بررسـی شان ، بایگانی وب را قسمتی از مأموریت خود پنداشته اند (RLG2006) که بسیار دلگرم کننده است.

امیـدواریم که بـازنمون های ایجاد شـده در این فصل، مباحث مهم و روش ها، متفقاً با منطق و دلیل، به آن ها و دیگران برای مشارکت در این تلاش گروهی، کمک خواهد کرد.

1- اکثر رادیو و تلویزیون های دنیا اصلا حفاظت نمی شوند.

منابع

Aarseth, E. J. (1997). Cybertext: perspectives on ergodic literature. Baltimore, MD: Johns Hopkins . 1
University Press

Abiteboul, S., Cobena, G., Masanè s, J., Sedrati, G. (2002). A first experience in archiving the French . 2
Web. Paper presented at the Proceedings of the 6th European Conference on Research and Advanced
Technology for Digital Libraries

Abiteboul, S., Preda, M., Cobena, G. (2003). Adaptive on-line page importance computation. Paper . 3
presented at the Proceedings of the twelfth international conference on World Wide Web

Antoniol, G., Canfora, G., Cimitile, A., De Lucia, A. (1999). Websites: files, programs or database. Paper .4
presented at the 1st International Workshop on Web Site Evolution, Atlanta, USA

Arvidson, A., Persson, K., Mannerheim, J. (2000). The Kulturarw3 project – The Royal Swedish Web . 5
Archiw3e – An example of "complete" collection of web pages. Paper presented at the 66th IFLA –
International Federation of Library Associations and Institutions, Jerusalem

Baeza-Yates, R. Castillo, C. (2005). Characteristics of the Web of Spain. Cybermetrics,9 .6

Baeza-Yates, R., Castillo, C., Efthimiadis, E. (2005a). Characterization of national Web domains .7

Baeza-Yates, R. A., Castillo, C., Marin, M., Rodriguez, A. (2005b). Crawling a country: better strategies .8
than breadth-first for Web page ordering. Paper presented at the WWW 05: Proceedings of the 14th
international conference on World Wide Web, Chiba, Japan

Balayé , S. (1988). La Bibliothè que nationale, des origines a 1800 (Histoire des idé es et critique . 9
litté raire; vol. 262). Genè ve: Droz

Battelle, J. (2005). Google Announces New Index Size, Shifts Focus from . 10
Counting.http://battellemedia.com/archives/001889.php

Benjamin, W. (1963). Das Kunstwerk im Zeitalter seiner technischen .11

Reproduzierbarkeit; drei Studien zur Kunstsoziologie. [Frankfurt am Main]: Suhrkamp .12

Bergman, M. I. K. (2001). The deep Web: Surfacing hidden value. The Journal of Electronic) .13

Publishing, 7(1

ص: 52

Bergmark, D. (2002). Collection synthesis. Paper presented at the 2nd ACM/IEEE-CS joint conference .14
on Digital libraries, Portland, USA

Bergmark, D., Lagoze, C., Sbityakov, A. (2002). Focused crawls, tunneling, and digital libraries. Paper .15
presented at the 6th European Conference on Research and Advanced Technology for Digital Libraries,
Roma, Italy

Berners-Lee, T. Connolly, D. (1995). Hypertext Markup Language - 2.0. RFC,1866 .16

Berners-Lee, T. (1994). Universal Resource Identifiers in WWW, A Unifying Syntax for the Expression .17
of Names and Addresses of Objects on the Network as used in the World- Wide Web. RFC 1630

Berners-Lee, T. (1998). Cool URIs don't change. http://www.w3.org/Provider/ Style/ URI.html .18

and ultimate destiny of the World Wide Web by its inventor (1st pbk. ed.). New York: HarperCollins .19

.BjÖrneborn, L. Ingwersen, P. (2001). Perspective of webometrics .20

Scientometrics, 50(1), 65-82 .21

Bolter, J. D. (2001). Writing space: Computers, hypertext, and the remediation of print (2nd ed.). . 22
Mahwah, NJ: Lawrence Erlbaum Associates

Borgman, C. L. (2000). From Gutenberg to the global information infrastructure: access to information .23
in the networked world (Digital libraries and electronic publishing). Cambridge, MA: MIT

Borgman, C. L. (2003). The Invisible Library: Paradox of the Global Information Infrastructure. Library .24
Trends, 51(4), 652-674

Boudrez, P. Eynde, V. D., Sofie. (2002). Archiving Websites .25

Boufkhad, Y. Viennot, L. (2003). The Observable Web. RR .26

Boyko, A. (2004). Test Bed Taxonomy. IIPC Reports, 16 .27

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2000). Graph . 28
structure in the web. Paper presented at the 9th International World Wide Web Conference (WWW9),
Amsterdam, Netherlands

29. Brown, A. (2006). Archiving the Web: A guide for information management professionals. Library Assn Pub.

30. Brügger, N. (2005). Archiving Websites, general considerations and strategies. A arhus, Denmark: Center for Internet Research

Bruns, A. (2005). Gatewatching: Collaborative online news production (Digital formations, v. 26). . 31
New York: P. Lang

Burner, M. (1997). Crawling towards Eternity Building An Archive of The World Wide Web. New . 32
Architect, 5

Canfora, L. (1989). The vanished library (Hellenistic Culture and Society; 7). Berkeley: University of . 33
California Press

Canfora, L. (1996). Les bibliothè ques anciennes et l'histoire des textes. In M. Baratin, C. Jacob (Eds.), . 34
Le pouvoir des bibliothè ques: la mé moire des livres en Occident. (pp. 338 p). Paris: A. Michel

Carlin, J. W. (2004). Harvest of agency public websites. NARA Bulletin, 2005-02 .35

Castells, M. (1996). The rise of the network society. Malden, MA: Blackwell .36

Castillo, C., Marin, M., Rodriguez, A., Baeza-Yates, R. A. (2004). Scheduling Algorithms for Web . 37
Crawling

Chakrabarti, S. (2002). Mining the Web: discovering knowledge from hypertext data. San Francisco, . 38
CA: Morgan Kaufmann

Chakrabarti, S., Berg, M. V. D., Dom, B. (1999). Focused crawling: A new approach to topic-specific . 39
Web resource discovery. Computer Networks (Amsterdam, Netherlands: 1999), 31, 1623-1640

Chang, K. C.-C., He, B., Li, C., Patel, M., Zhang, Z. (2004). Structured .40

,databases on the web: observations and implications. SIGMOD Record .41

61-70 ,(3)33 .42

Charlesworth, A. (2003). Legal issues relating to the archiving of Internet resources in the UK, EU, USA .43
and Australia

Cho, J., Garcia-Molina, H. (2000). The evolution of the web and implications for an Incremental . 44
Crawler. Paper presented at the Proceedings of the 26th International Conference on Very Large Data Bases

Cho, J., Garcia-Molina, H., Page, L. (1998). Efficient Crawling Through url ordering. Computer . 45
Networks and Isdn Systems, 30, 161-172

46. Christensen-Dalsgaard, B. (2001). Archive experience, not data. Paper presented at the Preserving the Present for the Future – Strategies for the Internet, The Royal Library, Copenhagen, Denmark

47. Crowston, K., Williams, M. (1997). Reproduced and emergent genres of communication

on the World-Wide Web. Paper presented at the 30th Annual Hawaii International Conference on System Sciences (HICSS-30), Wailea, USA

Cruse, P., Eckman, C., Kunze, J. (2003). Web-based government information: Evaluating solutions for .48 capture, curation, and preservation. An Andrew W. Mellon funded initiative of the California Digital Library

Dahn, M. (2000). Counting Angels on a Pinhead: Critically Interpreting Web Size Estimates. Online, .49 January/February, 35-40

Day, M. (2006). The long-term preservation of Web content. In J. Masanè s (Ed.), Web archiving. .50 Berlin Heidelberg New York: Springer

Dikaiakos, M. D. (2004). Intermediary infrastructures for the World Wide web. Computer Networks, .51 45(4), 421-47

Dobra, A., Fienberg, S. E. (2004). How Large Is the WorldWide Web?. In M. Levene, A. Poulovassilis .52 (Eds.), Web dynamics web dynamics - adapting to change in content, size, topology and use. (pp. 23-44). Berlin Heidelberg New York: Springer

Dubberly, H., Forlizzi, J., Hodge, C., Laurel, B., Lyman, P., Meggs, P. B., et al. (2002). Archiving .53 experience design, a virtual roundtable discussion. LOOP: AIGA Journal of Interaction Design Education, Number 6

Dumais, S. T., Cutrell, E., Cadiz, J. J., Jancke, G., Sarin, R., et al. (2003). Stuff I've seen: A system for .54 personal information retrieval and re-use. Toronto, Canada

Egghe, L. (2000). New informetric aspects of the Internet: some reflections - many problems. Journal of .55 Information Science, 26(5), 329-335

Eisenstein, E. L. (1979). The printing press as an agent of change: Communications and cultural . 56 transformations in early modern Europe. Cambridge [Eng.]; New York: Cambridge University Press

(Entlich, R. (2004). Blog Today, Gone Tomorrow? Preservation of Weblogs. RLG DigiNews, 8(4 .57

Eriksen, L. B. IhlstrÖm, C. (2000). Evolution of the web news genre - The slow move beyond the print .58 metaphor. Paper presented at the 33rd Hawaii International Conference on System Sciences (HICSS-33), Hawaii, USA

Estivals, R. (1961). Le dé pôt lé gal sous l'Ancien Ré gime, de 1537 a 1791. Paris: M. Riviè re .59

Estivals, R. (1965). La statistique bibliographique de la France sous la monarchie au .60

XVIIIe siècle. Paris: Mouton

Fauconnier, S. Frommé , R. (2004). Capturing unstable media, summary of research .61

Fayet-Scribe, S. (2000). Histoire de la documentation en France: Culture, science, et technologie de .62
l'information, 1895-1937 (CNRS histoire). Paris: CNRS

Featherstone, M. (2000). Archiving cultures. British Journal of Sociology, 51(1) .63

Febvre, L.P.V. Martin, H. J. (1976). The coming of the book: The impact of printing 1450-1800 ([New .64
ed.] ed.). London: NLB

Fielding, R. T., Gettys, J., Mogul, J., Nielsen, H. F., Masinter, L., J, P., et al. (1999). Hypertext Transfer .65
Protocol - HTTP/1.1. RFC, 2616

Fitch, K. (2003). Web site archiving: An approach to recording every materially different response . 66
produced by a website. Paper presented at the AusWeb (2003): The Ninth Australina World Wide Web
Conference, Sanctuary Cove, Australia

Florescu, D., Levy, A., Mendelzon, A. (1998). Database techniques for the World- Wide Web: A . 67
survey. SIGMOD Record 27, 59-74

Freeman, E. Gelernter, D. (1996). Lifestreams: A storage model for personal data. SIGMOD Record, .68
25(1), 80-86

Gemmell, J., Bell, G., Lueder, R., Drucker, S., Wong, C. (2002). MyLifeBits: fulfilling the Memex . 69
vision. Juan-les-Pins, France

Gibson, D., Punera, K., Tomkins, A. (2005). The volume and evolution of web page templates. Paper . 70
presented at the WWW'05 14th international conference on World Wide Web, Chiba, Japan

Gillies, J. Cailliau, R. (2000). How the Web was born: The story of the World Wide Web. Oxford: . 71
Oxford University Press

Golder, S. Huberman, B. A. (2005). The Structure of Collaborative Tagging Systems 73. Gomes, D. . 72
Silva, M. J. (2003). A Characterization of the Portuguese Web. Paper presented at the 3rd Workshop on Web
Archives (IWAW'03), Trondheim, Norway

Gulli, A. Signorini, A. (2005). The indexable web is more than 11.5 billion pages. Chiba, Japan .74

Halavais, A. (2004). Tracking Ideas in the Blogosphere .75

Hallgrímsson, T. (2006). Access and finding aids or web archives. In J. Masanès (Ed.), Web archiving. .76
Berlin Heidelberg New York: Springer

Hine, C. (2000). Virtual ethnography. London; Thousand Oaks, CA: Sage .77

Hofmann, M. Beaumont, L. R. (2005). Content networking: Architecture, protocols, and practice (The .78
Morgan Kaufmann Series in Networking). Amsterdam; Boston: Morgan Kaufmann

Ingwersen, P. (1998). The calculation of web impact factors. Journal of .79

(Documentation, 54(2 .80

Jones, S. Johnson, C. (2006). Web Use and Web Studies. In J. Masanè s (Ed.), Web Archiving. Berlin .81
Heidelberg New York: Springer

Jones, W., Bruce, H., Dumais, S. (2001). Keeping found things found on the web. Atlanta, GA, USA .82

Jones, W., Bruce, H., Dumais, S. (2003). How do people get back to .83

information on the Web? How can they do it better? Paper presented at the IFIP INTERACT'03 .84

Kahle, B. (1997). Preserving the Internet. Scientific American, 397, 82–84 .85

Kahle, B. (2002). The Internet Archive. RLG DigiNews, 6(3) .86

Kimpton, M., Braggs, M., Ubois, J. (2006). Year by Year: From an Archive of the Internet to an Archive .87
on the Internet. In J. Masanè s (Ed.), Web Archiving. Berlin Heidelberg New York: Springer

Koehler, W. (1999). Unraveling the ISSUES, ACTORS, ALPHABET SOUP of the Great Domain) .88
Name Debates. Searcher, 7(5

Koehler, W. (2004). A longitudinal study of Web pages continued: a consideration of document) .89
persistence. Information Research, 9(2

Krishnamurthy, B. Rexford, J. (2001). Web protocols and practice: HTTP/1.1, networking protocols, .90
caching, and traffic measurement. Boston, MA: Addison–Wesley 91. Lagoze, C., Dean B. K., Sandy, P., Jesurogaii, S. (2005). What Is a Digital Library Anymore, Anyway? Beyond Search and Access in the NSDL. D–Lib Magazine, 11-11 92. Lampos, C., Eirinaki, M., Jevtuchova, D., Vazirgiannis, M. (2004). Archiving the Greek Web. Paper presented at the 4th International Web Archiving Workshop (IWAW'04),
(Bath (UK

Landow, G. P. (1997). Hypertext 2.0 (Rev., amplified ed.). Baltimore: Johns Hopkins University Press .93

Lavoie, B. F. Schonfeld, R. C. (2005). The systemwide print book collection. Paper presented at the CNI .94

.Lawrence, S. Giles, C. L. (1998). Searching the Web. Science, 281, 175 .95

Lawrence, S. Giles, C. L. (1999). Accessibility of Information on the Web.Nature, 400, 107–109 .96

Lecher, H. E. (2004). Informant networks, alarm systems, and research contributors. Selection and . 97
ingest process for the Digital Archive for Chinese Studies. Paper presented at the Archiving Web Resources
Conference – Issues for Cultural Heritage Institutions, NLA, Canberra, Australia

Lecher, H. E. (2006). Academic Web archiving: DACHS. In J. Masanè s (Ed.), Web archiving. Berlin . 98
Heidelberg New York: Springer

Levy, P. (1997). Collective intelligence: Mankind's emerging world in cyberspace. Cambridge, MA: . 99
Perseus Books

Liu, C. Albitz, P. (1999). DNS BIND (3rd ed.). O'Reilly Associates .100

Lueg, C. Fisher, D. (2003). From Usenet to CoWebs: Interacting with social information spaces . 101
(Computer supported cooperative work). Berlin Heidelberg London New York: Springer

Lyle, J. A. (2004). Sampling the Umich.edu Domain. Paper presented at the 4th International Web) .102
Archiving Workshop (IWAW'04), Bath (UK

Lyman, P. (2002). Archiving the World Wide Web. In CLIR (Ed.), Building a national strategy for . 103
preservation: issues in digital media archiving. Council

on Library and Information Resources and the Library of Congress .104

Lyman, P. Kahle, B. (1998). Archiving digital cultural artifacts. D–Lib Magazine .105

Mantratzis, C. Orgun, M. (2004). Towards a peer2peer world–wide–web for the broadband–enabled . 106
user community

Masanè s, J. (2002). Towards continuous Web archiving: First results and an agenda for the future. D–) .107
Lib Magazine, 8(12

Masanè s, J. (2004). Site–first priority: Implementing the frontline .108

Masanè s, J. (2005). Web archiving methods and approaches: A comparative study. Library Trends .109

Masanès, J. (2006a). Collecting the hidden web. In J. Masanès (Ed.), Web archiving. Berlin . 110
Heidelberg New York: Springer

Masanès, J. (2006b). Selection for Web Archives. In J. Masanès (Ed.), Web archiving. Berlin . 111
Heidelberg New York: Springer

Mohr, G., Kimpton, M., Stack, M. Ranitovic, I. (2004). Introduction to Heritrix, an archival quality . 112
web crawler. Paper presented at the 4th

International Web Archiving Workshop (IWAW'04), Bath (UK) .113

Mueller, M. (2002). Ruling the root: Internet governance and the taming of cyberspace. Cambridge, . 114
MA: MIT

Najork, M. Heydon, A. (2001). High-performance Web crawling. SRC Research Report .115

Najork, M. Wiener, J. (2001). Breadth-first search crawling yields high-quality pages. Paper . 116
presented at the 10th World Wide Web Conference (WWW'10), Hong Kong

National Archives of Australia. (2001). Archiving Web resources: A policy for keeping records of . 117
web-based activity in the Commonwealth Government

Osborn, T. (1999). The ordinariness of the archive. History of the human sciences, 12(2) .118

Page, L., Brin, S., Motwani, R. Winograd, T. (1998). The Pagerank citation ranking: Bringing order to .119
the Web, 17

Pandey, S. Olston, C. (2005). User-centric Web crawling. Chiba, Japan .120

Pant, G., Srinivasan, P. Menczer, F. (2004). Crawling the Web. In M. Levene, A. Poulovassilis (Eds.), .121
Web Dynamics. (pp. 153-178). Berlin Heidelberg New York: Springer

Pastor-Satorras, R. Vespignani, A. (2004). Evolution and structure of the Internet: A statistical) .122
physics approach. Cambridge, UK; New York: Cambridge University Press 123. Phillips, M. E. (2005).
Selective archiving of Web Resources: A study of 124. acquisition costs at the National Library of Australia.
RLG DigiNews, 9(3

Qin, J., Zhou, Y. Chau, M. (2004). Building domain-specific web collections for scientific digital . 125
libraries: A meta-search enhanced focused crawling method. Tuscon, AZ, USA

Rekimoto, J. (1999). Time-machine computing: A time-centric approach for the information . 126
environment. Paper presented at the 12th annual ACM symposium on User interface software and
technology, Asheville, North Carolina, USA

Riché, P. (1996). La bibliothèque et la formation de la culture médiévale. In M. Baratin, C. Jacob) .127
.(Eds.), Le pouvoir des bibliothèques: la mémoire des livres en Occident (p

ص: 59

Paris: A. Michel .(338

Ringel, M., Cutrell, E., Dumais, S., Horvitz, E. (2003). Milestones in Time: The Value of Landmarks . 128
in Retrieving Information from Personal Stores. Paper presented at the IFIP INTERACT '03

?RLG. (2006). Web Archiving Program. http://www.rlg.org/en/page.php .129

Page-ID=399 .130

Roche, X. (2006). Copying web sites. In J. Masanè s (Ed.), Web Archiving. Berlin Heidelberg New . 131
York: Springer

Rosenfeld, L. Morville, P. (2002). Information architecture for the World Wide Web (2nd ed.). . 132
Cambridge, MA: O'Reilly

Scharl, A. (2000). Evolutionary Web development (Applied computing). Berlin Heidelberg New . 133
York: Springer

Shepherd, M. Polanyi, L. (2000). Genre in Digital Documents. Paper presented at the Proceedings of . 134
the 33rd Hawaii International Conference on System Sciences - vol. 3

Sonnenreich, W. (1997). A History of Search Engines. http://www.wiley.com/legacy/ . 135
compbooks/sonnenreich/history.html

Spinellis, D. (2003). The decay and failures of web references. Communications of ACM, 46(1), 71- . 136
77

Stack, M. (2005). Full Text Search of Web Archive Collections. Paper presented at the IWAW'05, . 137
Vienna, Austria

Star, S. L. Ruhleder, K. (1994). Steps towards an ecology of infrastructure: Complex problems in . 138
design and access for large-scale collaborative systems. Chapel Hill, NC, United States

Teevan, J. (2004). How people re-find Information when the Web changes. AIM- 2004-012 .139

Thelwall, M. (2001). Extracting macroscopic information from Web links. Journal of the American . 140
Society for Information Science and Technology, 52(13), 1157-1168 141. Thelwall, M. (2006).
Interpreting social science link analysis research: A theoretical framework. Journal of American Society of

Information Science and Technology 57(1), 60–68

Thelwall, M. Harries, G. (2004). Do the websites of higher rated scholars have .142

ص: 60

significantly more online impact? Journal of the American Society for Information Science and Technology, 55(2), 149-59

143. Thelwall, M. Vaughan, L. (2004). A fair history of the Web? Examining country balance in the Internet archive. Library Information Science Research, 26(2), 162- 176

144. Ubois, J. (2002). The Oakland archive policy. Recommendations for managing removal requests and preserving archival integrity

145. Voerman, G., Keyzer, A., Hollander, F. D., Druiven, H. (2002). Archiving the Web: Political Party) Web sites in the Netherlands. European Political Science, 2 (1

چکیده

پیشـرفت در رایـانه و ارتباطات این امکان را فراهم ساخته که ذخیره کتاب، صـفحه موسـیقی، فیلم، بسـته های نرم افزاری و تمام صـفحات عمومی وب که تاکنون ساخته شده اند هزینه – سودمندی داشـته باشـند و دسترسی به این مجموعه ها از طریق اینترنت برای همه، از جوان تا پیر، در سرتاسر دنیا فراهم شود. برای سال های آینده رسالت آرشیو مشخص است : ایجاد کتابخانه ای جهانی که تمام دانش به سهول در اختیار هر مرد و زن و کودک در سرتاسر جهان قرار گیرد . در مقاله حاضر بحث آرشیو اینترنت و آرشیو در محیط اینترنت ، دسترس پذیری دراز مدت تمام دانش برای همه افراد جهان پیشرفت های فن آورانه ایجاد آرشیو اینترنتی مورد بررسی قرار می گیرد. سپس برنامه ها و پروژه های نمونه آرشیو اینترنتی در دنیا آرشیو اروپا و پتاباکس معرفی می شوند.

ص: 62

از آرشیو اینترنت تا آرشیو در اینترنت

اشاره

*از آرشیو اینترنت تا آرشیو در اینترنت (1)

میشل کیمتون (2) | جف یوبویس (3)

ترجمه : مرضیه هدایت (4)

مقدمه

«آرشیو اینترنت» (5) ، از آغـاز کار خود در 1995، تاکنون، هـدف درازمدت فراهم کردن دسترسـی جهانی به تمام دانش در طول عمر را دنبال کرده است.

در 10 سال گذشـته، آرشـیو، در برنامه های کشورهای مختلف، برای کاربران متعدد و گوناگون، شرکت داشته و برای حفظ صفحات وبی، کتاب، موسیقی، نرم افزار، و تصاویر پویا فعالیت کرده تا آن ها را از طریق اینترنت دسترس پذیر سازد.

در حال حاضـر، ماشین Wayback آرشـیو، روزانه به 70,000 بازدیـد کننده، و درهر ثانیه به 200 درخواست پاسـخ می دهد. مجموعه 600 ترابایتی آن شامل 50 بیلیون صفحه وب، 30000 جلد کتاب 36000 قطعه موسیقی، و 15000 قطعه فیلم است.

حضور صدها مشترک این امر را امکان پذیر ساخته است. به واسطه پیشرفت های فنی در صنعت رایانه،

1- Year – by – Year: From an Archive of the Internet to an Archive on the Internet: in Masanes, Julien (ed.), Web Archiving. Berlin Heidelberg New York: Springer.pp.201-212.

2- Michele Kimpton

3- Jeff Ubois

4- عضو هیئت علمی سازمان اسناد و کتابخانه ملی ایران

5- Internet Archive

امکانات این سیستم هر 12 – 18 ماه دو برابر می شود. در جامعه آرشیوی 10 سال زمان زیادی نیست، اما زمان مناسب برای خلق یک عامل 100 برابر پیشرفت است: از سال 1997 تاکنون قیمت دیسک خام بیش از 99درصد کاهش یافته است؛ یعنی از 180 دلار به 50 سنت برای هر گیگابایت رسیده و بیش از 25 میلیون ارتباط پهنای باند فقط در ایالات متحده اضافه شده است.

با توجه به اینکه میزان پیشرفت رایانه بر اساس ذخیره و بازیابی است، می توان آن چه را که راج ردی (1) از دانشگاه کارنگی ملون (2) با عبارت «دستیابی جهانی به تمام دانش» بیان کرده است، محتمل دانست.

پیشینه: چاپ اینترنتی اولیه

از اواسط دهه 1980 معلوم بود که تغییری در دنیای چاپ الکترونیکی در حال ظهور است. در نیمه اول دهه 1990 ، حرکت روزنامه ها از شکل پایگاه های بسته با مالکیت انحصاری به شکل اینترنتی شروع شد و نظریه اینترنت به عنوان کتابخانه شروع به شکل گیری کرد. نظام های چاپی اینترنتی از قبیل WAIS (3) و Gopher ، به عنوان متمّم های صفحات وبی دیده شدند؛ استعاره اینترنت به عنوان کتابی با صفحات وبی که فهرست مندرجات آن توسط سرورهای Gopher و نمایه آن توسط سرورهای WAIS تهیه می شد، به عنوان جایگزینی برای استعاره اینترنت به عنوان «فراشاهراه» محسوب شد.

شروع به کار خدمات آلتا ویستا (4) در دسامبر 1995 ثابت کرد که تمام صفحات موجود در وب را می توان مجموعه ای یگانه دانست که قابل نمایه شدن و جست و جو کردن بر روی شبکه برای تمام کاربران است؛ اما معلوم نبود که این صفحات را چگونه باید حفظ و نگهداری کرد.

آغاز به کار آرشیو اینترنت

آرشیو اینترنت، به طور رسمی با همکاری بروس گیلیت (5) و بروستر کال (6) ، در آوریل 1996 شروع شد. در آن زمان، پیوندهای شکسته (404درگاه) یک مشکل روبه رشد بود و روشن بود که بیشتر صفحات وبی عمر کوتاهی دارند. برای این مشکل راه حلی مورد نیاز بود و نظامی برای آرشیو صفحات وبی قبل از اینکه پاک شوند یک راهکار ضروری به نظر می رسید.

این مسئله، به تصمیم گیری راجع به طرح اولیه در «آرشیو» درباره سیاستگذاری مجموعه ها منجر شد؛ مانند حریص بودن نسبت به جمع آوری مطالبی که در معرض خطر از بین رفتن بودند، و شکار فرصت ها برای جمع آوری و رقومی کردن مواردی از گذشته مثل پستینگ های یوزنت (7).

با این حال، هنوز در 1996، در جامعه اینترنتی اهمیت از دست رفتن صفحات وبی مشکل چندان حساسیت

1- Raj Reddy

2- Carnegie Mellon University

3- WAIS (Wide Area Information Servers)

Alta Vista –4

Bruce Gilliat –5

Brewster Kahle –6

Usenet –7

برانگیز نبود. چون وب سابقه تاریخی چندانی نداشت توضیح فواید صفحات از دست رفته مشکل بود.

برای نشان دادن ارزش بالقوه چنین صـفحاتی، آرشیو، با مؤسسه اسمیتسونین (1) در واشـنگتن دی.سی، برای جمع آوری نسخه فوری (2) وبگـاه های تمام نامزدهای ریاست جمهوری 1996 (3) همکـاری کرد. ابزار انجام این پروژه چنـدان مورد رضایت نبود. این ابزار به طور اساسـی عبارت بودند از: تجهیزات رایانه های شخصـی که بر پایه ضـبط وب سایت ها از طریق دنبال کردن پیوندها از صـفحه اصلی عمل می کردند

این داده ها، در نهایت، به آرشیو ریاست جمهوری اسمیتسونین منتقل شد که در حال حاضر شامل

صفحاتی از 5 حزب سیاسی و کاندیداهای پرشمار ریاست جمهوری از بیل کلینتون گرفته تا پت بوکانن (4) است. بسیاری از سایت ها در این آرشیو با حذف کاندیداها بسته شدند.

بر اسـاس این موفقیت، کتابخـانه کنگره به آرشـیو مـأموریت داد تـا مجموعه پیوسـته متمرکزی از انتخابـات سال 2000 ایجاد کنـد و این درخواست را نیز برای انتخابات 2002 تجدید کرد.

همچنین، در سـال 1996، آرشـیو، ارتبـاطش را بـا اینترنت الکسا شـروع کرد. اینترنت الکسا، مؤسـسه ای انتفاعی است که خزشـگری و آرشـیوسازی وب را در نوامبر شـروع کرد تا داده های پیشنهادی نوار ابزار مرورگر درباره سایت های دیده شده را Plug-in (وصلینه) کند و بر اساس داده های جمع آوری شده از سایر کاربران و پیشنهاد درباره صفحات مرتبط – که ممکن است مورد توجّه باشند – عمل می کرد.

دو پیشرفت دیگر از سال 1996 نیز قابل ذکرند. نخستین پیشـرفت، فناورانه است. در 1996، هنوز ارجحیت نوار نسـبت به دیسک از نظر قیمت قابل توجه بود و آرشیو اولین نسل زیرساختار را با استفاده از روبات های ذخیره نواری ساخت که با ADIC 50 شروع شد. با وجود همکاری های سـخاوتمندانه فروشـندگان اصـلی، در نهایت ثابت شد این کار قابل دفاع نیست. نیازهای دستیابی که توسط کاربران آرشـیو مطرح می شدند بیش از حد شدید و زمان بازیابی خیلی کند بود. همانگونه که بروس گیلیت به طنز می گوید : گرفتن یک صـفحه می تواند در چند ثانیه باشد... یا چند روز بعد.

دومین پیشـرفت به مسائل قانونی مربوط می شد. مسائل قانونی جمع آوری صفحات وبی با الگوریتم حریصانه و ارائه آن ها بر پایه «صرف نظر کردن» (5) آن گونه که آرشـیو در آن سـال شـروع به انجـامش کرد – روشن نبود. بهبود پروتکل ارتقای روبات های محـدودیت متنی متعلق به آلتا ویستا گامی مهم بود چون «پیش فرض ها» را تغییر داد – صاحبان صفحات وبی که از گذاشتن آن ها در نمایه موتور جست و جو یا در آرشـیو ابا داشـتند، روشی ساده برای صرف نظر کردن یا انتقال صفحاتی که صاحب آن بودند (اثبات آن از طریق توان آن ها برای مناسب سازی راهنمای پایه ای وبگاه مورد نظر بود) داشتند. روبات های متنی برای انتقال موارد مورد نظر برای آن ها که صاحب صفحات در آرشـیو بودنـد راه حلی ارائه داد، ولی نتـوانست مشـکل انتقـال صـفحات متعلـق به دیگران را رفع کنـد. این موارد انتقـالی، در کنفرانس کوچکی

1- Smithsonian

Snapshots –2

http://movie0.archive.org/96–Elections/index.htm –3

Pat Buchanan –4

Opt–out –5

توضیح داده شد، که توسط آرشیو اینترنت در 2002 ، در یوسی برکلی (1) برگزار شد (یوبویس 2002).

ساختار پیوندی و روبات های نواری

جمع آوری صفحات وبی، داده های پیوندی و «ردّ پاهای استفاده ای (2) » به عنوان نمونه، فرصت هایی که میلیون ها کاربر وب با رفتن از صفحه ای به صفحه ای ایجاد کرده بودند؛ در اوایل 1997 با معرفی نوار ابزار الکسا، یک مرورگر Plug-in از طریق تهیه اطلاعات روی سایتی که در حال دیده شدن بود و نیز پیشنهادهایی برای سایت های مرتبط به کاربران کمک می کرد تا به ناوبری در وب بپردازند.

داده های پیوندی و ردپاهای استفاده ای که توسط الکسا جمع آوری می شدند، به عنوان نظام فیلترینگ اشتراکی عمل می کرد و صفحاتی را برجسته می کرد که جامعه اینترنتی دارای بیشترین ارزش می دانست. پیوندها و کلیک ها بر اساس ارزش صفحه داده شده امتیاز می گرفتند.

توان مشخص کردن ارزش یک صفحه، به طور خودکار، تا حد زیادی با یک طرح اساسی دیگر مرتبط بود. در 1997، تعدادی از بزرگ ترین و موفق ترین کتابخانه های وبی، فهرست سایت هایی مانند یاهو بودند. اما در مورد نحوه سنجش راهکارهای فهرست نویسی دستنامه ای که ممکن بود زمانش بگذرد، اطمینانی نبود. آیا احتمال حذف فهرستنویسی دستنامه ای و پذیرش روند انتخاب های قطعی با دیدگاه «جمع آوری کامل» که با فراداده ایجاد شده توسط کاربر به جای فهرست ترکیب می شد، امکان پذیر بود؟

به نظر می رسید پاسخ مثبت باشد. به همین دلیل، الکسا بر اساس ثبت میزان استفاده، به خزش در صفحات پرداخت. سنجش آن از طریق داده های جمع آوری شده توسط نوار ابزار مرورگر انجام می شد. صفحاتی که بسیار دیده شده بودند، نخستین هایی بودند که برایشان نسخه پشتیبان تهیه می شد.

خزشگر الکسا، طوری تنظیم شده بود که هر 8 هفته یک نسخه فوری از وب تهیه کند، و هنوز این برنامه اجرا می شود، اگر چه اندازه هر خزش از یک ترابایت در سال 1997 به 100 ترابایت در 2004 رسیده است.

مورد دیگری که آرشیو در سال 1997 با آن روبه رو شد این بود که آیا به ذخیره روی نوار می توان اعتماد کرد یا ذخیره روی دیسک؟ از نظر قیمت هنوز نوار بر دیسک ارجح بود، ولی دستیابی کند بود.

همانگونه که در مقاله گری و شنوی (1999) (3) آمده «نسبت قیمت نوار، دیسک و RAM تقریباً 10:1 1000 است، به این معنی که ذخیره روی دیسک 10 بار گران تر از ذخیره روی نوار و ذخیره روی 100 RAM بار گران تر از ذخیره روی دیسک است».

اما وقتی هزینه دستیابی را با یارد آهنی بسنجیم، دیسک واقعاً بسیار ارزان تر است؛ «هزینه آرشیو نواری در مقایسه با هر ترابایت ذخیره روی دیسک نصف است، ولی دستیابی آسان به داده ها از طریق نوار ممکن نیست. هزینه هر دستیابی از طریق نوار، به صورت تصادفی، حدود 100 هزار بار بیشتر است (دستیابی از طریق دیسک : 100 دستیابی در ثانیه با هزینه 1 دلار در مقابل 10/000 دستیابی در ثانیه از طریق نوار با هزینه 10000 دلار است)» (گری و شنوی 1999).

UC Berkeley –1

Usage trails –2

Gray and Shenoy –3

1998:حضور داده های آرشیوی بر روی (تقریباً) هر دسکتاپ

پس از آن که در 1996، نت اسکیپ (1) در دسترس عموم قرار گرفت؛ در 1998 ،اینترنت و مشاغل وابسته به آن در یک تغییر جهانی در اولویت های سـرمایه گذاری مورد توجه قرار گرفت. همانطور که بیلیون ها دلار در بازارهای عمومی، سـرمایه گذاری های مخاطره آمیز و راه اندازی اینترنت هزینه می شد، شماره صفحات در آرشیو هر 3 – 6 ماه دو برابر و تعداد کاربران اینترنت نیز هر چند ماه دو برابر می شد. به تدریج دستیابی به یک دغدغه تبدیل می شد.

در تلاش برای ایجاد دسترسـی به موجودی و تأسـیس آرشیو و اینترنت الکسا، به عنوان بخشی از زیر ساختار اینترنت، الکسا قراردادهایی با مرورگر های مایکرو سافت و نت اسکیپ بست. این امر به معنای حضور الکسا در 90 درصد رایانه های دنیا بود و کاربران چه می دانستند چه نه، الکسا بود که دسترسی به داده های تهیّه شده توسط آرشیو اینترنت را به آن ها می داد.

برای آرشیو، ارائه داده ها برای ده ها میلیون کاربر تأثیر شدیدی بر زیر ساختار نوار مدار آن داشت.

در پایان 1998 دو امر مشخص بود:

– به دلیـل میزان زیاد تقاضا رفتن از نوار به دیسک نیاز مبرمی بود. همان طور که تقاضای دسترس افزایش می یافت توان روبات های نواری برای پاسخگویی مورد تردید قرار می گرفت.

– سیاسـتگذاری های مجموعه های دستی گران تر از دیسک بود، یعنی آرشیو سازی بر پایه فهرست نویسی دستی، به ویژه وبگاه ها، بسیار گران تر از آرشیوسازی تمام سایت های قابل دسترسی بر اساس داده های جمع آوری شده از کاربران نهایی توسط الکسا بود.

1999: از نوار تا دیسک، یک خزشگر جدید و تصاویر متحرک

موفقیت تجاری خدمات الکسا ، که به واسـطه حضورش در هر رایانه شخصـی مرتبط با اینترنت، بخش مشخصـی بود، باعث شد آمازون، در سال 1999، الکسا را خریداری کند. این امر، در نهایت، به تغییراتی در ساختار هر دو سازمان منجر شد.

یک پیشرفت فناورانه مهم در سال 1999، ابداع خزشگر جدید توسط اندی جوئل (2) بود. این خزشگر جدید، برای جمع آوری داده های وبی توانـایی بیشتری داشت و از طریق ماشـین هـای چنـدگانه قابل مـدیریت بود. این خزشـگر، به افزایش توانایی الکسا در فیلتر کردن 16 بیلیون URL و 4 بیلیون خزشگری و گسترش وسعت و عمق خزش خود بپردازد.

همچنین، تصمیماتی درباره فرمت فایل ARC اخذ شد، که برای ذخیره صـفحات وبی استفاده می شد. ویژه سازی فرمت فایل ARC – که برای نیازهـای متعـددی طراحی شـده بود – از آغـاز، در سـال 1996 توسـط مایک برنر (3) و بروستر کـال توسـعه یـافت (برنر و کال 1996). نیازهای مذکور عبارت اند از:

- فایل باید خود شمول باشد: فایل باید اجازه بدهد اشیای متراکم، بدون استفاده از فایل نمایه همراه،

Netscape –1

Andy Jewel –2

Mike Burner –3

شناسایی و باز شوند؛

- فرمت باید برای پذیرش فایل های بازیابی شده از طریق پروتکل های شبکه ای متنوع مانند FTP ، HTTP ، اخبار، گوفر، و پست الکترونیکی قابل گسترش باشد؛

- فایل باید قابلیت «جریان داشتن» (1) داشته باشد: فایل باید بتواند فایل های چندگانه آرشیو را در جریان

داده ها به هم پیوند دهد؛

- رکورد باید با یکبار نوشتن ماندگار: باشد یکپارچگی فایل نباید به ایجاد بعدی نمایه درون فایلی محتوا وابسته باشد.

در عین حال که خاص سازی به نمایه بیرونی محتوا و شیء – جبرانی نیاز ندارد، چنین نمایه ای به میزان زیادی قابلیت بازیابی اشیای ذخیره شده در این فرمت را افزایش می دهد. در حال حاضر، آرشیو چنین نمایه هایی را فراهم کرده و به دنبال استاندارد سازی فرمت های آن ها از طریق کنسرسیوم حفاظت بین المللی اینترنتی (2) است.

در 999 ، حرکت از صفحات وبی، به سایر انواع داده ها نیز شروع شد. در این سال، هزینه ذخیره آن قدر کاهش یافت که آرشیو توانست به گردآوری تصاویر متحرک بپردازد. از طریق شراکت با ریک پری لینگر (3) ، از آرشیو Prelinger، پروژه دیجیتالی کردن 1000 فیلم (با هزینه حداکثر 160000 دلار) و نیز آرشیو کردن اخبار تلویزیون در پایان سال شروع شد.

2000: ایجاد مجموعه های موضوعی وب

در سال 2000، آرشیو به سطحی از ثبات فناوری رسید. پذیرش داده های خزشگری روش معمول شد و مهاجرت از نوار به دیسک سپری مجموعه گشت.

عکس

شناسایی و باز شوند؛

- فرمت باید برای پذیرش فایل‌های بازیابی شده از طریق پروتکل‌های شبکه‌ای متنوع مانند HTTP ، FTP ، اخبار، گوفر، و پست الکترونیکی قابل گسترش باشد؛
- فایل باید قابلیت «جریان داشتن»[1] داشته باشد: فایل باید بتواند فایل‌های چندگانه آرشیو را در جریان داده‌ها به‌هم پیوند دهد؛
- رکورد باید با یکبار نوشتن ماندگار باشد: یکپارچگی فایل نباید به ایجاد بعدی نمایه درون فایلی محتوا وابسته باشد.

در عین حال که خاص‌سازی به نمایه بیرونی محتوا و شیء - جبرانی نیاز ندارد، چنین نمایه‌ای به میزان زیادی قابلیت بازیابی اشیای ذخیره شده در این فرمت را افزایش می‌دهد. در حال حاضر، آرشیو، چنین نمایه‌هایی را فراهم کرده و به دنبال استانداردسازی فرمت‌های آنها از طریق کنسرسیوم حفاظت بین‌المللی اینترنتی[2] است.

در ۹۹۹ ، حرکت از صفحات وبی، به سایر انواع داده‌ها نیز شروع شد. در این سال، هزینه ذخیره آنقدر کاهش یافت که آرشیو توانست به گردآوری تصاویر متحرک بپردازد. از طریق شراکت با ریک پری لینگر[3] ، از آرشیو Prelinger، پروژه دیجیتالی کردن ۱۰۰۰ فیلم (با هزینه حداکثر ۱۶۰۰۰۰ دلار) و نیز آرشیو کردن اخبار تلویزیون در پایان سال شروع شد.

۲۰۰۰ : ایجاد مجموعه‌های موضوعی وب

در سال ۲۰۰۰، آرشیو به سطحی ازثبات فناوری رسید. پذیرش داده‌های خزشگری روش معمول شد و مهاجرت از نوار به دیسک سپری گشت.

جدول ۱ . مجموعه‌های آرشیو اینترنت در مارس ۲۰۰۰

مجموعه	واحد	اندازه
وب (۱۹۹۶ تا ماه ۳ سال ۲۰۰۰)	۱ بیلیون صفحه	۱۳/۸ ترابایت
FTP (۱۹۹۶)	۵۰/۰۰۰ سایت	۰/۰۵ ترابایت
Usenet (۱۹۹۶- ۱۹۹۸)	۱۶ میلیون پستینگ	۰/۵۹۲ ترابایت

در سال ۲۰۰۰، انتخابات دیگری در ایالات متحده برگزار شد و این بار تمام انتخاب‌کنندگان دسترسی اینترنتی داشتند. از نظر سیاسی، روشن بود که حضور در اینترنت برای برنده شدن حیاتی است و این امر تمرکز بر برخط بودن سیاسی را افزایش داد. آرشیو اینترنت، با کتابخانه کنگره برای گردآوری سایت‌های سیاسی شریک شد.

این اقدام، اولین پروژه آرشیو با کتابخانه کنگره بود و برای بسیاری از کارکنان آن به معنی انتقال از

1. Streamable
2. International Internet Preservation Consortium (IIPC) http:// netpreserve.org
3. Rick Prelinger

جدول 1. مجموعه های آرشیوی اینترنت در مارس 2000

در سال 2000، انتخابات دیگری در ایالات متحده برگزار شد و این بار تمام انتخاب کنندگان دسترسی اینترنتی داشتند. از نظر سیاسی، روشن بود که حضور در اینترنت برای برنده شدن حیاتی است و این امر تمرکز بر برخط بودن سیاسی را افزایش داد. آرشیو اینترنت با کتابخانه کنگره برای گردآوری سایت های سیاسی شریک شد.

این اقدام، اولین پروژه آرشیو با کتابخانه کنگره بود و برای بسیاری از کارکنان آن به معنی انتقال از

Streamable -1

International Internet Preservation Consortium (IIPC) http://netpreserve.org -2

Rick Prelinger -3

یک پروژه تجربی به یک مؤسسه ثابت بود.

فکر ایجاد دسترسی به آثار ناپایدار نگهداری شده باعث حرکت و جنبش آنی و آرشیو تصاویر متحرک (1) در سال 2000 تأسیس شد. در حال حاضر، این آرشیو، با پروانه Creative Commons، شامل فیلم هایی از آرشیو Prelinger مجموعه ای بیش از 1900 فیلم ناپایدار (2) (پیام های بازرگانی، آموزشی، صنعتی و آماتور) است. این مجموعه، در حال حاضر دارای بیش از 10 درصد تمام فیلم های ناپایدار تولید شده در سالهای 1927 و 1987 در آمریکاست و یکی از کامل ترین و متنوع ترین مجموعه فیلم های ژانرهایی است که تعداد زیادی از آن ها نگهداری نشده اند.

2001: دسترسی از طریق ماشین Wayback: آرشیو یازده سپتامبر

طی یک بازه زمانی یکساله از مارس 2000 تا مارس 2001 ،آرشیو ، اندازه موجودی خود را 3 برابر کرد و به بیش از 40 ترابایت رساند. در این دوره، آرشیو هر ماه تقریباً 10 ترابایت رشد می کرد.

عکس

یک پروژه تجربی به یک مؤسسه ثابت بود.

فکر ایجاد دسترسی به آثار ناپایدار نگهداری شده باعث حرکت و جنبش آنی و آرشیو تصاویر متحرک[1] در سال ۲۰۰۰ تأسیس شد. در حال حاضر، این آرشیو، با پروانه Creative Commons، شامل فیلم‌هایی ازآرشیو Prelinger، مجموعه‌ای بیش از ۱۹۰۰ فیلم ناپایدار[2] (پیام‌های بازرگانی، آموزشی، صنعتی، و آماتور) است. این مجموعه، در حال حاضر، دارای بیش از ۱۰درصد تمام فیلم‌های ناپایدار تولید شده در سال‌های ۱۹۲۷ و ۱۹۸۷ در آمریکاست و یکی از کامل‌ترین و متنوع‌ترین مجموعه فیلم‌های ژانرهایی است که تعداد زیادی از آنها نگهداری نشده‌اند.

۲۰۰۱ : دسترسی از طریق ماشین Wayback: آرشیو یازده سپتامبر

طی یک بازه زمانی یکساله از مارس ۲۰۰۰ تا مارس ۲۰۰۱ ،آرشیو، اندازه موجودی خود را ۳ برابر کرد و به بیش از ۴۰ ترابایت رساند. در این دوره، آرشیو، هر ماه تقریباً ۱۰ ترابایت رشد می‌کرد.

جدول ۲ . مجموعه‌های آرشیو در مارس ۲۰۰۱

مجموعه	واحد	اندازه
۱۹۹۶- ماه ۳ سال ۲۰۰۱	٤ بیلیون صفحه	٤۰ ترابایت
آرشیو انتخابات ۲۰۰۰	۲۰۰ میلیون صفحه	۲ ترابایت
Usenet : ۱۹۹۶- ۱۹۹۸، ۲۰۰۰- ماه ۳ سال ۲۰۰۱	۱٦ میلیون پستینگ	۰/۵ ترابایت
فیلمهای‌آرشیو: حدود۱۹۰۳- حدود ۱۹۷۳	۳٦۰ فیلم	۰/۵ ترابایت
Arpanet : اسناد تاریخی	۵۰۰۰ صفحه	<۰/۱ ترابایت

اما سال ۲۰۰۱ برای بسیاری از سازمان‌های دارای فناوری بالا در محدوده سانفرانسیسکو، سال سختی بود. سقوط بازار سرمایه، واگذاری صدها شرکت محلی، و حمله به مرکز تجارت جهانی در نیویورک بر کارهای آرشیو اثر گذاشت. به‌ویژه، از دست رفتن مشاغلِ«های‌تک»[3] در واقعه ۱۱ سپتامبر باعث تمرکز برآن شد. همانگونه که برای توان آرشیو نیز درتهیه تصاویر متحرک و پاسخ به این وقایع حکم نوعی آزمون را داشت. در اوایل ۲۰۰۱ ، شاید مهم‌ترین پرسش پیش روی آرشیو این بود که چگونه به بهترین وجه، دسترسی به مجموعه را فراهم کند. داده‌های بسیاری، به‌طور مستقیم، از طریق خدمات الکسا در اختیار عموم قرار می‌گرفت، ولی دسترسی مستقیم به مجموعه‌ها هنوز نیازمند مهارت‌های برنامه‌ریزی یونیکس[4] بود.

برنامه‌ریزان الکسا، در قراردادی با آرشیو ، برنامه ماشین Wayback راساختند که خدمات دسترسی

1. Moving Images Archive
2. ephemeral
3. high tech
4. Unix

جدول2. مجموعه های آرشیو در مارس 2001

اما سال 2001 برای بسیاری از سازمان های دارای فناوری بالا در محدوده سانفرانسیسکو، سال سختی بود سقوط بازار سرمایه، واگذاری صدها شرکت محلی، و حمله به مرکز تجارت جهانی در نیویورک بر کارهای آرشیو اثر گذاشت. به ویژه، از دست رفتن مشاغلِ«های تک» (3) در واقعه 11 سپتامبر باعث تمرکز بر آن شد. همانگونه که برای توان آرشیو نیز در تهیه تصاویر متحرک و پاسخ به این وقایع حکم نوعی آزمون را داشت. در اوایل 2001، شاید مهم ترین پرسش پیش روی آرشیو این بود که چگونه به بهترین وجه، دسترسی به مجموعه را فراهم کند. داده های بسیاری، به طور مستقیم، از طریق خدمات الکسا در اختیار عموم قرار می گرفت، ولی دسترسی مستقیم به مجموعه

ها هنوز نیازمند مهارت های برنامه ریزی یونیکس (4) بود.

برنامه ریزان الکسا، در قراردادی با آرشیو ، برنامه ماشین Wayback را ساختند که خدمات دسترسی

1- Moving Images Archive

2- ephemeral

3- high tech

4- Unix

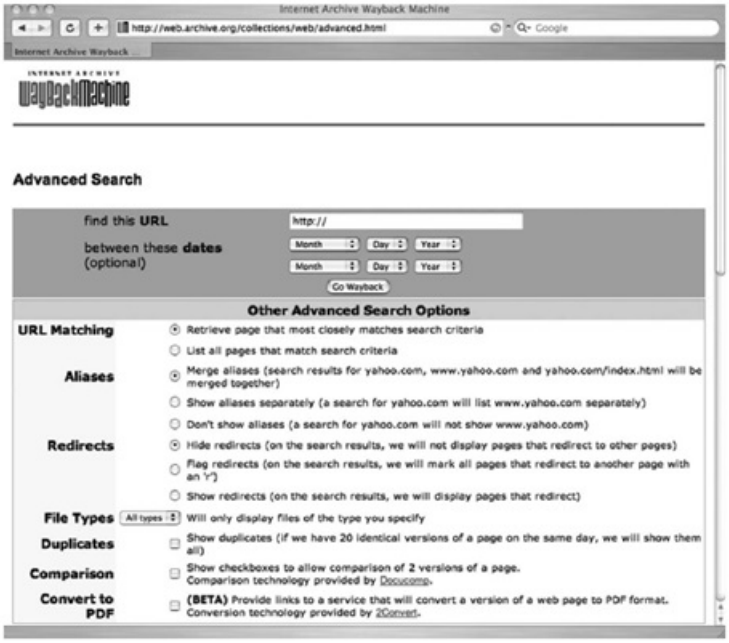
به محتوای آرشیو را بر پایه URL ها می داد 24 اکتبر 2001، ماشین Wayback به کار افتاد و دسترسی به بیش از 10 بیلیون صفحه وبی آرشیو شده و 100 ترابایت داده ممکن شد.

در آن زمـان، داده هـا روی Hewelett Packar ذخیره می شـد و سـرورهای uslab.com از سیسـتم های عامل FreeBSD و Linux استفاده می کرد. هر رایانه، حدود 512 مگابایت حافظه داشت و به طور کلی بیش از 300 گیگابایت روی دیسک های IDE بود.

عکس

به محتوای آرشیو را بر پایه URLها می‌داد. ۲٤ اکتبر ۲۰۰۱، ماشین Wayback به‌کار افتاد و دسترسی به بیش از ۱۰ بیلیون صفحه وبی آرشیو شده و ۱۰۰ ترابایت داده ممکن شد.

در آن زمان، داده‌ها، روی Hewelett Packar ذخیره می‌شـد و سرورهای uslab.com از سیستم‌های عامل FreeBSD و Linux اسـتفاده می‌کرد. هر رایانه، حدود ۵۱۲ مگابایت حافظه داشت و به‌طور کلی بیش از ۳۰۰ گیگابایت روی دیسک‌های IDE بود.



تصویر ۱. ۹. گزینه‌های فعلی جستوجو برای Wayback machine

پروژه مهم دیگر سال ۲۰۰۱ ، آرشیو ۱۱ سپتامبر[1] بود. با همکاری کتابخانه کنگره، آرشیو، تصاویری از بیش از ۳۰،۰۰۰ وبگاه منتخب را از ۱۱ سپتامبر ۲۰۰۱ تا ۱ دسامبر ۲۰۰۱ و صدها ساعت پخش اخبار را گردآوری کرد.

۲۰۰۲ : کتابخانه اسکندریه[2]، کتابخانه سیار، و حق مؤلف

در سال ۲۰۰۲، آرشیو،۵ برنامه مهم دیگر یعنی افزایش مجموعه‌ها، گزینه‌های دستیابی به این مجموعه‌ها، همکاری با سایر سازمان‌ها و تعیین راهکارها را بر عهده داشت.

1. September 11 Archive
2.The library of Alexandria

تصویر 1. 9. گزینه های فعلی جستجو برای Wayback machine

پروژه مهم دیگر سال 2001 ، آرشیو 11 سپتامبر (1) بود. با همکاری کتابخانه کنگره، آرشیو، تصاویری از بیش از 000,30 وبگاه منتخب را از 11 سپتامبر 2001 تا 1 دسامبر 2001 و صدها ساعت پخش اخبار را گردآوری کرد.

2002: کتابخانه اسکندریه، کتابخانه سیار، و حق مؤلف

*کتابخانه اسکندریه، کتابخانه سیار، و حق مؤلف (2)

در سال 2002، آرشیو،5 برنامه مهم دیگر یعنی افزایش مجموعه ها، گزینه های دستیابی به این مجموعه ها، همکاری با سایر سازمان ها و تعیین راهکارها را بر عهده داشت.

1- September 11 Archive

2- The library of Alexandria

نخستین و بزرگ ترین پروژه ایجاد سایت قرینه در کتابخانه اسکندریه در مصر بود. سرورها و بیش از 100 ترابایت داده، که بیش از 5 میلیون دلار ارزش داشتند به مصر برده شد و برای افتتاح کتابخانه در ماه آوریل نصب شد.

دومین پروژه مهم، ایجاد کتابخانه سیار اینترنتی (1) بود که برای نمایش چگونگی ترکیب اسکن های الکترونیکی کتاب ها، چاپ بر اساس فناوری مورد تقاضا، و اینکه ارتباط شبکه ای ماهواره ای چگونه می تواند مناسب کتابخانه ای هزار جلدی در پشت یک ون باشد، طراحی شد پروژه یک میلیون کتاب (Million Books Project (2) ، در تابستان 2002 ، با شراکت کارنگی ملون شروع شد. هدف این بود که حداقل یک میلیون کتاب دیجیتالی شود و به صورت رایگان روی اینترنت قرار گیرد. با تشویق سوزان مبارک (3) ، آرشیو، ساخت یک کتابخانه سیار را در ایالات متحده شروع کرد و با همکاری دیگران، در هند و کنیا، نیز نمونه هایی از آن ایجاد شد.

سومین پروژه مهم مربوط به سیاست گذاری بود. 30 سپتامبر 2002، در تلاشی برای افزایش آگاهی عمومی درباره اهمیت موارد راهکاری حق مؤلف و کتابخانه سیّاراینترنتی، آرشیو به سفری در سر تاسر کشور و چاپ و توزیع کتاب های رایگان مبادرت کرد کتابخانه سیار را در حیاط ساختمان دادگاه عالی ایالات متحده پارک کرد و به چاپ کتاب پرداخت؛ جایی که قضات، در 9 اکتبر مباحثه میان الدرد (4) و اشکرافت (5) را شنیدند این حادثه مهمی بود که بر اساس آن تصمیم گرفته شد چه تعداد کتاب باید بخشی از کتابخانه دیجیتالی کتابخانه سیار و سایر کتابخانه های دیجیتالی در ایالات متحده باشد. متأسفانه، الدرد شکست خورد و تعمیم حق مؤلف مؤثر واقع افتاد؛ ولی پروژه کتابخانه سیار شکوفا شد و در نهایت، به صورت کتابخانه ای غیر انتفاعی درآمد. اخیراً، دولت هند، ساخت 25 کتابخانه سیار برای استفاده در سر تاسر کشور هند را شروع کرده است.

چهارمین حوزه فعالیت شامل ایجاد اولین آرشیو مجموعه کتاب و موسیقی است در ژوئن اولین مجموعه های کتاب، به صورت پیوسته در اختیار گذاشته شد در آگوست آرشیو موسیقی زنده، شامل مجموعه ای از اجراهای کنسرت – که به طور قانونی قابل بارگیری بودند – به صورت پیوسته درآمد.

پنجمین پروژه مهم، تأسیس کتابخانه دیجیتالی بین المللی کودکان (6) ، با شرکت دانشگاه مری لند (7) ، بود که از سوی کتابخانه کنگره، NSF، IMLS بنیاد کال / اوستین (8) ، شرکت سیستم های ادوب (9) ، بنیاد مرکل (10) ، و اکتاوو (11) پشتیبانی شد. ICDL، بر ماهیت ذاتی اینترنت، مبتنی بر فراهم آوری دسترسی مستقیم و جهانی به محتوای کیفی برای کودکان متمرکز بوده و هست.

1- Internet Bookmobile

2- MBP)،

3- Suzanne Mubarak

4- Eldred

5- Ashcroft

6- International Children's Digital Library

7- University of Maryland

8- Kahle/Austin Foundation

.Adobe System Inc –9

Markle Foundation –10

Octavo –11

در پایان سال 2002، آرشیو، برای مراقبت از یکپارچگی آرشیوهای دیجیتالی از طریق استاندارد سازی معیارها، و جلوگیری از جابه جا یا غیر قابل دسترس شدن مواد تلاش کرد. در نشستی در یوسی برکلی، نمایندگانی از آرشیو با سایر کتابداران دیجیتالی، به منظور تکمیل Okland Archive Policy ملاقات کردند که روند انتقال مواد را بر اساس قانون یا بر اساس خواست مالکان سایت و سایرین با جزئیات بیان می کرد.

2003: گسترش دستیابی ما به کتابخانه های ملی و مؤسسات آموزشی

در سال 2003 ، آرشیو، رسیدن به کتابخانه های ملّی و مؤسسات آموزشی در سرتاسر دنیا را ادامه داد. آرشیو، با همکاری کنسرسیوم حفاظت بین المللی اینترنتی (IIPC) از نزدیک با سازمان های شریک روی استانداردهای جدید و یک خزشگر جدید منبع باز شروع به کار کردن کرد.

در جولای 2003 ، آرشیو، به کنسرسیوم حفاظت بین المللی اینترنتی برای آغاز به کار کمک کرد. گروهی متشکل از 12 کتابخانه ملی برای توسعه استانداردها، ابزار و راهکارهایی برای تهیه حفاظت و دسترس پذیری دانش و اطلاعات از اینترنت برای نسل های آینده در همه جا، ارتقای تبادل جهانی و ارتباطات بین المللی تلاش می کردند. برای انجام این رسالت، IIPC برای رسیدن به اهداف زیر کار می کند:

– رسیدن به مجموعه ای غنی از محتوای اینترنتی از سرتاسر دنیا که به گونه ای حفاظت شوند که بتوانند آرشیو و محافظت شده و در لحظه قابل دسترسی باشند؛

– توسعه و استفاده از ابزار عادی فناوری ها و استانداردهایی که ایجاد آرشیوهای اینترنتی را ممکن می سازند؛

– تشویق و حمایت از کتابخانه های ملی در همه جا برای آرشیو اینترنتی و حفاظت.

IPC در کتابخانه ملی فرانسه با 12 مؤسسه همکار مجاز شروع به کار کرد. اعضا موافقت کردند به اتفاق هزینه ها را پرداخت کرده و در برنامه ها و کارگروه ها برای رسیدن به اهداف مذکور شرکت کنند. موافقت اوّلیه برای 3 سال بود. طی این پروژه اعضاء محدود به مؤسسات مجاز بودند.

در سال 2003 ، آرشیو بودجه قابل توجهی از خارج یعنی از سازمان های دیگر از جمله بنیادهای هیولیت واسلون (1) دریافت کرد و شروع به کار بر مجموعه های خاص کرد. کاهش های بعدی هزینه ذخیره روی دیسک و پهنای باند اینترنت آرشیو را به عرضه دائم «پهنای باند نامحدود، برای همیشه و رایگان» برای سازمان ها و افراد با مواد دیجیتالی، راهبر شد.

این امر به شراکت با Etree منجر شد. سازمانی که داوطلبانه در سال 1998 ایجاد شد تا تجارت آزاد و قانونی کنسرت های موسیقی زنده را امکان پذیر سازد. حاصل همکاری با Etree این شد که آرشیو e اکنون میزبان بیش از 15000 کنسرت موسیقی زنده است.

آرشیو، برای حمایت از نیازهای رو به رشد هم ذخیره و هم پهنای باند، یک مرکز جدید در سانفرانسیسکو باز کرد. این مرکز داده جدید، از طریق پیوند 1 Gbps به اینترنت متصل است و بیش از 1500 رایانه شخصی را که از لینوکس استفاده می کنند، میزبانی می کند.

2004: آرشیو اروپا و پتاباکس

اشاره

*آرشیو اروپا و پتاباکس (1) (2)

آرشیو، در سال 2004 شـروع به انتقال داده ها به سومین نسل سـخت افزاری خود، موسوم به پتاباکس، کرد. طرح پتاباکس، بر پایه سخت افزاری rack-mounted و سیسـتم عامل لینوکس، ذخیره RAID برای هر ترابـایت به مبلغ تقریبی 2000 دلار یا 2 میلیون دلار برای هر پتابایت پیشنهاد کرد.

نخستین نصب این طرح جدید در آمستردام در آرشـیو تازه تشـکیل شده اروپا بود، مؤسسه ای که قرار بود پاسخگوی نیازهای جامعه اروپا باشـد و آرشـیو اینترنت بـا سایر شـرکای اروپایی در اوّلین سال تأسـیس، آن را حمایت می کرد. نصب آن در آمستردام برای ایجاد قرینه ای برای مجموعه های اسـکندریه و سانفرانسیسـکو است. ایجاد شـبکه ای از مؤسـسات مسـتقل در سرتاسر دنیا که هر یک قادر است به طور مستقل عمل کند به پیشگیری از نابودی های فاجعه آمیز اطلاعات کمک خواهد کرد.

همچنین، در سال 2004، کنسرسیوم حفاظت بین المللی اینترنتی (3) Heritrix)، یعنی خزشگر وبی منبع باز، قابل توسعه با قابلیت تغییر اندازه وبی، با کیفیت آرشیوی و بر پایه جاوا را شروع کرد.

در مسیر توسعه مجموعه در سال 2004 ، از طریق اسـتخدام کارکنان بیشتر، تکمیل پروژه های اسکن کتاب و فیلم و دریافت داده از سایر مؤسسه ها، گام های مهمی رو به جلو برداشته شد.

آینده

پیشـرفت در رایـانه و ارتباطات این امکان را فراهم ساخته که ذخیره کتاب، صـفحه موسـیقی، فیلم، بسـته های نرم افزاری و تمام صـفحات عمومی وب که تاکنون ساخته شده اند، هزینه – سودمندی داشـته باشـند و دسترسی به این مجموعه ها از طریق اینترنت برای همه از جوان تا پیر در سرتاسر دنیا فراهم شود.

همانگونه که در اعلامیه حقوق بشر، ماده 19 آمده است: «هر انسانی حق آزادی بیان و عقیده را دارد. این حق شامل آزادی داشـتن باور و عقیـده بـدون نگرانی از مداخله و مزاحمت و حق جست و جو، دریافت و انتشار اطلاعات و نظرها از طریق هر رسانه ای بدون ملاحظات مرزی است».

برای سال های آینده، رسالت آرشیو مشخص است : ایجاد کتابخانه ای جهانی که تمام دانش را به سهولت در اختیار هر مرد و زن و کودک در سرتاسر جهان قرار می دهد.

منابع

Burner, M. Kahle, B. (1996). The ARC File Format

Gray, J. Shenoy, P. (1999). Rules of Thumb in Data Engineering. Microsoft

Technical Report, MS-TR-99-100

Masanès, J. (2006). Web archiving: issues and methods. In J. Masanès (Ed.), Web Archiving. Springer, Berlin Heidelberg New York

Ubois, J. (2002). The Oakland Archive Policy. Recommendations for Managing Removal Requests and Preserving Archival Integrity

---

European Archive -1

Petabox -2

Heretrix -3

چکیده

پیشـرفت در رایانه و ارتباطات این امکان را فراهم ساخته که ذخیره کتاب، صـفحه موسـیقی، فیلم، بسته های نرم افزاری و تمام صفحه های عمومی وب که تاکنون ساخته شده اند، هزینه – سودمندی داشـته باشند و دسترسی به این مجموعه ها از طریق اینترنت برای همه، از جوان تـا پیر، در سـر تاسـر دنیـا فراهم شود. برای سال های آینـده، رسالت آرشـیو مشـخص است: ایجاد کتابخانه ای جهانی که تمام دانش را به سهولت در اختیار هر مرد و زن و کودک در سـر تاسـر جهان قرار دهد. در مقاله حاضـر بحث آرشـیو اینترنت و آرشیو در محیط اینترنت ، دسترس پذیری دراز مدت تمام دانش برای همه افراد جهان پیشـرفت های فناورانه ایجاد آرشـیو اینترنتی مورد بررسـی قرار می گیرد. سپس برنامه ها و پروژه های نمونه آرشیو اینترنتی در دنیا آرشیو اروپا و پتاباکس معرفی می شوند.

ص: 74

کاربرد وب و مطالعات مربوط به آن

اشاره

*کاربرد وب و مطالعات مربوط به آن (1)

استیو جونز (2) | کَمیل جانسون (3) (دانشگاه ایلینویز شیکاگو)

ترجمه: دکتر سید مهدی طاهری (4) | سید محمد موسوی (5)

خلاصه

اشاره

در سال 2002، مرکز کتابخانه رایانه ای پیوسته (OCLC) برآورد کرد که بیش از سه میلیون وبگاه در شبکه وب جهان گستر در دسترس عموم قرار دارد (اونیل و همکاران 2003، پاراگراف 9). در دنیای مادی اطلاعات، این تعداد وبگاه نزدیک به 14 تا 28 میلیون خواهد بود که با حجم کتاب های موجود در بزرگ ترین کتابخانه های دنیا برابری یا حتی از تعداد آن تجاوز می کند. این مجموعه گسترده اطلاعات، مجموعه ای بی پایان و دسترس پذیر از داده های دیداری و شنیداری را در اختیار پژوهشگران علاقه مند به مطالعه در حوزه فعالیت های پیوسته قرار می دهد. حجم خالص منابع دسترس پذیر بر روی وب همواره چالش هایی را به وجود آورده است چه چیزی را انتخاب و چگونه مطالعه کنیم؛ با این حال، شبکه وب ثابت کرده است این حجم اطلاعات برای کاربران دلهره آورتر هم شده است. در این مقاله، برخی رویکردهای روش شناختی را که پژوهشگران برای مطالعه در حوزه وب به کار گرفته اند، مرور خواهیم کرد.

1- Web Use and Web Studies: in Masanes, Julien (ed.), Web Archiving. Berlin Heidelberg New York: Springer. pp.55-67.

2- Steve Jones

3- Camille Johnson

4- عضو هیئت علمی پژوهشگاه علوم و فرهنگ اسلامی

5- کارشناس ارشد کتابداری و اطلاع رسانی پژوهشگاه علوم و فرهنگ اسلامی

هـدف از این کار، طبقه بنـدی جامع روش شناسـی ها نیست به جای آن امیـدواریم با درک روش های به کار گرفته شده در مطالعه وب و مطالعه کاربرد آن، افرادی که درصدد آرشیو کردن و نگهداری وب هستند بتوانند نیازهای جامعه علمی را هرچه بهتر درک کنند.

شبکه وب عبـارت است از طیف گسترده ای از انواع مواد که تنوع آن هـا را بـا دو بُعـد می توان بهتر درک کرد. نخست اینکه خود وب یک رسـانه است نه محتوا برای روشن تر شـدن مطلب بایـد گفت که وب هم رسـانه ای است که محتوا را از طریق پروتکـل های متعـدد (نظیر HTTP) منتقـل می کنـد و هم «ظرفی» برای محتواست که محتوا را شـکل داده و به کـاربران خود ارائه می دهـد. هر چنـد ارائه محتوا به تشـخیص کاربر وابسته است و گذشته از آن توسط ابزارهای دیداری مورد استفاده (مرورگر ها و برنامه های کاربردی دیگر) شکل می گیرد. به عبـارت دیگر، اگر چه محتوا ممکن است صـفحه وب یکسانی باشـد، دو کاربری که از مرور گر های متفاوت، یا از تنظیمات متفاوتی از یـک مرورگر اسـتفاده می کننـد، ممکن است در نهـایت صـفحه هـای متفـاوتی را ببیننـد. دیگر اینکه، بر خلاف دنیای مواد آنالوگ، همان تعریف رسـانه برای ذخیره سـازی وب مورد تردیـد است. وبگـاه هـا را می توان به صورت محلی ذخیره یـا پنهـان کرد، اطلاعات آن ها را منعکس نمود، یا ممکن است مجازی یا موقتی باشند همان طور که این ویژگی ها برای وب کم ها نیز صادق است.

به طور کلی، پژوهش درباره اینترنت و وب، توسط پژوهشـگرانی با تخصـص های مختلف از جمله متخصـصان رشـته های زبان شناسی، روزنامه نگاری، علوم سیاسـی، مدیریت بازرگانی، جغرافیا، تبلیغات، ارتباطات، هنر، و دیگر رشـته ها انجام می شود. از آن جا که با طیف وسـیعی از سـنت های آموزشی سر و کار داریم، انواع موادی که پژوهشگران از آن ها بهره می گیرند بسیار متنوع است و به همان اندازه نیز در روش های پژوهش تنوع مشاهده می شود. آن چه به عنوان تحلیل «قوم نگاری» متون وب، برای پژوهشـگر بازاریابی به حساب می آید، ممکن است از دیـدگاه پژوهشـگر رشـته ارتباطـات تحلیل محتوای کیفی تعبیر گردد. بنابراین هـدف دیگر این مقاله نه تنها ارائه مقوله های ثابت محتوای وب یا طرح هایی برای رویکردهای روش شـناختی است – آن گونه که برای مطالعات وب به کار می روند – بلکه ارائه طیفی از تعابیر و کاربردهای این روش ها نیز مورد نظر است، به نحوی که مفیدترین روش ها را برای پژوهشگران حوزه وب به اثبات رساند.

1-تحلیل محتوا

تحلیـل محتوا یکی از رایـج ترین روش هـای مطـالعه بر روی حوزه وب است. پژوهشـگری که از روش تحلیـل محتوا اسـتفاده می کنـد، محتوای وب را – هم محتوای متنی و هم تصـاویر – بر اسـاس معیارهای خاص رمزگـذاری کرده و آن ها را درون مقوله ها یا موضوع های مرتبط قرار می دهد؛ به عبارت دیگر، این نوع تحلیل بیشتر بررسی محتوای وب است تا کاربران وب. تحلیل محتوا در میان مطالعات وب، به عنوان ابزار مقـایسه مورد اسـتفاده قرار گرفته است به نحوی که این ابزار به پژوهشـگر اجازه می دهـد تا مقایسه های معنی داری از محتوا میان متون مشابه وب انجام دهد. در بررسی وبگاه های سازمان های ضد جهانی سازی

از روش تحلیـل محتوا اسـتفاده شـد تـا نشـان داده شود که آیا میان این سایت ها انسـجام پیام و مقصود وجود دارد (وَن الست و ،والگریو 2002) ؟ محتوا، بر اساس کارکرد اصـلی خود در چهار حوزه رمزگـذاری و مرتب شـده است ارائه اطلاعات درباره سازمان مسائل ضـد جهـانی سـازی افزایش تعامل با مؤسـسان گروه و اعضای دیگر و افزایش بسـیج افراد در مورد مسائلی چون اهـدای پول که در یک فراخوان اینترنتی اعلاـم شـده است وَن الست و والگریو، از طریق تحلیل خود به این نتیجه رسـیدند که در حقیقت سازمان های ضـدجهانی سازی پیوسته، عموماً به وسیله وبگاه های خود و از روش های مشابهی برای آگاه نمودن و مشغول کردن اعضای خود استفاده می کردند.

مطـالعه تطبیقی دیگری به تحلیـل محتوای وبگاه های مربوط به ایسـتگاه های رادیویی پرداخت، تا چگونگی پاسـخگویی صـنعت تجاری رادیو را به انـدازه دسترس پـذیری وب جهـان گستر برای ارتقـای ایسـتگاه هـای خود مشـخص کنـد. این ارزیـابی بـا تجزیه و تحلیل انواع اطلاعـات کـاربر – محور، که در وبگـاه های ایسـتگاه هایی رادیویی فراهم شـده بود برای مثال نقشه های ترافیکی و لاگ های برنامه ها)، اسـتفاده از وبگاه ها به عنوان ابزارهایی برای بهبود ایستگاه خود (مانند اطلاعات رقابتی و بایوس های DG.)، و ترکیب ویژگی های تعاملی آن هـا (نظیر نشانی ایمیل ها برای اعضای ایسـتگاه ها و نظرسـنجی های شـنوندگان) به دست آمـد (پیتز و هارمز 2003). این مطالعه در حوزه وب منبع اطلاعاتی ارزشمندی را برای ارزیابی کاربردهای کنونی و بالقوه وب در اختیار ایستگاه های رادیویی قرار داد.

همچنین، روش تحلیـل محتوا، برای مطـالعه اثرات تغییرات خـط مشـی های سازمانی بر وبگاه های مرتبط به کار گرفته شـده است. برای مثال، خط مشی های آموزشی، اهمیت الحاق فناوری اطلاعات و ارتباطات (ICT) به مأموریت های سازمانی مدارس راهنمایی را افزایش دادنـد ،مطالعه ای، پیشـرفت های صورت گرفته برای نیل به این هـدف را همخوان با طرح ریزی انجام شـده توسط شـبکه ملی یادگیری در بریتانیـا از طریق تحلیـل محتوا در وبگاه های 150 مـدرسه راهنمایی مورد ارزیابی قرار داد (هسـکت و سولیوان 1999). تصاویر و متون رمزگذاری شدند، و به شـناسایی وبگاه ها در یکی از مقوله های روشـی زیر پرداختند: فعال، دانش آموز–محور، و منفعل. پژوهشـگران از طریق این مقوله هـا و تحلیل پرونـده های پژوهشـی مـدارس، رابطه ای را بین تعهـد به یکپارچگی فناوری اطلاعات و ارتباطات (ICT) و سرمـایه های سازمانی درک نمودنـد؛ هر قـدر سـرمایه اقتصادی و اجتماعی مـدرسه بیشتر باشـد احتمال اینکه محتوای وبگاه آن منعکس کننده گرایش های مثبت و فعال الحاق فناوری اطلاعات و ارتباطات با مأموریت های مدرسه باشد بیشتر خواهد بود.

کمیسـیون بازرگـانی فـدرال امریکا، از روش تحلیل محتوا، برای ارزیابی پیاده سازی برنامه های حفظ امنیت اطلاعات و حریم خصوصـی پیوسـته بر روی وبگاه های بازرگانی استفاده کرده است (Milne and Culnan). مطالعه انجام شده از سوی این کمیسـیون، دارای تحلیل دراز مـدت چهـار نظر سـنجی وب با گستره زمانی 1998 تا 2001 بود از معیارهـایی نظیر حفـظ یـا حـذف اعلان ها، با توجه به افشای اطلاعـات بازدیدکننـدگان بـا اشـخاص ثـالث و اسـتفاده از «کوکی هـا»، مجموعه ای از اطلاعـات کاربران به غیر از نشانی ایمیل ها، و به کارگیری امنیت اطلاعاتی در سایت ها مورد استفاده قرار گرفت تا مشخص شود که آیا

وبگاه ها ملزومات شیوه های استفاده بی طرفانه از اطلاعات را تأمین می نمایند؟ مقایسه نتایج به دست آمده از تحلیل انجام شده در سال 2001 با تحلیل های سال های پیشین نشان می دهد، در میان وبگاه های تجاری که به گردآوری اطلاعات شخصی کاربران خود دست می زنند، روند افزایشی در اجرای شیوه های استفاده بی طرفانه از اطلاعات وجود داشته است.

2-بررسی ها

بررسی های مربوط به وب را می توان به دو روش متمایز تعریف کرد: بررسی هایی که بر روی وب انجام می گیرد، و بررسی های منتشر شده در وب. بررسی های نوع نخست که از تداول بیشتری برخوردار است به وسیله پژوهشگرانی انجام شده است که مطالعه درباره وبگاه ها را به منظور دسترس پذیری برای جمعیت علاقه مند ارسال می کنند. در بیشتر موارد، هدف از این نوع بررسی ها، گردآوری اطلاعاتی درباره موارد استفاده کاربران از اینترنت (1) است (مانند بررسی کاربران وب که در مرکز گرافیک، دیداری سازی، و کاربرد پذیری (The Graphic, Visualization, and Usability Center (2) انجام گرفته است (GVU ، 1998). از سال 1994، مطالعات مداومی درباره کاربرد اینترنت انجام داده است. این مرکز، شرکت کنندگانی را از طریق ثبت گروه های خبری مرتبط با اینترنت ،تابلوهای تبلیغاتی در رسانه های خبری، وبگاه های موتورهای کاوش و اطلاعیه های موجود در رسانه های ناپیوسته نظیر مجلات و روزنامه ها به کار می گیرد. چندین بررسی متمرکز توسط GVU از سال 1994 تولید شد که از آن میان می توان به بررسی سرشماری عمومی، بررسی مرتبط با سرشماری فناوری- که اطلاعاتی را نظیر سرعت ارتباط کاربران و مرورگر مورد استفاده گردآوری کرد، بررسی مربوط به گرایش کاربران به سوی حریم خصوصی و امنیت اینترنتی، بررسی عمومی رایانه ها، کاربرد اینترنت و وب، و نیز بررسی جست و جوی محصولات و فعالیت های مربوط به خرید به صورت پیوسته اشاره کرد. جامعه آماری، معمولاً بالای 1000 [نفر]بودند که بررسی مربوط به سرشماری عمومی مشتمل بر 5000 شرکت کننده بود.

سازمان ها نیز بررسی های مربوط به وب را ابزار مفیدی برای ارزیابی کیفیت تجربیات کاربرانی می دانستند که از وبگاه ها یا کارکرد های وب بهره می برند. چنین مطالعه ای، بررسی بازدیدکنندگان حدود 450 وبگاه موزه ها را هدایت می کند به نحوی که هم اطلاعات مربوط به سرشماری، و هم پاسخ های مربوط به کیفیت و قابلیت استفاده از این وبگاه ها را گردآوری می نماید (Sarraf 1999). بسیاری از موزه ها به تدریج وب گاه هایی را تهیه کردند تا ابزاری برای ارائه اطلاعات درباره مجموعه های موجود در موزه ها و نیز ایجاد ارتباطی تعاملی و قابل دسترس برای بازدیدکنندگان کنونی و پیشین باشد.

1- با دسترسی 59 درصد از افراد بالغ آمریکایی به اینترنت در سال 2002، پژوهشگران در آمریکا نیز بررسی های مبتنی بر وب را به جای انواع بررسی های سنتی از قبیل بررسی های تلفنی و کاغذی درباره موضوعاتی به جز کاربردهای اینترنت و نگرش های موجود مد نظر قرار دادند به نظر می رسد پژوهشگران علاقه مند به گردآوری هنگام واکنش ها نسبت به وقایع اخیر به ویژه به بررسی های اینترنتی هستند، به نحوی که این علاقه مندی در بررسی های انجام گرفته درباره اثرات حملات تروریستی نیویورک در 11 سپتامبر سال 2001 بر شرکت کنندگان مشهود است (برای مثال لی و همکاران 2003) درست 9 روز پس از این حملات، گروهی از پژوهشگران توانستند بررسی از طریق ارزیابی وب-پایه پاسخ های روان شناختی آمریکایی ها نسبت به این واقعه تلخ را توزیع کنند.

GVU)، -2

به منظور ارائه بازخوردی جامع برای موزه های مورد بحث به پاسخ های کیفی و کمی نیاز بود. کارزول و ونکاتش (2002)، با استفاده از بررسی وب محور توانستند پاسخ های ارزیابی بیش از 500 دانشجوی مقطع کارشناسی را که در کلاس های پیوسته غیر همزمان شرکت داشتند درخواست کنند. آن ها حمایتی نسبی را در مورد فرضیه های خود یافتند مبنی بر اینکه پذیرش، و در آینده قصد استفاده از فناوری در دوره ای پیوسته و غیر همزمان می تواند به صورت مثبت متأثر از گرایش ها و درک دانشجویان از فناوری باشد. در تمامی موارد از جمله GVU، مزایای استفاده از بررسی مبتنی بر وب آشکار بود: جمعیت های علاقه مند در بررسی ها همان کاربران وب بودند، و داده های گردآوری شده با فعالیت های وب ارتباط مستقیم داشت.

3-تحلیل بلاغی

*تحلیل بلاغی (1)

زیر مجموعه دیگری از پژوهش در حوزه وب نگاهی جدی به متون وب دارد تا راهکارهای متقاعد کننده ای را از طریق تحلیل بلاغی مشخص کند. همان طور که وارنیک توصیف می کند:

روش انتقادی بلاغی این مسئله را در نظر دارد که چگونه متن به برخی عناصر موجودیت می بخشد- در حالی که با برخی دیگر مخالفت می کند، چگونه ساختارهای گزارشی تجربه مرورگر را به روش های مشخص پیکربندی می کنند، و اینکه چگونه گفتمان، تمایلات و عادات ذهنی مخاطب خود را نقش می دهد.

وی، در بررسی خود به وبگاه های سیاسی در جریان مبارزات انتخاباتی رئیس جمهوری آمریکا در سال 1996 به ویژه راهکارهای بلاغی وبگاه های مقلد پرداخت – وبگاه هایی که از طراحی و محتوای وبگاه های قانونی مربوط به مبارزات انتخاباتی تقلید می کردند تا در صورت امکان از اعتبار آن سایت ها بکاهند و تا حدی آن ها را مورد تمسخر قرار دهند. بر اساس نتیجه گیری ،وارنیک چنین سایت هایی بر گزارش های نادرست در دولت، تحریک مشارکت سیاسی از طریق ویژگی های تعاملی - نظیر شکایاتی که هرگز به حزب مورد نظرشان تسلیم نشد و اشاراتی درباره سوابق سیاسی و جنایی - که نامزدها بر آن ها متکی هستند تا خوانندگان شان را به پذیرش بینش بلاغی خود مجاب کنند - تکیه می کنند. در نهایت، وارنیک، این راهکارها را ریاکارانه می داند؛ زیرا از رفتارهای غیر اخلاقی مشابه بسیاری که ظاهراً از سوی نامزدهای مورد نظر به کار گرفته می شود بهره می گیرند.

از تحلیل بلاغی وبگاه هایی که برای شناسایی اجتماعات بلاغی چنین استفاده می گردد که گروه هایی از وب برای شکل گیری جهان بینی خود استفاده می کنند که ممکن است همسو با جریانات اخیر باشد یا نباشد کروبر، به بحث و تبادل نظر در این مسئله می پردازد که گروهی از وب گاه های شکل گرفته توسط مادرانی با گرایش های فمنیستی، خط مشی [خود] را به مخاطبان القا می کند تا به طرز مؤثری در برابر تصورات منفی فمینیستی در رابطه با مادر بودن مقاومت ورزند. زنان، توانستند از طریق متون و تصاویر موجود در وب گاه های شان دوباره به بحث و تبادل نظر در خصوص فمینیسم بپردازند و برای درک معنای مادری به عنوان بخش پر نفوذ و حتی لازم فمینیست بودن به بحث و جدل می پردازند.

بررسی دیگر، نگاهی به موارد استفاده از وبگاه ها از سوی گروه های نفرت دارد، مانند شوالیه های

Rhetorical analysis -1

کوکلاکس کلان (1) و اتحاد ملی سازمان نئونازیسم که مقاصد ترغیبی را دنبال می کردند (Duffy 2003). در این مورد، دافی، از روش انتقادی بلاغی به نام تحلیل موضوعی فانتزی استفاده کرد که بر «نظریه همگرایی نمادین» استوار است. همانطور که دافی مشخص کرد، نظریه همگرایی نمادین «نظریه کلی علم معانی است که در آن، گروه ها تصوراتی را درباره گروه و گروه های خارج شکل می دهند و به اشتراک می گذارند، و در نتیجه هویت مشترکی را به وجود می آورند» (صفحه 293) (2) نویسنده، به وسیله تحلیل متون برگزیده موجود در در وبگاه هر گروه توانست گزارش های بلاغی متعددی را شناسایی کند که «ادعای انصاف و عدالت»، «نظم طبیعی و رستاخیز انسان» و «ساکنان اولیه زمین نامیده شدند تا مفهوم جدیدی را به نژاد انتقال دهد» از آن جمله هستند (صفحه 295-305).

4-تحلیل گفتمان

*تحلیل گفتمان (3)

مک کوئیل (2000)، تحلیل گفتمان را استعمال آن «در تمامی اَشکال کاربردی زبان و اَشکال نوشتاری» می داند، با استناد بر این طرز تفکر که «ارتباط در قالب نوشتار و گفتار رخ می دهد که با موقعیت اجتماعی خاص، موضوع ها، و انواع شرکت کنندگان سازگار می شود» (صفحه 494). تحلیل گفتمانی متون وب، موقعیت اجتماعی - فرهنگی وبگاه ها را در نظر می گیرد؛ یعنی به ساختار معنی محلی آن ها از طریق عناصر گفتاری و دیداری می پردازد. دگربار، متون وبِ متعلق به گروه های نفرت ظاهر می شوند؛ چنان که بیلینگ (2001) قراردادهای زبانی طنز را همانطور که از سوی وبگاه های «طنز» متعلق به گروه کوکلوکس کلان دیده می شود کشف کرد. او پی برد که در این وبگاه ها از تکذیب کنندگان برای هشدار علیه تفسیر طنزهای نژادپرستانه استفاده می کردند این طنزهای نژاد پرستانه به اقدامات خشونت آمیز فرمان می دادند در حالی که محتوای آن ها برای افراد «شوخ طبع» تعبیری طنزگونه داشت (صفحه 274). هر چند نویسنده به این نتیجه رسید که بخش عمده محتوای اینگونه سایت ها طنزگونه نبوده بلکه از جنس حقیقت است و فراگفتمانی را که او مطرح می سازد «انکار می کند که طنز طنز است» (صفحه 278). متون و تصاویر نژادپرستانه که هم خوانندگان این «طنزها» با گرایش های نژادپرستانه و هم مقاصد اخلاقی اینگونه طنزها را مخاطب قرار می داد شواهدی از گفتمان اهریمنی تری را فراهم می نمود، گفتمانی که بررسی بیلینگ را به تعاملات اجتماعی اوباش لینچ (4) تشبیه می کند (صفحه 287).

گفتمان در سطح کلان نیز در وب مورد تجزیه و تحلیل قرار گرفته است، و تأکید آن بر وبگاه هایی است که صدای دولت و ملت ها هستند. در بررسی پورسل و کورداس از وبگاه دولت اسلوونی، آن ها با متون موجود در این وبگاه به عنوان واکنشی به خطاهای مشاهده شده در معرفی اسلوونی به عنوان دولت بالکان - منطقه ای که گرفتار جنگ داخلی شد و بنابراین از سوی دولت های غربی به عنوان دولتی

1- Ku Klux Klan

2- فاس (1996)، نظریه همگرایی نمادین را بر اساس دو ادعای اصلی توصیف می کند: «ارتباط واقعیت را می سازد» و « نه تنها واقعیت را برای افراد می سازد بلکه معانی افراد را از نشانه ها می تواند با هم یکی کند تا واقعیت مشترکی را برای شرکت کنندگان به وجود آورند» (صفحه 122).

3- Discourse analysis

نامطلوب به خاطر اقدامات تروریستی و سرمایه گذاری های تجاری دیگر محکوم شد – برخورد کردند. پورسل و کورداس اظهار می دارند که اسلوونی تلاش می کند تا هویت از دست رفته خود را از طریق متون و تصاویر در وبگاه دولتی احیا کند که نوعی راهکار بلاغی محسوب می شود. اما بر اساس نتیجه گیری آن ها این سایت، سرانجام بخشی از تلاشی مستمر برای مذاکره موقعیت [اسلوونی]» در «گفتمان جهانی» است.

شندلر (1998)، بیان کرده است که علاوه بر وبگاه های سازمانی، صفحه های شخصی نیز ممکن است شکلی از گفتمان باشند. شندلر در تحلیل خود از متون وب تولید شده توسط نوجوانان ولز از طریق مصاحبه با سازندگان سایت ها، روش های بسیاری را بررسی کرد که بیان گر تصور نوجوانان از مخاطبان شان بود و درباره مرز بین فضاهای عمومی و خصوصی زندگی شان صحبت می کردند. بسیاری از نویسندگان جوان وب بیان کردند که سعی دارند تا در سایت های خود با مخاطبانی که احتمالا علاقه های خود را به اشتراک می گذارند ارتباط برقرار کنند در حالی که افراد دیگر وبگاه را چیزی توصیف می کنند که صرفاً برای خودشان ایجاد کرده اند برخی ،دیگر انگیزه ای پیدا می کنند تا بخشی از این اینترنتِ «پهناور» باشند و عقایدشان را درباره زندگی با دیگران به اشتراک بگذارند.

5-تحلیل دیداری

*تحلیل دیداری (1)

به دلیل قابلیت های چندرسانه ای شبکه جهانی وب، برخی پژوهشگران در صدد برآمده اند تا تحلیل های سنتی از متون وب گفتاری را با تحلیل متون دیداری تکمیل کنند یا بر آن ها غلبه کنند. ایجادکنندگان وبگاه ها نیز از گرافیک ها و تصاویر خلاقانه فراوان با چاشنی سلیقه در وبگاه ها بهره می گیرند، مانند آن چه در JenniCam.org به کار رفته است. از ویژگی های این وبگاه شخصی دوربین 24 ساعته ای با چشم انداز اتاق خواب جنی رینگلی نویسنده وب است. جیمروگلو (1999)، از سایت جنی کَم (2) به عنوان بررسی موردی برای کشف مفهوم علم فرمانشی ترکیبی از ماشین و ارگانیسم هاراوی (3) استفاده می کند (صفحه 149). همانند سایبورگ جیمرو گلو مشاهده کرد که مرزهای بین بدنه جنی و فناوری از طریق دوربین پیوسته او محو می شود، مثلاً تصویر جنی که در مقابل صفحه نمایش رایانه اش نشسته است. سایت وبٌ کَم او تصویر جنی در پشت رایانه اش به نمادی برای آن تبدیل شده است، به طوری که جسمش ذوب و تبدیل به صفحه کلید می شود» (صفحه 441). برای بررسی معنای جنی کَم، به عنوان موضوعی جنسیتی و دیداری از نظریه فمینیستی فیلم استفاده شد. تعهد رینگلی در ارائه نگاهی اجمالی به زندگی واقعی او در برگیرنده لحظاتی بود که از جلوی وب کم به دلیل برهنگی یا داشتن فعالیت جنسی کنار می رفت. جیمر و گلو، مشکل منحصر به فرد این تصاویر را تشخیص داده است: جنی موفق شد تا مرزهای سنتی اختصاص یافته به اندام زن را بین فضاهای عمومی و خصوصی بشکند، حال آن که با انتقاد برخی فمینیست ها روبه رو شد. آن ها بر این باور هستند که جنی در دام عینیت بخشی اندام زن، که در وب شایع

1- Visual analysis

2- JenniCam

3- Haraway

است، گرفتار شـده است. در هر دو مورد «انـدام جنی به عنوان کانون معنا، محل کمال، و ریشه معنی واحـدی از جنی کم عمل می کند» (صفحه 449).

پژوهشـگران دیگر، از تحلیل دیداری برای ارزیابی کاربرد اینترنت از سوی رسانه های خبری برای ارائه و توزیع تصویر استفاده کرده اند. فِر و پارکز (2001)، تحلیـل انتقـادی را از تصاویر مربوط به بحران نسل کشـی روآنـدا در سال 1994 ارائه می دهنـد که رسـانه های خبری آمریکا از طریق تلویزیون و اینترنت در دسترس عموم قرار دادند. پوشـش زمینی این بحران، که در آن دوربین ها تصاویری را از پناهندگان و حشت زده گرفته بودنـد و امدادگران غربی سـفید پوست به آن ها یاری می رساندند، و نیز تصاویر ماهواره ای از مردمی که دسـته جمعی از روآنـدا می گریختنـد، با مخالفت روبه رو شدند، زیرا از قربانیان این بحران اختیار را سـلب کرده بود. [آن ها]موقعیت سیاسـی این واقعه را بیش از اندازه ساده انگاشـته بودند و بیننده را از مردم درگیر و مسائل موجود دور نگه می داشتند. فِر و پارکز، می افزایند که استفاده از وبگاه ها برای ارائه و پخش تصاویر روآنـدا بخشـی از رونـد گسترده تر سـلطه و مالکیت فناوری های دیداری از سوی رسانه های خبری در غرب است. در واقع، محتوای تولیـد شده به مصـرف مخاطبان غربی یعنی کاربران اصـلی اینترنت می رسد. فِر و پارکز، به این نتیجه رسـیدند که انتخـاب تصـاویر از سوی رسـانه های خبری در پوشـش خبری خود از بحران روآنـدا به همراه پخش گسترده تصاویر از طریق فناوری های رسانه ای غرب دست به دادند تا «آشفتگی» موجود امریکا را از طریق فرهنگ و سیاست آفریقا تقویت کنند (صفحه 42).

دو مطالعه یادشده، توجه خاصـی به تصاویر مورد استفاده در وبگاه ها، به منظور انتقال تفاسیر شان مبذول داشته اند. اما تحلیل های متون دیـداری و گفتـاری، جـدا از منحصـر به فرد بودن دو جـانبه شـان ، معمولاً توسط پژوهشـگران تلفیق می گردنـد. همان گونه که در مطالعه هسـکچ و سالیوان، درباره هویت بخشی مدارس از طریق تصاویر و متن ها در وبگاه ها، یا مطالعه و ارنیک درباره وبگاه های مقلد مربوط به مبـازرات انتخاباتی و اسـتفاده آن ها از متون گفتاری و غیر گفتاری برای پیشبرد دسـتور جلسات سیاسـی ویژه مشاهـده گردیـد. طراحی دیداری وبگاه ها، از قبیل اسـتفاده از تصاویر و متن ها، و گذاشـتن آن ها در صفحه وبگاه، نمونه دیگری از تلفیق این روش هاست (2000 Rivett). بررسی موردی وبگاه «نیوبیتل» ولکس واگِن (1) ، در مورد صـفحه آرایی برگزیده، را که در صفحه اصلی این سایت قرار داشت مورد بحث قرار داد، و نویسـنده سایت آن ها را با قراردادهای طراحی مجلات ناپیوسته مقایسه نمود. به علاوه، این سایت، با استفاده از یک پیش زمینه سفید، «هویت دیداری» پیوسته شرکت را منتقل می کند (صفحه 50). تحلیل دیداری وبگاه نیوبیتل مکمل تحلیل متنی است. در صورتی که داسـتان «حمله بیگانگان» در سرتاسـر صـفحه هایش به چشم می خورد. سپس، سایت ولکس واگِن با طرح دیداری سایت دانشـگاهی دانیل شندلر مقایسه گردید. سایت دانشگاهی دانیل شندلر، که مملو از متن بود بر هدف خود به عنوان منبع اطلاعات تأکید می کرد. در این سایت از تصویر کلاسوری به عنوان تصویر زمینه این سایت استفاده شد که بر ماهیت آموزشی آن تأکید می کند.

1- Volkswagen

6-قوم نگاری

*قوم نگاری (1)

قوم نگاری، اغلب به عنوان «توصیف و تفسیر گروه یا نظام فرهنگی یا اجتماعی» درک می شود (Creswell 1998P58). تعریف موضوع در پژوهش های قوم نگاری به فرآیند پیچیده ای تبدیل شده است. زیرا در خلال مطالعه فرهنگ های اینترنت و وب انجام می گیرد. جنبه های عملی قوم نگاری، پیش از اینترنت، مطالعه فرهنگ های تک محله ای بوده است که از لحاظ جغرافیایی متمرکز شده اند. با وجود این انسجام پیوسته و پراکندگی فیزیکی ناپیوسته شرکت کنندگان در فرهنگ های اینترنتی مستلزم روشی چند مکانه برای قوم نگاری است. مفهوم جامعه چند مکانه نیز باید در موقعیت های شرکت کنندگان جامعه پیوسته به کار برده شود و ارائه توصیفی تیره از جامعه پیوسته مستلزم روشی متنوع تر از مشاهده ساده گفتمان گروه درون اتاق گفت و گو یا تحلیل کردن محتوای وبگاه های مربوط به آن جامعه است. در عوض، برخی پژوهشگران قوم نگار وب استدلال کرده اند که نیازمند روشی کلی نگر تر هستند، روشی که پژوهشگر تمامی سایت های مربوط به مشارکت و تجربه اعضای جامعه پژوهش را هم به صورت پیوسته و هم ناپیوسته بررسی کند (Howard 2002, Miller 2000 Slater).

میلر و اسلَتر (2000) نمونه سودمندی را برای اجرای الگوی چند مکانه از قوم نگاری اینترنتی در مطالعه خود درباره اهالی ترینیداد (2) و اینترنت ارائه کردند. پژوهشگران تالارهای گفت و گو، وب گاه ها، و نیز خانه ها و کافی نت های مردم را بازدید کردند تا از روش های تلاقی اینترنت با زندگی های سیاسی، اقتصادی، و مذهبی «ترینی ها» آگاهی یابند. هر سایت، ارائه دهنده چشم انداز منحصر به فردی از آن فرهنگ لست. اتاق های گفت و گو مکان هایی هستند که بینشی را از «ترینی» بودن به صورت پیوسته به دست می دهند. وب گاه ها، به عنوان نمایندگان ترینیداد به صورت پیوسته بودند (صفحه 13). میلر و اسلَتر، تحلیل محتوایی کیفی را به کار گرفتند تا درباره وب گاه ها به بحث و گفت و گو بپردازند. در حالی که، تلفیق نمادهای ملی نظیر پرچم و تصاویر نقشه ها را که با هویت های شخصی نویسندگان وب سراسر سایت های تحلیل شده اند - متذکر می شوند. افزون بر آن، آن ها به بحث درباره بازتاب دستور جلسات سیاسی در وبگاه ها، با سود آور ترین جاذبه توریستی ترینیداد، یعنی جشنواره کارنیوال، پرداختند که به محور اصلی در این سایت ها تبدیل شده بودند.پژوهشگران علاقه مند به جوامع طرفدار اینترنت، قوم نگاری را روشی سودمند برای مطالعات شان می دانند. بلوستین (2002)، تحلیل های انجام گرفته از وب گاه های طرفدار و انجمن های مربوط را در ترسیم خود از تعصب، و نمایش تلویزیونی «بافی، قاتل خون آشام» گنجاند. او با استفاده از تحلیل های خود از متون وب و اطلاعات گردآوری شده از مصاحبه های شخصی و بازدید از اتاق های طرفداران توانست به نتیجه گیری هایی برسد که چگونه طرفداران بافی (که اغلب نوجوان بودند) بین جلوه فانتزی و جلوه های سحرآمیز نمایش تمایز قائل می شوند تا تخیل و «تفریح» جوانی را حفظ کنند، مادامی که به فضای جدی و ترسناک ابهام اخلاقی موجود در بزرگسالی وارد می شوند (صفحه 440).

1- Ethnography

2- ترینیداد بزرگ ترین جزیره کشور ترینیداد و توباگواست. این جزیره جنوبی ترین جزیره دریای کارائیب است که تنها 11 کیلومتر با کرانه های ونزوئلا فاصله دارد. ترینیداد 4،768 کیلومتر مربع مساحت دارد.

7-تحلیل شبکه

*تحلیل شبکه (1)

به عقیده گارتون و همکارانش (1997)، «شبکه اجتماعی، به مجموعه افراد (یا سازمان ها یا نهادهای اجتماعی دیگر) اطلاق می شود که از طریق مجموعه ای از روابط اجتماعی نظیر دوستی، همکاری یا تبادل اطلاعات به هم مرتبط می گردند» (پاراگراف 2). هنگامی که این روابط اجتماعی با استفاده یا ایجاد وبگاه ها بنا برقرار یا ایجاد می گردند، فراپیوندها وسیله ای را برای مطالعه الگوهای ارتباطی میان اعضای شبکه ارائه می دهند. پارک (2003)، این نوع تحلیل در پژوهش وب را تحلیل فراپیوند می داند که وبگاه ها را نمایان گر افراد، گروه ها، سازمان ها، و دولت ها و ملت ها، و فراپیوندهای بین سایت ها را «پیوند[های] رابطه ای» تصور می کنند (صص 50-51). برای مثال، هالاویس (2000)، به منظور درک بهتر این موضوع که آیا وب به درستی شبکه «جهان گستر» سایت ها را رواج می دهد یا اینکه مرزهای ملی در حال پیوستن به مرزهای اینترنت هستند؟، به مطالعه فراپیوندهای موجود در وب گاه ها پرداخت. او با مطالعه 4000 وبگاه به بررسی پیوندهای آن وبگاه ها با صفحه های خارج از آن وبگاه ها پرداخت که از هویت میزبان ملی وبگاه هایی که به آن ها پیوند می یابند، پیروی می نمایند. هالاویس، با استفاده از این فرآیند اینگونه نتیجه گرفت که در حقیقت بیشتر وبگاه های موجود مورد بررسی اصولاـ به سایت های درون فرهنگ های ملی مرتبط می شوند. با وجود این، در مقایسه با انواع دیگر فناوری های اطلاعاتی مانند پست زمینی و تلویزیون او بر این باور بود که اینترنت بین المللی ترین فناوری اطلاعاتی است، و نسبت به دیگر رسانه ها در داخل و خارج از آمریکا در سراسر مرزهای بین المللی از تعداد مراجعات بیشتری برخوردار است صفحه (23).

همان طور که پیشتر ذکر شد، تحلیل شبکه در سطح سازمانی، همزمان با تحلیل محتوایی در مطالعه وبگاه های ضدجهانی سازی، برای درک بهتر این مسئله مورد استفاده قرار گرفت که این سایت ها تا چه حد با هم تلفیق شده اند (Van Alst and Walgrave 2002). پیوندهای موجود بین 17 سایت، با استفاده از نرم افزار دیداری سازی شبکه (Pajek: http://vlado.fmf.uni-lj.si/pub/networks/pajek (2)، مورد تجزیه و تحلیل قرار گرفت. نرم افزار یاد شده، نقشه گرافیکی پیوندهای میان وبگاه ها را ایجاد می کرد. سایت های بسیار مرتبط و دارای ارجاعات بسیار با گردآوری پیوندها در موقعیت خودشان، و خروج از آن ها به آسانی شناسایی شدند. حال آن که، سایت های پراکنده تنها به وسیله یک یا دو پیوند با دیگر سایت های شبکه پیوند می یابند، دور افتاده به نظر می رسند. ون اِلست و والگریو دریافتند که سایت های ضد جهانی سازی تا اندازه ای یکپارچه شده اند، اما یادآور می شوند که ارزیابی ماهیت روابط بین سازمان هایی که صرفاً مبتنی بر موجودیت پیوندهای بین سایت ها می باشند، دشوار است.

در موارد دیگر، فراپیوندها به منظور کشف شبکه های اجتماعی و اطلاعاتی افراد پیوسته مورد تجزیه و تحلیل قرار گرفته اند بیشتر مرورگرهای وب، مانند اینترنت اکسپلورر (3) و Netscape Navigator ، مجهز

1- Network analysis

2- Pajek)

3- Internet Explorer

به کـار کرد «تاریخچه (1) » هسـتند که نشانی های اینترنتی را که شـخص در هنگام بازدید از سایت های شـبکه به آن ها دسترسـی داشته است، ذخیره می کنند. تاچر و گرینبرگ (1997)، از این کار کرد برای تحلیل رفتارهای مرورگرانه 23 نفر طی یک دوره 6 هفته ای استفاده کردند. آن ها به دنبال الگوهایی در بازدید صفحه های وب بودند. این دو پژوهشگر، متوجه شدند که تقریباً دو سوم بازدیدهای صفحه های وب از سوی آزمودنی ها، صـفحه هایی هسـتند که قبلاً بازدیـد شده بودند (صـفحه 112). همچنین، تعداد صـفحه هایی که اغلب توسـط مخاطبان دوباره بازدید می شوند، نسبتاً کم بودند. تاچر و گرینبرگ، فرض می کنند که کارکرد «بازگشت (2) » در مرورگرهای وب، که 30 درصد از فعالیت های مرورگرایانه آزمودنی ها را تشکیل می دهد، ممکن است عامل مؤثری برای علاقه کاربران اینترنت برای بازدید دوباره صفحه ای واحد در وب باشد (صفحه 131).

علاوه بر شـبکه های اجتماعی مطالعهٔ شـبکه های مدارک، به دلیل بزرگی شبکه وب رو به فزونی رفته است، و این بدان معناست که اَشکال گوناگون تحلیل علمی از شـبکه اجتماعی امکان پـذیر است. تحلیل شـبکه های مـدارک می توانـد فرصت هایی را برای فراتحلیل محتوا به وجود آورد. هنزینگر و لورنس (2004)، درباره روش های نمونه برداری از صـفحه های وب به بحث و گفت و گو می پردازنـد تا «به طور خودکـار نمونه وسـیعی از علایق و فعالیت های انجام گرفته در این دنیا را ... به وسـیله تجزیه و تحلیل ساختار پیونـد وب و چگونگی روی هم انباشته شدن پیوندها در طول زمان تجزیه و تحلیل کنند» (صفحه 5186). آیزن و همکارانش (2004)، بر این ادعا هستند که استفاده از داده ها در وبگاهی با ترافیک بالا می توانـد اطلاعاتی را درباره وقایع خارجی و احساسات ناگهانی در دسترس عموم قرار دهد که ممکن نیست به تنهایی از تحلیل محتوا و ساختار پیوند دسترس پذیر شود (صفحه 5254).

8-ملاحظات اخلاقی

اهمیت نیاز به پژوهش با رعایت اصول اخلاقی برای پژوهشـگران اینترنت پوشـیده نیست. اما خط مشـی ها و ابزارهای پژوهش اخلاقی به روشـنی مشـخص نیسـتند. از دیدگاه پژوهشگران وب، و به ویژه آن هایی که مواد آرشیوی وب را استخراج می کنند، ملاحظات اخلاقی به صورت ضـرورت در جریان پژوهش نمود می یابند. برخی اطلاعات موجود در وب و اطلاعات در دست آرشـیو، که محرمانه به شمار می آیند، ممکن است سـهواً دسترس پذیر شده، و سپس توسط موتورهای کاوشی نظیر گوگل نمایه و ذخیره شوند. دیدگاه های مربوط به امکان اسـتفاده از اطلاعات خصوصـی، که به صورت عمومی آرشـیو شده اند، متعدد و خارج از حوصـله این مقاله می باشند. با وجود این، باید خاطرنشان کرد که تمامیت موتورهای کاوش در جست و جوی وب می تواند منجر به ایجاد آرشیوهای اطلاعاتی شود که کاربران عموماً نه قصـد فـاش کردن آن ها را دارنـد و نه آن ها را آرشـیو کرده انـد. در حالتی خوش بینانه، برآوردی سـرعتی که موتورهای کاوش وب را نمایه سازی می نمایند، نشان داد که سرعت نمایه سازی و آرشیوسازی گوگل از ایجاد صفحه های جدید

History -1

Back -2

وب فراتر رفته است (Whelan 2004). احتمالاً مسائلی که پژوهشگران حوزه مطالعات اینترنت مدت ها به بحث و تبادل نظر گذاشته اند، و نیز مطالعات انجام شده بر روی ارتباطات با واسطه رایانه، با توجه به تقابل داده های عمومی با داده های خصوصی، در آینده بسیار نزدیک در میان پژوهشگران حوزه مطالعات وب و حرفه مندان حوزه آرشیو وب و دیگر مواد الکترونیکی از اهمیت بالایی برخوردار خواهند شد.

9-نتیجه گیری

پژوهش های حوزه وب به وضوح سودمندی خود را نشان داده اند. این پژوهش ها بدون آن که در سایه تحلیل های متنی جوامع مبتنی بر متن قرار داشته باشند، ثابت کرده اند که مفاهیمی نظیر جامعه، فرهنگ، ،رفتار و ساختارهای معنایی می توانند در اینترنت به طور مؤثری بررسی شوند. در حالی که غنا و گوناگونی پژوهش های وب نوید بخش است، هنوز مسائل زیادی وجود دارد که باید کشف شوند.

میراث تحلیل متنی به جامانده از ارتباطات با واسطه رایانه در سوگیری شدید نسبت به تحلیل های زبانی در مطالعات حوزه وب مشهود است. «مطالعات مفقود» به مطالعاتی اطلاق می شود که وب را مانند چشم اندازی چندرسانه ای بررسی می کند، و توصیف می کند که چگونه تصاویر و صدا برای از میان برداشتن ارتباطات گفتاری مورد استفاده قرار می گیرند تا بر موانع زبانی در رسانه جهانی فائق آیند. به استثنای اشاره به کاربرد صدا در وبگاه های ترینیدادی، در مطالعه قوم نگاری میلر و اسلتر (2000)، وب گاه های شنیداری کاملاً از بخش کنونی پژوهش وب غایب بودند. احتمال دارد که پروژه های آینده به کاربردهای بلاغی صدا در وب از طریق تحلیل های وب گاه های تجاری یا سیاسی بپردازند. دیگر مطالعاتی که به حوزه چند رسانه ای ها گرایش دارند توانستند روش هایی را که از صدا و تصاویر برای ایجاد هویت افراد موجود در اینترنت از افراد حقیقی و سازمان ها گرفته تا دولت های ملی استفاده می شدند، بررسی کنند.

به منظور افزایش تعداد پژوهشگران این حوزه دسترسی به همه انواع محتوای وب مربوط به تمامی دوره های زمانی، به ویژه بازآفرینی پیوندهای درون اشیای محتوایی و میان وبگاه ها الزامی است. به علاوه، باید مرورگرها و ابزارهای دیگری در دسترس قرار گیرند که محتوای وب به وسیله آن ها رصد شود تا بتوان تجربه کاربر را به بهترین شکل درک کرد. بنابراین، چالش ها فراتر از حفظ و نگهداری محتوا رفته، و شامل حفظ ساختار و رویارویی با محتوا نیز می شود.

منابع

Aizen, J., Huttenlocher, D., Kleinberg, J., Novak, A. (2004). Traffic-based feedback on the Web. . 1 Proceedings of the National Academy of Sciences of the United States of America, 101(1), 5254-5260

Billig, M. (2001). Humour and hatred: The racist jokes of the Ku Klux Klan. Discourse .2

Society, 12(3), 267–289

Bloustein, G. (2002). Fans with a lot at stake: Serious play and mimetic excess in Buffy the Vampire .3
Slayer. European Journal of Cultural Studies, 5(4), 427– 449

Carswell, A. D. Venkatesh, V. (2002). Learner outcomes in an asynchronous distance education .4
environment. International Journal of Human–Computer Studies, 56, 475-494

Chandler, D. Roberts–Young, D. (1998). The construction of Identity in the Personal Homepages of .5
Adolescents. Retrieved April 15, 2002 from http://www.aber.ac.uk/media/
Documents/short/strasbourg.html

Creswell, J. W. (1998). Qualitative inquiry and research design. Thousand Oaks, CA: Sage .6

Duffy, M. E. (2003, July). Web of hate: A fantasy theme analysis of the rhetorical vision of hate groups .7
online. Journal of Communication Inquiry, 27(3), 291-312

Ess, C. (2002). Ethical decision making and Internet research: Recommendations from the AoIR working .8
committee. Association of Internet Researchers. Retrieved July 12, 2004, from http://www.aoir.org/reports/
ethics.pdf

Fair, J. E. Parks, L. (2001). Africa on Camera: Television news coverage and aerial imaging of Rwandan .9
refugees. Africa Today, 48(2), 34-57

Foss, S. K. (1996). Rhetorical criticism: Exploration and practice. Prospect Heights, IL: Waveland .10

Garton, L., Haythornthwaite, C., Wellman, B. (1997). Studying online social networks. Journal of .11
Computer–Mediated Communication, 3 (1). Retrieved April 4, 2004 from http://www.ascusc.org/jcmc
/vol3 /issuel / garton.htm

Geertz, C. (1973). The interpretation of cultures. New York: Basic Books Graphic, Visualization and .12
Usability (GVU) Center (1998). 10th WWW User Survey. Atlanta, GA: Georgia Institute of Technology.
./Retrieved March 25, 2004 from http://www. cc.gatech.edu/gvu/user–surveys/survey–1998–10

Halavais, A. (2000). National Borders on the World Wide Web. New Media Society, 1(3), 7–28 .13

Haraway, D. (1991). Simians, cyborgs, and women: The reinvention of nature. New York: Routledge .14

Henzinger, M. Lawrence, S. (2004). Extracting knowledge from the World Wide Web. Proceedings of .15
,the National Academy of Sciences of the United States of America

ص: 87

5186-5191 ,(1)101

Hesketh, A. J. Selwyn, N. (1999). Surfing to school: The electronic reconstruction of institutional . 16
identity. Oxford Review of Education, 25(4), 501-520

Howard, P. N. (2002). Network Ethnography and the Hypermedia Organization: New Media, New . 17
Organizations, New Methods. New Media and Society, 4(4), 550-574

Jimroglou, K. M. (1999). A Camera with a view: JenniCam, visual representation, and . 18
cyborgsubjectivity. Information, Communication and Society. 2(4), 439-453

Kroeber, A. (2001). Postmodernism, Resistance, and Cyberspace: Making Rhetorical Spaces for . 19
Feminist Mothers on the Web. Women's Studies in Communication, 24(2), 218-240

Lee, W., Hong, J., Lee, S. (2003). Communicating with American consumers in the post 9/11 climate: .20
An empirical investigation of consumer ethnocentrism in the United States. International Journal of
Advertising, 22, 487-510

McQuail, D. (2000). McQuail's mass communication theory (4th ed.). London, UK: Sage .21

Miller, D. Slater, D. (2000). The Internet: An ethnographic approach. Oxford, UK: Berg .22

Milne, G. R. Culnan, M. J. (2002). Using the content of online privacy notices to inform public policy: .23
A longitudinal analysis of 1998-2001 US Web surveys. The Information Society, 18, 345-359

O'Neill, E. T., Lavoie, B. F., Bennett, R. (2003). Trends in the evolution of the public Web, 1998-2002. .24
./D-Lib Magazine, 9(4). Retrieved March 29, 2004 from http://wcp. oclc.org

Park, H. W. (2003). Hyperlink network analysis: A new method for the study of social structure on the . 25
Web. Convergence, 25(1), 49-61

Pitts, M. J. Harms, R. (2003). Radio websites as promotional tools. Journal of Radio Studies, 10(2), . 26
270-282

Purcell, D. Kodras, J. E. (2001). Information technologies and representational spaces at the outposts of .27
the global political economy. Information, ommunication and Society, 4(3), 341-369

Rivett, M. (2000). Approaches to analyzing the Web text: A consideration of the Web site as an . 28

emergent cultural form. Convergence, 6(3), 34–56

?Sarraf, S. (1999). A survey of museums on the Web: Who uses museum Websites .29

Curator, 42(3), 231–243

Silver, R. C., Holman, E. A., McIntosh, D. N., Poulin, M., Gil-Rivas, V. (2002). Nationwide . 30 longitudinal study of psychological responses to September 11. Journal of the American Medical Association, 288(10), 1235-1244

Spooner, T. (2002). Internet use by region in the United States. Pew Internet American Life Project. . 31 Washington, DC. Retrieved March 25, 2004 from http://www.pewinternet. org/reports/pdfs/ PIP-Regional-Report- Aug-2003.pdf

Tauscher, L. Greenberg, S. (1997). How people revisit Web pages: Empirical findings and implications . 32 for the design of history systems. International Journal of Human- Computer Studies, 47, 97–137

Van Aelst, P. Walgrave, S. (2002, December). New media, New movements? The role of the Internet in . 33 shaping the 'anti-globalization' movement. Information Communication and Society, 5(4), 465-493

Warnick, B. (1998). Appearance or reality? Political parody on the Web in Campaign '96. Critical . 34 Studies in Mass Communication, 15, 306-324

Whelan, D. (2004, 16). Google Me Not. Forbes, 174(3), 102-104 .35

چکیده

اطلاع از مشـخصات وب ملی یکی از نیازمندی های اصـلی کشور در رابطه با سیاست گذاری در حوزه فناوری اطلاعات است. از آن جا که متأسفانه تا به حال، تحقیقی در این خصوص، انجام نشده است، نیاز مبرمی برای پژوهش و کسب اطلاع از وضعیت فعلی وب فارسی، احسـاس می شود. بـا درک این ضـرورت، در این مقاله نتایـج حاصل از مطالعات صورت گرفته در مورد خصوصـیات و ویژگی های وب فارسـی مطرح خواهد شد. برای این کار سامانه نیمه خودکاری طراحی و پیاده سازی شده است. هدف از این سیسـتم، اسـتخراج شاخص های مختلف از قبیل حجم محتوای فارسـی، تنوع محتوا و نیز عمر محتوا می باشـد. از این آمار می توان جهت هدفمند نمودن برنامه های آتی در خصوص ساماندهی وب فارسـی و نیز ارائه راهکارهایی برای حل معضـلات فعلی، بهره گرفت. آنالیزهای انجام شده توسط سامانه فوق الـذکر بر روی حدود یازده هزار وبگاه ثبت شده در دامنه IR، مشتمل بر اغلب سازمان ها و ارگان های دولتی، وزارتخانه ها، شرکت ها و دانشگاه ها است که بالغ بر حدود دو میلیون صفحه می باشند.

کلید واژه: دامنه IR، وب فارسی، مشخصات وب

خصوصیات وب ایران: مریم پیروزمند

اشاره

*خصوصیات وب ایران: مریم پیروزمند (1)

1.مقدمه

توسعه و رشد نمایی وب باعث شده است تا با حجم عظیمی از اطلاعات شامل اسناد با فرمت های متفاوت از جمله متن، صوت، تصویر و غیره در مکان ها و سازمان های مختلف مواجه شویم. در عین حال، گسترش روز افزون نیازمندی به وب که زاییدهٔ فزونی عرضه خدمات از طریق وب است، باعث شده تا همواره کاربران بیشتری به استفاده از وب، راغب شوند. بنا بر آمارهای موجود در حال حاضر بالغ بر 130 میلیارد صفحه در محیط وب از طریق جویشگر های مطرح دنیا قابل دسترسی و جستجو هستند [1]. به دلایل مختلف از جمله شروع توسعه وب توسط کاربران انگلیسی زبان، غالب سرویس ها و خدمات عرضه شده از طریق وب بویژه سرویس های جستجو، بصورت انگلیسی ارائه می شوند. گر چه در طی سال های اخیر، موتورهای جستجوی عمده ای مانند Yahoo و Google، سرویس جستجو را به زبان های دیگر نیز عرضه کرده اند، اما متأسفانه به دلایل مختلف بالاخص تحریم های سیاسی و اقتصادی، این خدمات هیچگاه به حوزه زبان فارسی گسترش پیدا نکرده است. از سوی دیگر، در داخل کشور نیز به علت عدم وجود شناخت کافی از مختصات وب فارسی موتورهای جستجوی توانمندی بوجود نیامده اند.

بطور کلی، شناخت مختصات وب ملی یک کشور، نمایان گر شاخص های توسعه یافتگی در آن کشور

1- دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی امیر کبیر piroozmand@aut.ac.ir

نیز محسوب می شود. به عنوان مثال، در کشورهای توسعه یافته، حجم وب ملی شامل تعداد وبگاه های مختلف و همچنین تنوع خدمات قابل عرضه در این رسانه، بسیار زیاد است. همچنین کیفیت و به روز بودن اطلاعات عرضه شده نیز در دامنهٔ وب این قبیل کشورها در مقایسه با دیگر کشورها بسیار بهتر است.

از سوی دیگر، با توجه به اینکه زیر ساخت اصلی اشاعه دولت الکترونیک، محیط وب می باشد، لذا مطالعه و شناسایی ویژگی های این محیط، کمک شایانی به انجام برنامه ریزی مناسب جهت تدوین سیاست های اجرایی جهت تحقق اهداف دولت الکترونیک طی مراحل مختلف، خواهد کرد. با درک این مطلب، در این مقاله سعی می شود تا به ویژگی های اصلی وب ایران پرداخته شود. شاخص های مورد مطالعه، به گونه ای انتخاب شده است تا بتوان از آن ها در جهت بهبود عملکرد سیستم های جستجوی وب، بهره گرفت.

با توجه به مسائل فوق در این مقاله سعی می شود تا حد ممکن، محتوای وب فارسی کشور را به صورت خودکار، تحلیل و ارزیابی کرد (ابزار نظارت خودکار) تا در مقاطع زمانی مختلف، بتوان عمل ارزیابی را با کمترین هزینه، تکرار نمود. لازم به ذکر است که در این مقاله حدود 11 هزار وبگاه با پسوند IR، که شامل تقریباً دو میلیون صفحه است پردازش و تحلیل شده اند. البته تعداد وبگاه های فارسی بیش از این مقدار است اما به دلیل دشواری یافتن وبگاه های فارسی زبان، در این پژوهش به وبگاه های ثبت شده در دامنه IR بسنده شده است. علت این دشواری این است که همانطور که در قسمت نتایج، به تفصیل ذکر خواهد شد. متاسفانه دامنه IR به عنوان تنها دامنه فارسی زبان نیست و بسیاری از وبگاه های فارسی زبان، در دامنه های مختلف و حتی غیر مرتبطی از قبیل .com, net, org و دامنه های دیگر ثبت شده اند.

همانگونه که پیش از این ذکر شد، هدف اصلی این مطالعه، تعیین وضعیت فعلی وب فارسی است.

بر این اساس، هدف نهایی، استخراج شاخص های زیر از "وب ایران" می باشد:

1. پاسخگویی به سؤالات کلی درباره وب ایران از جمله:

1-1. درصد پیوندهای معتبر چقدر است؟ تعیین این شاخص بطور ضمنی، نرخ تغییرات وب را نیز تعیین خواهد کرد و زمان بندی برای خزشگر وب را دقیق خواهد کرد.

1-2. توزیع صفحات وب فارسی از لحاظ محتوا، چگونه است؟ یعنی چه حجمی از محتوای وب در کلاس های علمی، تجاری، روزنامه، خبر، وبلاگ و غیره، قابل طبقه بندی است؟

1-3. ساختار محتوایی صفحات وبگاه ها چگونه است؟ بطور مشخص می خواهیم بدانیم که:

- درصد کدینگ های زبانی مختلف استفاده شده نظیر Windows- 1252, Windows - 1256 و UTF-8 به چه صورت است؟ بدین ترتیب مشخص خواهد شد که تبدیل کدینگ ها در سیستم های بازیابی وب چقدر حائز اهمیت است؟

- چه کسری از صفحات، عنوان (1) مناسب دارند؟ وجود صفحات با عنوان مناسب و گویا، موجب بهبود کیفیت بازیابی توسط سیستم و نیز سهولت دسترسی توسط کاربران خواهد شد.

Title -1

1-4. فایل های غیر متنی مانند Doc PPT PDF و Image چند درصد از صفحات را تشکیل می دهند؟

1-5. نرخ به روزآوری، تغییر، ایجاد (عمر صفحه) چقدر می باشد؟ استخراج این اطلاعات، نرخ تغییرات وب را تعیین خواهد کرد و موجب برنامه ریزی و زمان بندی برای خزش گر وب را فراهم خواهد آورد.

1-6. سرعت دسترسی به وبگاه چقدر است؟ این آگاهی نیز در تنظیم عملکرد خزشگر وب موثر خواهد بود.

1-7. تعداد دسترسی به وبگاه ها و صفحات در مقاطع زمانی مختلف یا بعبارت دیگر الگوی دسترسی به وبگاه ها و صفحات چگونه می باشد؟

1-8. تعداد کل صفحات فارسی، میانگین تعداد صفحات هر وبگاه و حجم آن ها چقدر است؟

1-9. تعداد لغاتی که در تمام صفحات محاسبه شده چقدر می باشد؟

1-10. چند درصد صفحات شامل هر دو محتوای فارسی و انگلیسی است؟

لازم به ذکر است در این فاز به دلایلی مانند پیچیدگی کار استخراج شاخص های مربوط به فازهای تراکنش و تبدیل به کارهای آینده موکول شده است. نتایج این مقاله را می توان در تعریف و پیاده سازی پروژه هایی مانند موتور جستجوی ملی و درگاه دولتی استفاده کرد. به علاوه این پروژه در تدوین راهکارهای آینده جهت تحقق سریع دولت الکترونیک در کشور مفید فایده خواهد بود.

2. کارهای مرتبط در داخل و خارج

تاکنون در داخل کشور مکانیزم ارزیابی وب بدین صورت انجام نشده است، اما بعضی از کشورها کاری شبیه به این پروژه را انجام داده اند. برای مثال در تایلند پروژه ای تحت عنوان "ابزار نظارت خودکار بر پروژه دولت الکترونیک تایلند" [2] انجام شده است. هدف این کار استخراج تمام شاخص های دولت الکترونیک از وب تایلند می باشد. بر اساس نتایج این مطالعه که حاصل بررسی حدود 150 وبگاه دولت تایلند در سال 2002 است، حدود 31% این سایت ها، صرفاً به ارایه اطلاعات می پردازند، حدود 57% امکان تعامل محدود کاربران را با وبگاه فراهم می کنند و تنها حدود کنند و تنها حدود 11%، بستر لازم برای اجرای تراکنش های مورد نیاز کاربران را در اختیار وی قرار می دهند.

به صورت مشابه، کارهایی برای استخراج مشخصه های وب در کشورهای اسپانیا [3] ، کره جنوبی [4]، استرالیا[5] ، پرتغال [6] ، اروپا [7، 8] و آرژانتین [9] انجام شده است. در کنار این تحقیقاتی نیز در مورد مطالعه ویژگی های کلی وب صورت گرفته است [10، 11، 12] در فعالیت های فوق بیشتر وب کشورها از دیدگاه ساختاری و شکل گراف وب مورد بررسی قرار گرفته است و به علاوه پارامترهایی مانند توزیع اندازه صفحات و وبگاه ها، نرخ بروزآوری آن ها، رتبه آن ها در موتورهای جستجو و عمر صفحات مورد بررسی قرار گرفته ولی از منظر محتوا و سرویس های ارائه شده کاری انجام نگرفته است. در این مقاله علاوه بر استخراج پارامترهای فوق برای وب فارسی، از دید دولت الکترونیک نیز به وب

ایران توجه گردیده است.

در تحقیق صورت گرفته در مرجع [10]، ساختار کلی گراف وب جهانی مورد مطالعه قرار گرفته است. بررسی های صورت گرفته حاکی از وجود ساختاری موسوم به مدل پاپیونی (1) در مورد گراف وب است. بعنوان مثال، طی بررسی انجام شده در سال 2000 میلادی روی حدود دویست میلیون صفحه وب، همانگونه که شکل شماره یک نشان می دهد، گراف وب، شامل چهار بخش کلی است که عبارتند از:

الف- بخش هسته مرکزی (2) که حدود 25% کل وب را تشکیل می دهد.

ب- بخش ورودی (3) که لینک هایی به بخش هسته مرکزی دارد و حدود 25% کل وب را تشکیل می دهد.

ج- بخش خروجی (4) که لینک هایی از بخش هسته مرکزی دارد و حدود 25% کل وب را تشکیل می دهد.

د- بخش های پراکنده که یا به صورت جزایر پراکنده (و بعضاً به سختی قابل دسترس) هستند یا واسط هایی بین بخش های ورودی و خروجی هستند. این قسمت نیز حدود 25% کل وب را تشکیل می دهد.

عکس

ایران توجه‌گردیده است.

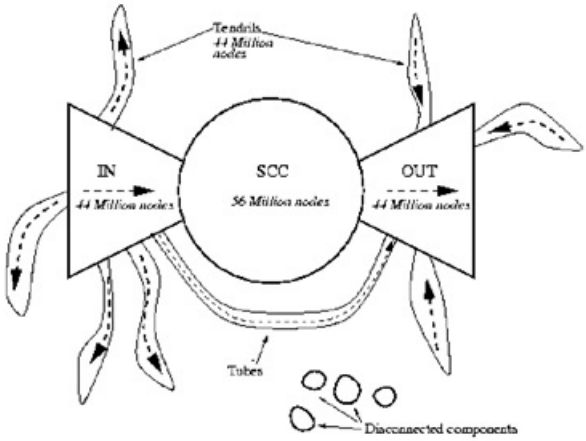
در تحقیق صورت گرفته در مرجع [۱۰]، ساختار کلی گراف وب جهانی مورد مطالعه قرار گرفته است. بررسی‌های صورت گرفته حاکی از وجود ساختاری موسوم به مدل پاپیونی[1] در مورد گراف وب است. بعنوان مثال، طی بررسی انجام شده در سال ۲۰۰۰ میلادی روی حدود دویست میلیون صفحه وب، همانگونه که شکل شماره یک نشان می‌دهد، گراف وب، شامل چهار بخش کلی است که عبارتند از:

الف-بخش هسته مرکزی[2] که حدود ۲۵٪ کل وب را تشکیل می‌دهد.

ب-بخش ورودی[3] که لینک‌هایی به بخش هسته مرکزی دارد و حدود ۲۵٪ کل وب را تشکیل می‌دهد.

ج-بخش خروجی[4] که لینک‌هایی از بخش هسته مرکزی دارد و حدود ۲۵٪ کل وب را تشکیل می‌دهد.

د-بخش‌های پراکنده که یا به‌صورت جزایر پراکنده (و بعضاً به سختی قابل دسترس) هستند یا واسطه‌هایی بین بخش‌های ورودی و خروجی هستند. این قسمت نیز حدود ۲۵٪ کل وب را تشکیل می‌دهد.



شکل ۱. مدل پاپیونی گراف وب

۳. سامانه خودکار ارزیابی وب ایران

در این قسمت، ابتدا معماری سیستم ارزیاب خودکار، تشریح می‌شود. این سامانه، نسخه گسترش یافته ابزار استفاده شده در مرجع [۱۳] است که با هدف بررسی مشخصات وبگاه‌های ثبت شده در دامنه IR اعم از دولتی و غیر دولتی می‌باشد. شکل شماره دو شمای کلی این سامانه و تعامل اجزای آن را با

1. Bow-Tie model
2. Central core
3. In
4. Out

شکل 1. مدل پاپیونی گراف وب

3.سامانه خودکار ارزیابی وب ایران

در این قسمت، ابتدا معماری سیستم ارزیاب خودکار، تشریح می شود. این سامانه، نسخه گسترش یافته ابزار استفاده شده در مرجع [13] است که با هدف بررسی مشخصات وبگاه های ثبت شده در دامنه IR اعم از دولتی و غیر دولتی می باشد. شکل شماره دو شمای کلی این سامانه و تعامل اجزای آن را با

Bow-Tie model -1

Central core -2

In -3

Out -4

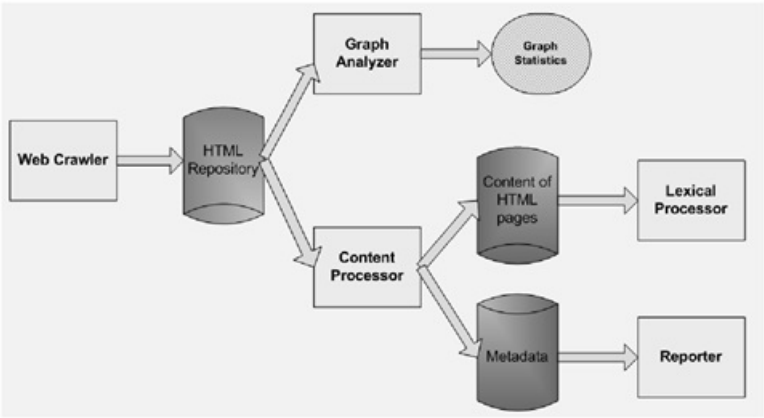
یکدیگر، نشان می دهد.

همان گونه که در شکل شماره دو دیده می شود، ابتدا یک خزشگر وب، بر اساس یک لیست اولیه از وبگاه های ثبت شده در دامنه IR که از قبل تهیه شده است و با توجه به پارامترهای تنظیم عملکرد آن از قبیل عمق خزش، حداکثر صفحات وبگاه و غیره، خزش را با توجه به گراف حاصل از پیوندها، انجام می دهد و صفحات بدست آمده را در یک مخزن موقت، شاخص بندی و ذخیره می کند. در مرحله بعد، این مخزن توسط واحد تحلیل گر گراف وب، مورد بررسی قرار می گیرد و آمارهای مختلفی را مورد ویژگی های گراف متناظر از قبیل قطر گراف، متوسط فاصله بین هر دو گره و غیره، استخراج می کند.

عکس

یکدیگر، نشان می‌دهد.

همانگونه که در شـــکل شماره دو دیده می‌شود، ابتدا یک خزشگر وب‌بر اساس یک لیست اولیه از وبگاه‌های ثبت شده در دامنه IR که از قبل تهیه شده است و با توجه به پارامترهای تنظیم عملکرد آن از قبیـــل عمق خزش، حداکثر صفحات وبگاه و غیره، خزش را با توجه به گراف حاصل از پیوندها، انجام می‌دهد و صفحات بدست آمده را در یک مخزن موقت، شاخص‌بندی و ذخیره می‌کند. در مرحله بعد، این مخزن توســـط واحد تحلیل‌گر گراف وب، مورد بررســـی قرار می‌گیرد و آمارهای مختلفی را مورد ویژگی‌های گراف متناظر از قبیل قطر گراف، متوسط فاصله بین هر دو گره و غیره،استخراج می‌کند.



شکل ۲. معماری کلی سامانه خودکار ارزیابی وب ایران

از ســـوی دیگر، واحد تحلیل‌گر محتوا با اســـتفاده از تجزیه کنندهٔ[1] HTML، صفحات این مخزن را بررســـی نموده و محتوا و داده‌های توصیفی آنها را اســـتخراج می‌کند. این قسمت که توسط زبان جاوا پیاده‌ســـازی شده است، بصورت مســـتقل و از طریق پارامترهای قابل تنظیم، اجرا می‌شود. ورودی این بخش، کل صفحات HTML است که شامل برچسب‌های مختلف و نیز نویسه[2]های موجود در صفحه می‌باشد. شایان ذکر است که در هنگام جمع‌آوری صفحات از وب، توسط خزشگر وب بصورت خودکار به هر صفحه، یک شناســـه عددی نسبت داده می‌شـــود که آن صفحه را بصورت یکتا مشخص می‌کند. جدول شماره یک، نمونه‌ای از ورودی‌های تجزیه کننده را نشان می‌دهد.

محتوای استخراج‌شده از صفحات، برای پردازش در اختیار واحد تحلیل‌گر واژگان، قرار داده می‌شود تا کلمات فارسی را از آن استخراج کند و در یک Lexicon ذخیره نماید. همچنین از اطلاعات بدست‌آمده می‌توان برای استخراج آمار متنوع از کلمات استفاده کرد.

1. Parser
2. Script

شکل2. معماری کلی سامانه خودکار ارزیابی وی ایران

از سوی دیگر، واحـد تحلیل گر محتوا با اسـتفاده از تجزیه کننـدهٔ (1 Parser) صـفحات این مخزن را بررسـی نموده و محتوا و داده های توصـیفی آن ها را اسـتخراج می کند. این قسـمت که توسط زبان جاوا پیاده سازی شده است، بصورت مستقل و از طریق پارامترهای قابل تنظیم اجرا می شـود ورودی این بخش کـل صـفحات HTML است که شامـل برچسب هـای مختلـف و نیز نویسه (2) های موجود در صفحه می باشد. شایان ذکر است که در هنگام جمع آوری صفحات از وب توسط خزشگر وب بصورت خودکار به هر صفحه، یک شناسه عددی نسـبت داده می شود که آن صـفحه را بصورت یکتا مشـخص می کند. جدول شـماره یک، نمونه ای از ورودی های تجزیه کننده را

نشان می دهد.

محتوای استخراج شده از صفحات برای پردازش در اختیار واحد تحلیل گر واژگان، قرار داده می شود تا کلمات فارسی را از آن استخراج کند و در یک Lexicon ذخیره نماید. همچنین از اطلاعات بدست آمده می توان برای استخراج آمار متنوع از کلمات استفاده کرد.

ص: 95

HTML -1

Script -2

بخش مهم دیگر این معماری، واحد تهیه گزارش است. داده های توصیفی استخراج شده از صفحات ورودی به عنوان ورودی به این واحد، ارسال می شود تا مورد بررسی قرار گیرد. در نتیجه این فرآیند آمارهای ارزشمند مختلفی نظیر درصد استفاده از کدینگ های مختلف توزیع صفحات و وبگاه ها در طبقه بندی های مختلف سرویس های ارائه شده نظیر جستجو، امنیت و غیره بدست می آید.

عکس

بخش مهم دیگر این معماری، واحد تهیه گزارش است. داده‌های توصیفی استخراج‌شده از صفحات ورودی به عنوان ورودی به این واحد، ارسال می‌شود تا مورد بررسی قرار گیرد. در نتیجه این فرآیند، آمارهای ارزشمند مختلفی نظیر درصد استفاده از کدینگ‌های مختلف، توزیع صفحات و وبگاه‌ها در طبقه‌بندی‌های مختلف، سرویس‌های ارائه شده نظیر جستجو، امنیت و غیره، بدست می‌آید.

جدول ۱. نمونه‌ای از ورودی‌های تجزیه کننده به ازای وبگاه http://www.whc.ir

```
<!-- DOCID: 166182 URL: www.whc.ir/-->
<head>
 <meta http-equiv="Content-Language" content="fa">
 <meta http-equiv="Content-Type" content="text/html; charset=utf-8">
 <meta name="author" content="xx">
 <meta name="copyright" content="2004-2005 by xxx">
 <title>Send to friend</title>
 <style type="text/css"></style>
 <script type="text/javascript"></script>
</head>
<body>
<div id="container">
 <form method="POST" action="">
 <table>
  <tr>
   <td>.....</td>
  </tr>
  <tr>
   <td><input type="submit" name="submit">
  </td> <td> </td> </tr>
 </table>
 <input type="hidden" name="action" value="1">
 <input type="hidden" name="newsid" value="1117">
</form>
</div>
</body>
```

٤. نتایج بدست‌آمده

ورودی سامانه تحلیل‌گر محتوا، بدین‌صورت انتخاب شد که به‌ازای هر وبگاه، حداکثر بیست‌هزار صفحه مورد بررسی قرار گرفت. علت این امر این است که وبگاه‌های با تعداد صفحات بالاتر، معمولاً فقط وبگاه‌های خبری نظیر IRNA، ISNA، IRIB و غیره هستند. در این خصوص، آمار حاکی از آن است که از حدود ۸ میلیون صفحه موجود در دامنه IR. حدود ٦ میلیون آن، مربوط به وبگاه‌های خبری است. لذا محتوای مناسب به جز خبر، تنها حدود ۲ میلیون صفحه است که این رقم در مقابل اغلب کشورها که بیش از ۱۰۰ میلیون صفحه دارند، رقم بسیار ناچیزی می‌باشد (به‌عنوان مثال، دولت الکترونیکی کره جنوبی، شامل بیش از ۱۰۸ میلیون صفحه است).

بواسطه حجم محاسباتی بالای به‌منظور انجام ارزیابی‌های مختلف، یک جامعه آماری شامل حدود ٦۰۰ هزار صفحه، در نظر گرفته شد و بررسی‌های مختلف، روی این مجموعه، انجام شد.

در زیر آمار صفحات و وبگاه ها از نظر حجم، عمر و نوع محتوا برای تمام صفحات غیر از اخبار شامل دو میلیون صفحه ارائه شده‌است.

جدول1. نمونه ای از ورودی های تجزیه کننده به ازای وبگاه http://www.whc.ir

4. نتایج بدست آمده

ورودی سامانه تحلیل گر محتوا، بدین صورت انتخاب شد که به ازای هر وبگاه، حداکثر بیست هزار صفحه مورد بررسی قرار گرفت. علت این امر این است که وب گاه های با تعداد صفحات بالاتر، معمولاً فقط وبگاه های خبری نظیر IRIB ISNA IRNA و غیره هستند. در این خصوص، آمار حاکی از آن است که از حدود 8 میلیون صفحه موجود در دامنه IR. حدود 6 میلیون آن، مربوط به وبگاه های خبری است. لذا محتوای مناسب به جز خبر، تنها حدود 2 میلیون صفحه است که این رقم در مقابل اغلب کشورها که بیش از 100 میلیون صفحه دارند، رقم بسیار ناچیزی می باشد (به عنوان مثال دولت الکترونیکی کره جنوبی شامل بیش از 108 میلیون صفحه است).

بواسطه حجم محاسباتی بالای به منظور انجام ارزیابی های مختلف، یک جامعه آماری شامل حدود 600 هزار صفحه، در نظر گرفته شد و بررسی های مختلف، روی این مجموعه، انجام شد.

در زیر آمار صفحات و وبگاه ها از نظر حجم، عمر و نوع محتوا برای تمام صفحات غیر از اخبار شامل دو میلیون صفحه ارائه شده است.

4,1.آمار وب گاه ها

جدول شماره دو، آمار میانگین حجم و سن (1) و عمق و تعداد صفحات وبگاه ها را نشان می دهد.

عکس

٤،١ آمار وبگاه‌ها

جدول شماره دو، آمار میانگین حجم و سن[1] و عمق و تعداد صفحات وبگاه‌ها را نشان می‌دهد.

جدول ۲. خلاصه آمار سایت

تعداد وبگاه‌های معتبر	۱۰۰۰۰
درصد وبگاه‌های .ir در داخل کشور	٪۲۷
درصد وبگاه‌های .ir در خارج کشور	٪۷۳
میانگین تعداد صفحات در هر سایت	۲۳۰
میانگین تعداد صفحات ایستا در هر سایت	۱۲۰
میانگین تعداد صفحات پویا در هر سایت	۱۱۰
میانگین سن مسن‌ترین صفحه (بر حسب ماه)	۱۳/۶
میانگین سن جوان‌ترین صفحه (بر حسب ماه)	۶/۸۳
میانگین سن متوسط‌ترین صفحه (بر حسب ماه)	۸/۵
میانگین عمق ماکزیمم سایت	۴/۵۲
میانگین حجم سایت (بر حسب مگابایت)	۳/۰۶

شکل شماره سه تعداد صفحات وبگاه‌ها را بر حسب درصد سایت نشان می‌دهد. همانطور که نشان داده شده است بیشتر وبگاه‌ها دارای تعداد صفحات بین ۱۰۰ و ۱۰۰۰ (میانگین ۲۳۰) صفحه می‌باشند. نمودار دارای توزیع تقریباً خطی می‌باشد. شکل شماره چهار، توزیع تجمعی شکل شماره سه را نشان می‌دهد.

شکل شماره پنج، توزیع تجمعی حجم محتوای وبگاه‌ها را نشان می‌دهد که از این آمار می‌توان برای مدل‌سازی محتوای وبگاه‌ها استفاده کرد.

۱. سن صفحه عبارت است از مدت زمان میان ایجاد یا تغییر محتوای صفحه و زمان فعلی

شکل شماره سه تعداد صفحات وبگاه ها را بر حسب درصد سایت نشان می دهد. همان طور که نشان داده شده است بیشتر وبگاه ها

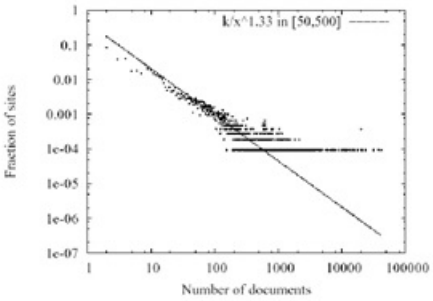
دارای تعداد صفحات بین 100 و 1000 (میاگین 230) صفحه می باشند. نمودار دارای توزیع تقریباً خطی می باشد. شکل شماره چهار، توزیع تجمعی شکل شماره سه را نشان می دهد.

شکـل شـماره پنـج، توزیع تجمعی حجم محتوای وبگـاه هـا را نشان می دهـد که از این آمار می توان برای مـدل سازی محتوای وبگاه ها استفاده کرد.

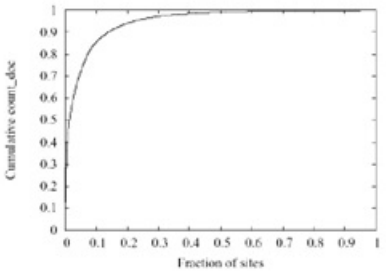
ص: 97

1- سن صفحه عبارت است از مدت زمان میان ایجاد یا تغییر محتوای صفحه و زمان فعلی

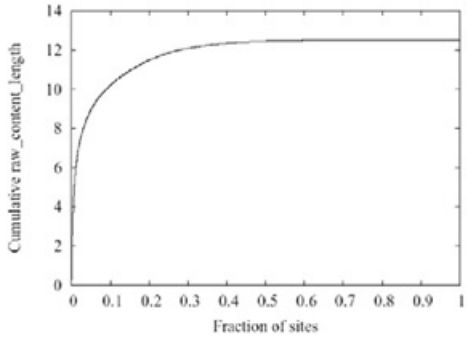
عکس



شکل ۳. تعداد اسناد در وبگاه‌ها

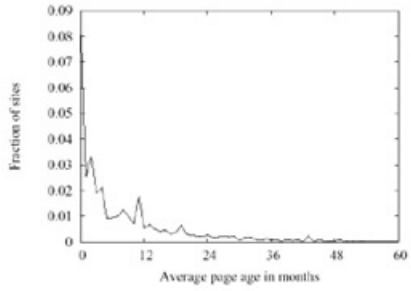


شکل ٤. توزیع تجمعی اسناد در وبگاه‌ها

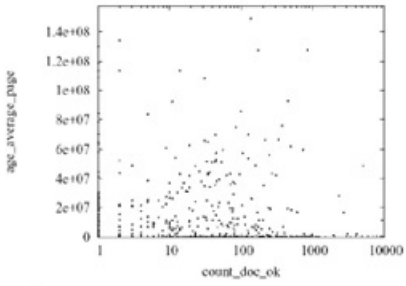


شکل ٥. توزیع تجمعی حجم محتوای وبگاه‌ها

شکل شماره شش نیز میانگین سن صفحات موجود در وبگاه‌ها را نشان می‌دهد. همچنین رابطه بین تعداد اسناد و سنِ یک سایت در شکل شماره هفت، نشان داده شده است.



شکل ٦. میانگین سن صفحات وبگاه‌ها بر حسب ماه



شکل ۷. تعداد اسناد بر حسب میانگین سن آنها

شکل 3. تعداد اسناد در وبگاه ها

شکل4. توزیع تجمعی اسناد در وب گاه ها

شکل 5. توزیع تجمعی حجم محتوای وب گاه ها

شکل شـماره شـش نیز میانگین سن صـفحات موجود در وبگاه ها را نشان می دهـد. همچنین رابطه بین تعداد اسـناد و سنِ یک سایت در شکل شماره هفت، نشان داده شده است.

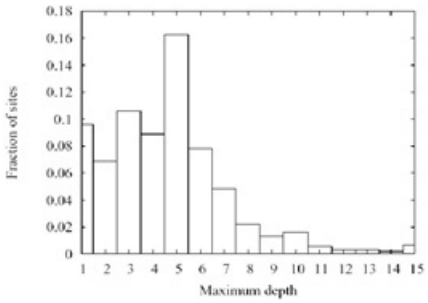
شکل 6. میانگین سن صفحات وب گاه ها بر حسب ماه

شکل 7. تعداد اسناد بر حسب میانگین سن آن ها

شکل شـماره هشت، توزیع عمق وب گاه ها را نشان می دهـد. درصد قابل توجهی از وب گاه ها دارای عمق 5 (23%) می باشند. شایان ذکر است که با توجه به شاخص های بـدست آمـده از وب گاه ها نظیر عمق، تعداد صـفحات، حجم محتوا و سن صـفحات، به راحتی می توان یک سایت را مـدل سازی کرد. با داشـتن یک مـدل مناسب از وبگاه ها می توان الگوریتم های مختلف را به راحتی تحت آزمون قرار داد.

عکس

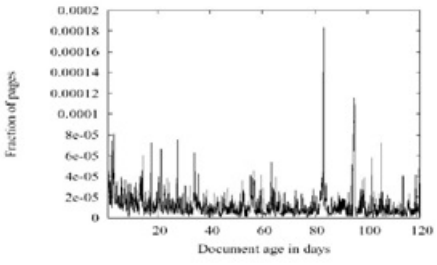
شکل شماره هشت، توزیع عمق وبگاه‌ها را نشان می‌دهد. درصد قابل توجهی از وبگاه‌ها دارای عمق ۵ (۲۳٪) می‌باشند. شایان ذکر است که با توجه به شاخص‌های بدست آمده از وبگاه‌ها نظیر عمق، تعداد صفحات، حجم محتوا و ســـن صفحات، به راحتی می‌توان یک سایت را مدل‌سازی کرد. با داشتن یک مدل مناسب از وبگاه‌ها می‌توان الگوریتم‌های مختلف را به راحتی تحت آزمون قرار داد.



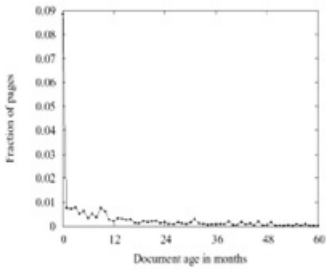
شکل ۸ بیشینهٔ عمق وبگاه‌ها

٤,٢. آمار صفحات

شـــکل‌های ۹، ۱۰ و ۱۱، به ترتیب، نمایانگر ســـن درصد صفحات برای حسب روز،ماه و سال است. همانطور که نشان داده شده سن صفحات نسبتاً بالا می‌باشند. به عبارت دیگر نرخ بروزآوری صفحات، پایین است.



شکل۹. عمر صفحات بر حسب روز



شکل ۱۰. عمر صفحات بر حسب ماه

شکل 8. بیشینه عمق وب گاه ها

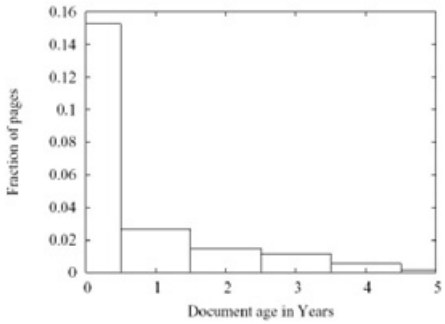
4,2.آمار صفحات

شکل های 9 ، 10 و 11 به ترتیب نمایـان گر سن درصـد صـفحات برای حسب روز،ماه و سال است. همانطور که نشان داده شـده سـن صفحات نسبتاً بالا می باشند. به عبارت دیگر نرخ بروزآوری صفحات، پایین است.

شکل 9. عمر صفحات بر حسب روز

شکل 10. عمر صفحات بر حسب ماه

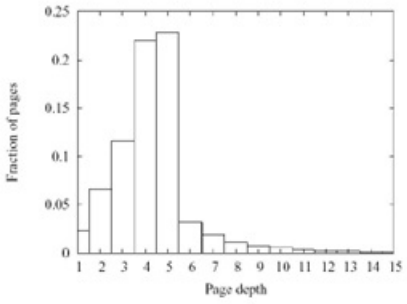
عکس



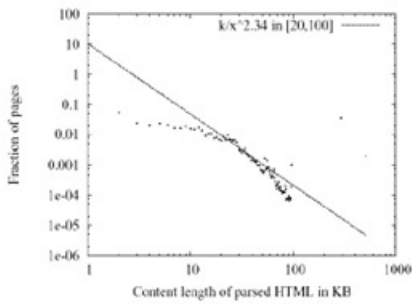
شکل ۱۱. عمر صفحات بر حسب سال

شکل شماره دوازده، عمق صفحات را نشان می‌دهد. با توجه به شکل، اکثر صفحات در عمق ٤ و ٥ قرار دارند. قابل توجه است که عمق صفحات، دارای توزیع پواسون می‌باشد.

حجم محتوای صفحات که یک توزیع تقریباً خطی است، در نمودار شماره ۱۳ نشان داده شده است. بیشتر صفحات دارای حجم بین ۱۰ تا ۱۰۰ کیلوبایت را دارا می‌باشند.



شکل ۱۲. عمق صفحات (توزیع پواسون)



شکل ۱۳. حجم محتوای صفحات بر حسب کیلوبایت (تقریباً خطی)

شـــکل شماره چهارده رابطه بین سن و حجم صفحات را نشان می‌دهد. همچنین رابطه بین عمق با سن و حجم صفحه در شکلهای ۱۵ و ۱٦ نشان داده شده است. با داشتن شاخص‌های فوق (عمر، حجم، عمق)، به راحتی می‌توان مدل مناسبی از وب ایران بدست آورد.

شکل 11. عمر صفحات بر حسب سال

شکل شـماره دوازده، عمق صفحات را نشان می دهد. با توجه به شکل، اکثر صفحات در عمق 4 و 5 قرار دارند. قابل توجه است که عمق ،صفحات دارای توزیع پواسون می باشد.

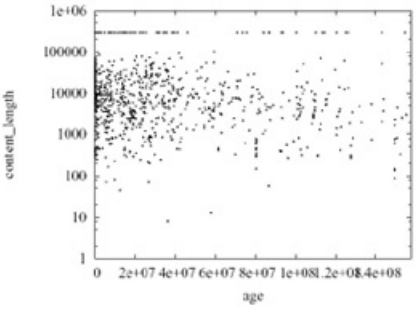
حجم محتوای صفحات که یک توزیع تقریباً خطی است، در نمودار شماره 13 نشان داده شده است. بیشتر صفحات دارای حجم بین 10 تا 100 کیلوبایت را دارا می باشند.

شکل 12. عمق صفحات (توزیع پواسون)

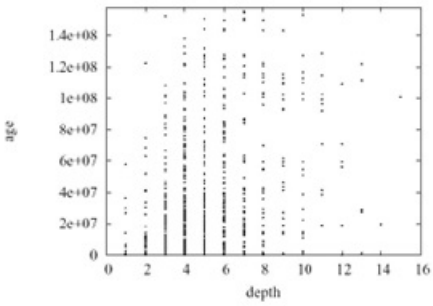
شکل 13. حجم محتوای صفحات بر حسب کیلو بایت (تقریباً خطی)

شکل شماره چهارده رابطه بین سن و حجم صفحات را نشان می دهد. همچنین رابطه بین عمق با سن و حجم صفحه در شکل های 15 و 16 نشان داده شده است. با داشتن شاخص های فوق (عمر، حجم، عمق)، به راحتی می توان مدل مناسبی از وب ایران بدست آورد.

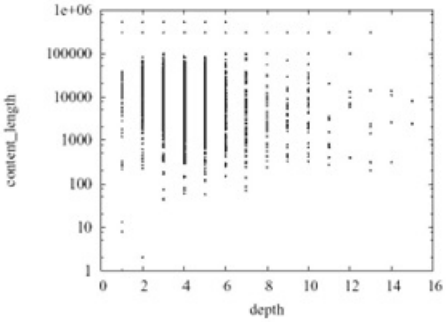
عکس



شکل ۱٤. سن صفحات بر حسب حجم محتوای صفحات



شکل ۱۵. سن صفحات بر حسب عمق



شکل ۱٦. عمق صفحات بر حسب حجم آنها

٤،۳. آمار نوع صفحات

جـــداول ۳، ٤، ٥ ، ٦، ۷، ۸ و ۹ بـــه ترتیب، آمارهای مختلفی از صفحات، اعم از ایســـتا، پویا ، صوتی، ویدیویی، تصویری و اجرایی را نشان می‌دهد. طبق جدول ۵، حدود ۵۰ درصد از صفحات، پویا هستند. در مجموع حدود ۸ میلیون آدرس صفحه ایستا و ۱۷ میلیون آدرس صفحه پویا کشف شده است. از بین صفحات پویا، بیشترین درصد را انواع PHP و ASP تشکیل داده‌اند.

جدول ۳. آمار انواع صفحات

تعداد صفحات (بر حسب میلیون)	۲
درصد صفحات ایستا	۵۰/۹۴
درصد صفحات پویا	۴۹/۰۶
درصد صفحات دونسخه‌ای (Duplicate)	۱

شکل 14. سن صفحات بر حسب حجم محتوای صفحات

شکل 15. سن صفحات بر حسب عمق

شکل 16. عمق صفحات بر حسب حجم آن ها

4,3.آمار نوع صفحات

جداول ،3، 4، 5، 6، 7، 8 و 9 به ترتیب، آمارهای مختلفی از صفحات، اعم از ایستا، پویا ، صوتی، ویدیویی، تصویری و اجرایی را نشان می دهد. طبق جدول 5، حدود 50 درصد از صـفحات، پویا هستند. در مجموع حدود 8 میلیون آدرس صفحه ایستا و 17 میلیون آدرس صفحه پویا کشف شده است. از بین صفحات پویا، بیشترین درصد را انواع PHP و ASP تشکیل داده اند.

جدول 3. آمار انواع صفحات

ص: 101

عکس

جدول ٤. آمار نوع صفحات پویای کشف‌شده

Filename extension	Number of links found	Percent
php	8,643,428	49.45%
asp	8,429,913	48.22%
jsp	158,477	0.91%
cfm	108,619	0.62%
cgi	81,769	0.47%
shtml	45,847	0.26%
pl	7,507	0.04%
dll	3,130	0.02%
fcgi	1,046	0.01%

جدول ٥. آمار درصد صفحات ایستای کشف‌شده

Filename extension	Number of links found	Percent
html	7,807,297	95.84%
pdf	220,004	2.70%
xml	63,679	0.78%
doc	15,493	0.19%
rdf	13,951	0.17%
ppt	5,731	0.07%
mso	5,715	0.07%
txt	5,335	0.07%
xls	2,763	0.03%
ps	2,328	0.03%
tex	1,641	0.02%
dvi	1,379	0.02%

جدول ٦. آمار اسناد صوتی کشف‌شده

Filename extension	Number of links found	Percent
mp3	27,839	44.14%
ram	16,436	26.06%
wma	16,216	25.71%
mid	1,744	2.77%
wav	513	0.81%
pls	173	0.27%
asf	78	0.12%
au	64	0.10%

جدول ۷. آمار ویدیوی کشف‌شده

Filename extension	Number of links found	Percent
swf	324,678	91.51%
wmv	26,666	7.52%
mpg	2,181	0.61%
avi	1,179	0.33%
mov	99	0.03%
qt	6	0%

جدول 4. آمار نوع صفحات پویای کشف شده

جدول 5. آمار درصد صفحات ایستای کشف شده

جدول 6. آمار اسناد صوتی کشف شده

جدول 7. آمار ویدیوی کشف شده

ص: 102

عکس

جدول ۸ آمار فایل‌های تصویری کشف‌شده

Filename extension	Number of links found	Percent
gif	11,441,330	81.13%
jpg	2,310,009	16.38%
png	300,407	2.13%
ico	38,371	0.27%
bmp	11,536	0.08%
wmf	694	0%
tiff	198	0%
img	2	0%
pbm	1	0%

جدول ۹. آمار نرم‌افزارهای کشف‌شده

Filename extension	Number of links found	Percent
exe	40,344	97.11%
iso	1,066	2.57%
rpm	72	0.17%
patch	42	0.10%
diff	9	0.02%
pdb	9	0.02%
deb	4	0.01%

۵. نتیجه‌گیری و کارهای آینده

هدف این مقاله، بررسی ویژگی‌های وب فارسی است. بر این اساس، سامانه نیمه خودکاری جهت انجام ارزیابی‌های مورد نظر، طراحی و پیاده‌سازی شد. از خصوصیات این سامانه خودکار بودن آن می‌باشد که در زمان‌های مختلف می‌توان آنرا اجرا و آمارهای مورد نظر را استخراج کرد. بررسی‌ها بر روی حدود ۱۱ هزار سایت با پسوند IR و مشتمل بر دو میلیون صفحه انجام شد. این مجموعه، اغلب سازمان‌های دولتی و غیر دولتی را پوشش می‌دهد. هدف اصلی پژوهش، استخراج و تعیین ویژگی‌های وب اعم از خصوصیات ساختاری نظیر ویژگی‌های گراف وب و نیز خصوصیات محتوایی وب فارسی است. شاخص‌های استخراج شده شامل توزیع محتوای وب‌وبگاه‌ها، حجم محتوای فارسی، نوع محتوا، سرویس های ارائه شده (جستجو، RSS و غیره) نرخ بروزآوری محتوا، توزیع مکانی وبگاه‌ها در داخل و خارج کشور وغیره می‌باشد. با انجام این پروژه بهتر می‌توان راهبردهای آینده مربوط به ICT را تعیین و تبیین کرد. از نتایج بارز این آمار عبارتند از: کم بودن حجم محتوای فارسی وب در مقایسه با سایر کشورها (نسبت ۱۰%)، نیاز به یک مرکز داده اینترنتی در کشور (۷۳%) سایت ها در خارج از کشور هستند)، نیاز به موتورهای جستجوی بومی(محتوا بیشتر از کدینگ های عربی استفاده کرده است) ، کم بودن سرویس‌های دولت الکترنیک (جستجو، امنیت و غیره). علاوه بر موارد فوق خروجی‌هایی مانند مجموعه لغات فارسی موجود در وب برای استفاده در خطایاب‌ها و موتورهای‌جستجو بدست آمده است.

در ادامه این کار، در نظر است تا با مطالعه تطبیقی وب فارسی در طی دوره‌های مختلف زمانی،

جدول 8. آمار فایل های تصویری کشف شده

جدول 9. آمار نرم افزارهای کشف شده

5. نتیجه گیری و کارهای آینده

هدف این مقاله، بررسی ویژگی های وب فارسی است. بر این اساس، سامانه نیمه خودکاری جهت انجام ارزیابی های مورد نظر طراحی و پیاده سازی شد. از خصوصیات این سامانه خودکار بودن آن می باشد که در زمان های مختلف می توان آن را اجرا و آمارهای مورد نظر را استخراج کرد. بررسی ها بر روی حدود 11 هزار سایت با پسوند IR و مشتمل بر دو میلیون صفحه انجام شد. این مجموعه اغلب سازمان های دولتی و غیر دولتی را پوشش می دهد. هدف اصلی پژوهش، استخراج و تعیین ویژگی های وب اعم از خصوصیات ساختاری نظیر ویژگی های گراف وب و نیز خصوصیات محتوایی وب فارسی است. شاخص های استخراج شده شامل توزیع محتوای وب وب گاه ها، حجم محتوای فارسی، نوع محتوا، سرویس های ارائه شده (جستجو، RSS و غیره) نرخ بروزآوری محتوا، توزیع مکانی وبگاه ها در داخل و خارج کشور و غیره می باشد. با انجام این پروژه بهتر می توان راهبردهای آینده مربوط به ICT را تعیین و تبیین کرد. از نتایج بارز این آمار عبارتند از: کم بودن حجم محتوای فارسی وب در مقایسه با سایر کشورها (نسبت 10%)، نیاز به یک مرکز داده اینترنتی در کشور (73%) سایت ها در خارج از کشور هستند)، نیاز به موتورهای جستجوی بومی (محتوا بیشتر از کدینگ های عربی استفاده کرده است) ، کم بودن سرویس های دولت الکترنیک (جستجو، امنیت و غیره). علاوه بر موارد فوق خروجی هایی مانند مجموعه لغات فارسی موجود در وب برای استفاده در خطایاب ها و موتورهای جستجو بدست آمده است.

در ادامه این کار، در نظر است تا با مطالعه تطبیقی وب فارسی در طی دوره های مختلف زمانی،

نرخ تغییرات شـاخص هـای مختلف تعیین گردد. این اطلاعـات، جهت گیری تغییرات وب فارسـی را نشان خواهـد داد و کمک شایانی به تصـمیم گیری های کلان حوزه فناوری اطلاعات خواهـد کرد. در این عین حال، در نظر است تا با بررسـی های بیشتر در رابطه با محتوای صفحات، اطلاعات بیشتری نظیر متوسط طول حروف در کلمات فارسـی، متوسط تعداد کلمات در اسـناد، مشخصات گراف (مانند توزیع درجه های ورودی و خروجی) برای زبان فارسی و آمارهای متنوع دیگر، تعیین گردد.

منابع

اشاره

.http://www.worldwidewebsize.com/, January 2013

Krootkaew C., A. Vongpakaymas, A. Jeawpoung 2002. Services E-readiness Explorer (SEE): Automatic ..monitoring tool for thailand e-government project in proceeding of EurAsia-ICT, Shiraz, Iran, Oct

Baeza-Yates R., C. Castillo and V. LÓpez, 2005. Characteristics of the Web of Spain. Journal of) .Cybernetics, 9(1

Baeza-Yates, R. and F. Lalanne.2004.Characteristics of the korean web. Technical report, Korea Chile IT .Cooperation Center ITCC, 2004

Bordino, I. and D. Donato.2009 .Dynamic characterization of a large Web graph, Dynamic characterization .of a large Web graph. Proceedings of the WebSci'09: Society On-Line, pp. 198-202

Broder A. Z., S.R. Kumar ,F. Maghoul ,P., Raghavan,S. Rajagopalan ,R. Stata , A. Tomkins and J.L.Wiener ..2000.Graph structure in the web. Computer Networks, 33(1): 309-320

Gabriel Tolosa, G., F. Bordignon, R. Baeza-Yates, C. Castillo .2007.Characterization of the argentinian) .web. International Journal of Scientometrics, Informetrics and Bibliometrics, 7(1

Gomes, D. and M.J. Silva.2003. A characterization of the Portuguese Web.In Proceedings of 3rd ECDL .Workshop on Web Archives, Trondheim, Norway

Keyhanipour A.H., A.M. Zare Bidoki, M. Mahmoudi and M. Azadnia.2007.Evaluation of Iran's web content from e-government perspective. Proceedings of the 12th International CSI Computer Conference, .pp. 2081-2086, Tehran, Iran

Rauber, A., O. Aschenbrenner, O. Witvoet, R.M. Bruckner, and M. Kaiser.2002.Uncovering information)

.hidden in Web archives. D-Lib Magazine, 8(12

Thelwall, M. and D. Wilkinson.2003.Graph structure in three national-academic webs: Power laws with anomalies. Journal of the American Society for Information Science and

.Technology, 54(8): 706–712

Thelwall, M. and A. Zuccala. 2008. A university–centred European Union link analysis. Scientometrics, .75(3): 407–420

Shestakov, D.2011.Sampling the national deep web, database and expert systems applications. Lecture .Notes in Computer Science, 6860: 331–340

چکیده

مقاله حاضـر گزارشـی است که توسط پژوهشـگران مؤسسهٔ اینترنت آکسفورد برای کنسرسیوم بین المللی حفاظت اینترنت تهیه شده است هدف از آن ایجاد انگیزه بحث و تبادل نظر بیشتر مابین آرشـیویست ها و پژوهشـگران وب در مورد روش های آرشـیو وب است در بخش اول مقـاله نمای کلی 4 سـناریو احتمالی از آرشـیوهای وب شامل نیروانا، آپو کالیپس انفرادی و آرشـیوهای غبار آلوده را معرفی می کنـد. سپس به توصـیف انواع مختلفی از پژوهش هایی که در مورد وب پویا در حال انجام هستند، می پردازد. و در بخش آخر چالش های فعلی و پیش روی آرشـیوهای وب را پوشـش می دهد و در نهایت به ارائه پیشنهادهای برای رفع چالش ها و ارائه راه حل های میان مدت و دراز مدت برای مقابله با تغییرات احتمالی را ارائه می کند.

آینده آرشیو وب

اشاره

*آینده آرشیو وب (1)

نوشته: اریک تی. مه یر (2)، آرتور توماس (3)، رالف شرودر (4) | مؤسسه اینترنت آکسفورد (5)

ترجمه: رضا خانی پور (6)، محبوبه قربانی (7)

خلاصهٔ اجرایی

این گزارش، توسط پژوهشگران مؤسسهٔ اینترنت آکسفورد برای کنسرسیوم بین المللی حفاظت اینترنت (8) به نگارش در آمده است. هدف از آن ایجاد انگیزه بحث و تبادل نظر بیشتر مابین آرشیویست ها و پژوهشگران وب در مورد روش های آرشیو وب است که می تواند مورد استفاده پژوهشگران قرار گیرد.

بخش اول. نمای کلی از چهار سناریوی احتمالی برای آینده:

اشاره

- نیروانا (9) : مکانی که آرشیوهای وب توسط بسیاری از گروه ها به طور گسترده استفاده، استاندارد سازی، و قابل مرور می شود و رابط کاربر قوی برای دسترسی دارند.

1- Web archives: the future(s)

2- Eric T. Meyer

3- Arthur Thomas

4- Ralph Schroeder

5- Oxford Internet Institute

6- عضو هیئت علمی سازمان اسناد و کتابخانه ملی ج. ا. ا.

7- دانشجوی دکترای علم اطلاعات و دانش شناسی واحد علوم و تحقیقات تهران دانشگاه آزاد اسلامی

8- (the International Internet Preservation Consortium (IIPC

9- Nirvana

– آپوکالیپس (1) : آرشیوها تجزیه می شوند و استاندارد نشده باقی می مانند، به سختی بازیابی شده و در دسترس قرار می گیرند و در نتیجه مفید نیست و به سختی استفاده می شوند.

– انفرادی: در این سناریو، آرشیوها مانند یک هوش هم بسته منفرد که قادر است بین اشیای رقومی و افراد ارتباط برقرار کند، غیر ضروری می شوند.

– آرشیوهای غبارآلود: در این سناریو، جامعه آرشیوی وب هرگز پاسخگوی سؤال (که چی؟) نیست آرشیوهای وب به طور گسترده، مورد استفاده قرار نمی گیرد، بلکه در حال جمع آوری غبارهای و رقومی هستند.

این سناریوها ما را به تفکر دربارهٔ تعامل بین آرشیوها پژوهشگران و محققان در روش های گوناگون قادر می سازد.

بخش 2. انواع مختلف پژوهش هایی را توصیف می کند که در مورد وب پویا در حال انجام هستند. فنی که حالا به طور گسترده ای، بیشتر از استفاده از آرشیوهای وب مورد بهره برداری قرار می گیرد.

مقصود این است که استفاده از وب پویا می تواند الهام بخش تفکر درباره کاربردهای بالقوه آرشیوهای وب باشد. این کاربردها عبارت اند از:

مصورسازی: که از طریق آن پیوندها، نه تنها بین وبگاه ها، بلکه میان انواع مختلف اطلاعات نیز

می توانند ایجاد شوند، همانگونه که سازمان ها و دیدگاه های کلی آرشیوها را توانمند می سازد.

سنجش های دگرساز (2) : دانشمندان حوزه علم سنجی شروع به کسب داده ها از منابع مجزا، از طریق تحلیل های استنادی نموده اند. به طور مثال: بلاگ های محققان و پیوندهای بین این بلاگ ها.

فنون دیگری، همچون نقشه برداری محتوای ایجاد شده توسط کاربر و تحلیل های شبکه اجتماعی در اینجا ارائه شده است.

بخش 3. چالش های فعلی و پیش رو را پوشش می دهد. نخستین قسمت از این بخش روش های تغییر وب، پیشنهادهایی در این مورد و راه حل های میان مدت و دراز مدت برای مقابله آرشیوها با این تغییرات را توصیف می نماید. این بخش با پیشنهادهایی درباره گام های پیش رو پایان می یابد.

مقدمه

این گزارش، به صورت پیش نویس در می 2012 نگارش یافت و در مجمع عمومی آی آی. پی. سی. (3) در

1- Apocalypse

2- ALT- metrics : ایجاد و مطالعه سنجش های جدید بر اساس تحلیل های شبکه های اجتماعی و اطلاع رسانی علمی (دسترسی در http://altmetrics.org/about)

3- IIPC (آی آ پی سی) کنسرسیوم بین المللی حفاظت اینترنت (http://www.netpreserve.org که این کار را پایه گذاری و زیربنایی برای بحث های بیشتر ارائه نمود که با ارائه در نشست مجمع عمومی IIPC در سال 2011 و برگزاری کارگاهی که ویرایشی نهایی از این گزارش خواهد بود آغاز می شود.

لاهه (1) ، هلند در تاریخ 10-9 می 2011 ارائه شده است.

این گزارش، در یک نشست جامع، تلخیص و در یک کارگاه نیز ارائه شده است. همچنین، از طریق پست الکترونیک برای جامعه محققان اینترنت و جامعه کتابداری و اطلاع رسانی ارسال شده است. هدف این بود که پیش نویس گزارش منجر به تهییج و ترغیب شود، افکار را برانگیزد. برای اینکه آرشیویست های وب و پژوهشگران را به خروج از سکون وادار کند. افراد مشابه و دیگران را به فعالیت وادار کند. برای ایجاد تغییر، پیش از این برای بحث و تبادل نظر ایجاد انگیزه شده است؛ که فعالیت را به سوی تغییری ملموس ترغیب می کند.

چرا تغییر ضروری است؟ وقتی که IIPC توجه ما را به اجرای این پروژه جلب کرد، این ضرورت که جامعه آرشیوی وب و IPC به طور ویژه، باید روش هایی نوین برای تشویق کاربران و کاربردهای جدید آرشیوهای وب و مدل های جدید آرشیوسازی وب و سبک های جدید تعامل با محققان بررسی نماید، احساس می شد.

این مباحث، پیش تر در قالب دو گزارش به وسیله JISC (2) تدوین شده بود که بر روی شرایط فعلی هنر آرشیوهای وب (دو گرتی و دیگران، 2010) (3) و فرصت های سرمایه گذاری جدید متمرکز شده بود.

بعضی از نتایج این دو مقاله در ادامه آمده است، اما موضوعات عمومی سرتاسر اثر این بود که هنوز فاصله و شکافی بین جامعه پژوهشگران بالقوه ای که دلیل خوبی برای استفاده، تحلیل و اشتراک آرشیو وب دارند و جامعه واقعی (به طور کلی هنوز کوچک) پژوهشگران که در حال حاضر فعالیت می کنند وجود دارد (دو گرتی و دیگران، 2010 ، ص.5). تجربه کار ما در این گزارش و صحبت با اعضای آی.آی. پی.سی. و جامعه پژوهشگران اینترنتی اندکی باعث تغییر نظر ما در مورد این موضوع شده است، مسلماً ما از اینکه همیشه کاربردهای آرشیوهای وب به خوبی مورد بحث قرار گیرد و جامعه پژوهشی هنوز به صورت معنی داری به آن دست نیافته اند، اطمینان بیشتری یافته ایم. این گزارش، به نوبه خود این [دیدگاه] را کمی تغییر خواهد داد، اما اگر پاره ای از پیشنهادهای ارائه شده در آن به طور جدی توسط جوامع مرتبط، به کار گرفته شود احتمال اینکه در آینده آرشیوهای منابع اینترنتی برای پژوهشگران اهمیت بیشتری یابد، افزایش خواهد یافت.

این گزارش، در ابتدا به منظور اینکه افکار نظری را به آینده احتمالی وب معطوف کند، به صورت یک تمرین که ما را وارد به تفکر درباره آن چه که برای انجام در حال حاضر به آن نیاز داریم و اطمینان از اینکه بتوانیم به صورت مطمئن و پر ثمر از آرشیو وب در آینده استفاده کنیم، شکل یافته است. سپس، ما بر روی روش های ابزاری مورد استفاده برای پژوهش وب پویا تمرکز می کنیم، همانند یک اشاره گر بر روی انواع چیزهایی که می تواند برای کمک در فهم وب آرشیوی، گسترش یابد. در ادامه، ما بر روی یک سری از موضوعات و سؤالاتی که پژوهشگران می خواهند یا ممکن است بخواهند برای استفاده از وب آرشیوی

1- the Hague

2- ISC کمیته سیستم های اطلاعاتی مشترک (http://www.jisc.ac.uk/) که بخش پژوهش حوزه اطلاعات و ارتباطات و توسعه زیربنایی آموزش و پژوهش در بریتانیا را پایه گذاری نمود.

3- Dougherty

به آن ها بپردازند تمرکز می کنیم. در پایان این بخش، برخی از چالش های انفرادی، سازمانی و سازمان های بین المللی که می توانند برای افزایش توانایی ما برای توضیح این موضوعات و پاسخ به این سؤالات، مورد توجه قرار گیرند را تبیین می نماییم.

ما گزارش مذکور را با نتایجی بر اساس آن چه از این تجربه آموخته ایم، به پایان می بریم.

ساختن آینده

(بهترین راه برای پیش بینی آینده، ساختن آن است) (کای، 1995) (1)

برای آغاز بحث از آینده شناسی بهره می بریم. هدف پیش بینی آینده نیست. که اشتباه خواهد بود. در واقع، ما کاملاً در مورد تلاش هایی که برای پیش بینی آینده انجام شده است، اطمینان نداریم. ساختن سناریوها و تلاش های دیگری که برای پیش بینی آینده (ادعاهایی درباره آینده) طراحی شده اند، معمولاً در دانش، نهفته می مانند، به طوری که بیشتر چنین تلاش هایی، هرگز جدی گرفته نخواهد شد.

با وجود این، حداقل یک نوع از آینده شناسی که از نظر ما مناسب است وجود دارد. [و] آن زمانی است که هدف از تجربه ، پیش بینی آینده نباشد؛ بلکه الهام بخشی باشد برای افرادی که مسئول ساختن نظام هایی هستند که زیربنای آینده خواهند بود. به منظور تفکر دربارهٔ نتایج تصمیمات فعلی در دوره هایی با تأثیرات احتمالاً طولانی مدت. آی.آی.پی.سی. متشکل از بسیاری از افرادی است که در حال حاضر عهده دار توسعه و گسترش نظام ها، ابزارها، استانداردها، و تفاهم نامه هایی برای حفاظت از محتوای اینترنت با نیم نگاهی به سوی مفید ساختن آن برای فهم جامعه ای که در آن زندگی می کنیم، هستند.

در گسترش نظام های رایانه ای، «انتخاب های معماری» بسیاری وجود دارد. (کلینگ، مک کیم و کینگ 2003) (2) ، در طول مسیر نقاطی که تصمیمات در آن جا اتخاذ می شوند، همان هایی هستند که یکی در دو راهی های طول مسیر به غیر از انتخاب های دیگر گزیده می شوند. شواهدی وجود دارد از اینکه جامعه آرشیوی وب اهمیت نقاط انتخاب در حال حاضر و در آینده نزدیک را می پذیرد. (پذیرفته است). یک سری از نقاط انتخاب با پیشنهادهایی که نوید تغییر زلزله گونه ای در آرشیو وب می دهند، مرتبط هستند: گذر از دسترسی به وبگاه ها و صفحه های انفرادی به سوی ساختن و استفاده از یک مجموعه همانند یک مجموعه در مقایسه با ایجاد دسترسی به بخش هایی از یک مجموعه. به چه تصمیماتی برای افزایش احتمال اینکه آرشیو – در – جعبه، مفید، قابل استفاده، ادامه پذیر و تاثیر گذار باشد، نیاز خواهد بود؟ما دوباره به ایدهٔ آرشیو – در یک جعبه در این مقاله، موازی با چالش های دیگری که انتخابها را می طلبد، باز خواهیم گشت.

در ادامه، هدف این تمرین این است که جامعه آرشیوی وب کدام راه را برای آن انتخاب ها که بر آینده تأثیر خواهند گذاشت و برای پیشنهاد گام ها و انتخاب هایی که آینده را در یک جهت یا جهت گیری هدایت [ترغیب] می کند، درخواست می کند.

1- Kay

2- Kling, McKim, King

سناریوها [طرح ها]

اشاره

ما می توانیم آینده های بسیار مختلفی را برای آرشیوهای وب و کاربردهای شان تصور نماییم، برای فرض بحث، ما چهار سناریوی بالقوه را که می توانند در دهه یا دو دهه آینده اجرا شوند، بررسی الزامات شان و راه های پیشنهادی که جامعه آرشیوی وب باید با آن ها هماهنگ شوند ، ارائه داده ایم:

در ادامه این متن، به طور ویژه تعدادی از عناصر که این سناریوها را تکمیل خواهند کرد، نشان می دهیم، چالش هایی را که در سر راه به کارگیری آن ها وجود دارد شناسایی می نمائیم، و مثال هایی از گستره متنوع ابزارهای توسعه وب پویا را نشان می دهیم. اگر آن ها برای داده های تاریخی به کار گرفته شوند، بین بهترین و بدترین طرح می توانند تفاوت ایجاد نمایند.

سناریوی نیروانا

در بهترین شرایط، آرشیوهای وب، مستحکم، استاندارد شده و به صورت ایمن، حفاظت شده خواهند بود. در حالی که، در زمان مشابه به صورت باز، انعطاف پذیر و گسترده، و به عنوان بخشی از ابزار استاندارد پژوهشی در علم اینترنت، علوم سیاسی، اقتصادی، جامعه شناسی، تاریخ معاصر (و در آینده، تاریخ اواخر قرن بیستم و اوایل قرن بیست و یکم) روزنامه نگاری، زبان شناسی، ارتباطات، تجارت، مطالعات رسانه و دیگر رشته ها استفاده می شوند.

علاوه بر دانشگاه، آرشیوهای وب برای عموم، حکومت ها، واحدهای سیاسی، و گروه های مشاوران

(متفکران) و سازمان های غیر دولتی مفید و کاربردی خواهد بود. متأسفانه، از خیلی از جنبه های بسیاری، کم احتمال ترین سناریوست، به این سبب که به نظر می رسد در حال حاضر، برای امکان پذیر ساختن ورودآن به جامعه آرشیوی وب به تلاش گسترده تر و منابع وسیع تری نیاز خواهد بود.اگر چه، ممکن است برای اینکه آن را در ذهن خود به صورت مطلوب نگهداری کنیم، مفید به نظر برسد، همان گونه که ایجاد توازن را بین آن چه که می توانیم و آن چه باید باشد بررسی می کنیم.

به منظور ارائه این سناریو، حتی در طرح کلی، نیاز است که اتفاق هایی رخ دهد (در بخش بعدی، مثال هایی از وب پویا در زمان وقوع، ارائه خواهیم کرد) این موارد عبارت اند از:

– توسعه قوی تر و مؤثرتر ابزارهای جست و جوی متن، تحلیل و استخراج اطلاعات، مصورسازی، وبلاگ نویسی اجتماعی، تحلیل های طولی و تحلیل های نظری.

– توسعه روش های بهتر برای اینکه کاربران، گشتالت مجموعه های چندتایی و انفرادی را درک کنند. در حالی که محتوای متنی می تواند جست و جو شود، نیاز به فراداده غنی برای پشتیبانی خطوط کلی محتوا و یا روش های جدید سازماندهی آن می باشد.

افراد به طور ویژه در تشخیص الگوهای بصری توانمند هستند، و به نظر می رسد ابزارهای گرافیکی یکی از بهترین راه های کسب آن می باشند. می توانیم ایجاد فضای مجازی را تصور کنیم که به (1 fly throgh) های سه بعدی و سایر روش های فضایی حسی امکان

سازماندهی محتوا را می دهد. در بهترین

1- Fly through: نوعی شبیه سازی کامپیوتری است که در فضاهای مجازی به کاربران اجازه مشاهـده مدل های مجازی سایت ها را می دهد. (دسترسی در:http://dictionary.reference.com/browse/fly-through)

حالت، محیط های مجازی کاملاً همه جانبه از نوع غار (1) (شرودر ، 2011) می توانند همکاری گروه های توزیع شده فضایی (مکانی) افراد را پشتیبانی کرده و امکان اشتراک مؤثر و تعامل اجتماعی را ایجاد کننـد. از این دیـدگاه ، تمامیت آرشـیوهای وب می تواند به صورت یـک کلیـت بزرگ در نظر گرفتـه شـود، کـه افراد می تواننـد در آن به صـورت منفرد یـا گروهی سـیر نماینـد. وب فعلی (و به همین ترتیب آرشـیوهای وب) درک نـاچیزی از سازمانـدهی فضـایی ارائه می دهنـد و نشـانه های مناسب فضایی یا قابلیت های دیگر کمک به مرور و کاوش را محدود (سـلب) می کنند. بر خلاف کتابخانه فیزیکی، اسناد رقومی در محیط جدا سازی شده از یکدیگر نگهداری می شوند، در مکانی که اسناد در دنیایی سه بعدی سازماندهی می شوند تا امکان مرور آن ها به واسطه گردش (جست و جو) در محیط را میسر ساخته و نشانگرهایی همچون همجواری فضایی (مکانی) به کشف آن ها کمک می کنند.

– در حالی که ابزارهای وب نوشت نویسی اجتماعی پدید می آیند، آرشیو های وب، فاقد سایر ابزار های همکاری هستند، مانند موتورهای پیشنهادی (مخزن پیوسته آمازون مثال عالی از این امکان است).

– به طور فزاینـده ما نیازمنـد محتوای آماده شده توسط کاربران (وب 2) در مقیاسـی بزرگ (فیس بوک (2)) هستیم. امـا ساختاری حیرت انگیز بر اسـاس چنین محتوای سازمـان نیـافته نامتجانس و ذاتاً سازمان نیافته ای، در این مقیاس با مشـکل مواجه است. انجام آن به وسـیله ماشـین از نظر فنـاوری بسـیار چـالش برانگیز است، بخشـیدن غنـای معنـایی، به عنوان جایگزینی برای پشتیبانی رهیافت انباشت منبع، به کاربران آرشیو اجازه می دهد که محتوا را سازماندهی کنند این شکل نهایی وب نوشت نویسی (حاشیه نگاری) اجتماعی است که کاربران نه فقط داده بلکه فراداده را نیز تولید می کنند (گازان، 2008) (3) .

در نیروانا، انتخاب های امروز توسط پژوهشـگران آینـده سـتایش خواهـد شـد، کسانی که به اتکای اطلاعات و شواهد، تلاش های انسانی انجام شـده در اینترنت را، تجسم داده، حفاظت کرده، و افزایش داده انـد تا به همه روش های فنون قدرتمنـد پژوهش دست یابند و توانمند شوند.

سناریوی آپوکالیپس (آخر الزمان)

در بـدترین حالت، تغییرات روزافزون اینترنت برای بسط و توسـعه فناوری های جدیـد (HTML5، محتوای قابل اجرا، ویـدئوی جاسازی شده و اشیای تعاملی، وبگاه های مبتنی بر پایگاه داده، نرم افزارهای موبایل غیر وابسته به HTML.TTP و مانند آن) با سرعتی گیج کننده، ادامه خواهند یافت، و ابزارهای آرشـیو سازی وب در حفظ همگامی با آن ها شـکست خواهند خورد، و بیش از پیش عقب می افتند. حتی اگر فناوری های آرشـیوسازی وب بتوانند سـرعت را حفظ کنند، تغییرات دائمی (پیوسـته) همه اشـکال مطرح شده، چالشی حل نشدنی را ایجـاد می نمایـد. در این طرح به طور صادقـانه ای، فقـط انـدکی از محتوای واقعی ذخیره می شود، و حتی اگر ذخیره شود، افزودنی هـای اختصاصی برای مشاهده آن حفظ نشده یا قابل حفظ

---

1 – CAVE: فضـایی شبیه غـار در محیـط مجازی (دسترسـی در: http://en.wikipedia.org/wiki/Cave-automatic-virtual-environment)

2– Facebook

3– Gazan

شـدنی نیسـتند، و مشاهـده محتوا غیر ممکن می شود. بیشتر پیشـینه های پیوسـته در طول دوره ما سـرانجام همچون کارت پانچ های دهه 1960 و نوارهای مغناطیسـی ریل به ریل (1) ، غیرقابل خواندن (اسـتفاده) خواهند شد. علاوه بر مشـکل قالب، مشکل حجم زیاد نیز در حال رشد و فزونی است.همان طور که اینترنت به سـمت تکامل استفاده از IPV6 (2) حرکت می کند، اشیای قابل نشانه گذاری (به طور روزافزونی شامل، اشیای فیزیکی در «Web of Things» (3) به راستی، به وسـیله دستورات متعدد ذخیره سازی - حتی نشانی ها - بیش از ظرفیت موجودمان، در حال ازدیاد حجم هسـتند (1038)، که به تنهـایی اجـازه آماده سازی همه محتوا را می دهـد. در نتیجه، ما نمی توانیم موارد بیشتری را جست و جو کنیم، و به این دلیل نمایه سازی و فناوری جست و جوی ما ناامیدانه شکست می خورد.

همزمان با توسـعهٔ وب معنایی، همه مفهوم محتوا تغییر می کنـد. محتوا چیزی بیشتر از متن و تصویر نیست، اما در حال حاضـر اقلام داده قراردادی و پیوندهای بین آن ها را در بر می گیرد. حتی حالا در سال 2011، جهان داده های پیوند خورده عمومی (4) (که شامل مجموعه هایی همچون data.gov و data.gov. uk است). ده ها میلیون اقلام داده ای (به طور روز افزون قالب RDF (5)) را در برمی گیرد که به وسـیله صدها میلیون پیوند قابل ارجاع مجدد، به هم پیوسـته اند. چالش آرشیوسازی این سری داده ها از زمان نشانی دهی - آغاز می شود (برای مثال: به وسیله پروژه پرونوم (6) آزمایشگاه های آرشیوهای ملی بریتانیا (7))، اما احتمالا جهان داده های پیوند خورده سریع تر از اینکه از نظر مجموعه سازی یا تحلیل، قابل مدیریت باشد، رشد خواهند کرد.

در این سـناریو، حتی منابع عظیم شـرکتی چون گوگل تحت الشـعاع مشـکلات قرار می گیرد. بنابراین، پاسخ بدیهی برای آرشـیوسازی این حجم («بگذارید گوگل کارش را بکند» (8)) چیزی بیشتر از یک راه حل نیست.

اگر انتخاب ها ما را به این مسـیر هدایت کند، پژوهشـگران آینده می آموزند که گذشته وب را به صورت غیر قابل دسترسی و بدون اعتماد و داسـتانی که حکـایت می شود و مـدرک دست دومی از زمان، تصور کننـد. امروزه، حجم بزرگی از اطلاعات به مقیاس جهانی ایجاد می شود که ممکن است به صورت نوشـته های روی تکه های کاغـذ در میلیون ها جعبه کفش ذخیره شـده باشـد، که همگان، پیشـرفت های صورت گرفته در دنیا را همان گونه که در محتوای اینترنت منعکس می شود به نحو مناسبی درک خواهند کرد.

1- reel to reel

2- IPV6 آخرین نسخه پروتکل ارتباطات در اینترنت (دسترسی در http://en.wikipedia.org/wiki/IPv6)

3- Web of Things (http://en.wikipedia.org/wiki/Web_of_Things)

4- Linked Data - Connect Distributed Data across the Web, at www.linkeddata.org

5- RDF (Resource description Framework) (چارچوب توصیف منابع)

6- PRONOM

7- http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

8- let Google do it

سناریوی انفرادی

در دنیایی کاملاً متغیر، افراطی ترین سناریو سناریوی انفرادی است که در آن اینترنت همانگونه که ما می شناسیم به سمت پدیده ای کاملاً جدید و احتمالا در نوع خود هوشمند (کورزویل 2005) (1) تکامل می یابد. در مسیر منحصر به فرد شدن ، به سمت توسعه و تبدیل شدن به موجود مجازی پیچیده ای حرکت می کند، که ما فهم کمی از آن داریم و برای آرشیو کردن آن، در حال حاضر روشی بهتر از آن چه که با هوش انسانی خود از آرشیو می دانیم، وجود ندارد. حتی اکنون، در سال 2011، برای غلبه بر آن آغاز به تشخیص تمایز بین پردازش انسانی و مصنوعی کرده ایم. خدماتی همچون: ریکاپچا (2) (ون آون (3) ، مورر (4) ، مک میلن (5) ، آبراهام (6) و بلام 2008) (7) و آمازونز مکانیکال ترک (8) که از موجودیت های انسانی «در یک چرخه» برای حل مشکلاتی که برای ماشین ها سخت است، استفاده می کنند، به ما راه هایی را به سوی دنیایی که در آن هوش انسان و ماشین به صورت دوقلوهای جدایی ناپذیری هستند، نشان می دهند و مرز بین آن ها نامشخص است. در چنین دنیایی، مشخص نیست که آرشیوسازی چه مفهومی می تواند داشته باشد. بنابراین، همزمان با حرکت زمان به سمت جلو، گذشته به ناچار و به صورت جبران ناپذیری از دست می رود. این سناریو، ممکن است شبیه به یک داستان علمی به نظر برسد. اگر چه، ارزشمند است که به خاطر داشته باشیم آینده غیر قابل پیش بینی است، حتی اگر تأثیر بر روی شرایط انتخاب در طول مسیر را برای اینکه در یک مسیر یا مسیر دیگری قرار گیرد – مدیریت نماییم. انتخاب هایی که ما انجام می دهیم ممکن است، برای اجرای وظیفه ای جدید، همچون اینترنت هوشمند ناکافی باشد. این بدین معنا نیست که ما باید انتخاب های صحیح را در نظر بگیریم و تلاشی از خود به خرج ندهیم.

سناریوی غبار آلود

متأسفانه این سناریو، احتمالاً در ابتدای امر [این اندیشه را به ذهن متبادر می کند] که آرشیوهای وب معادل آرشیو رقومی غبارآلود است، اگر چه اغلب به خوبی حفاظت و نگهداری می گردد. به سختی از آن استفاده می شود.

حتی اگر جامعه آرشیوسازی ،وب، به توسعه استانداردها و تجربیات برای حفظ بخشهای اینترنت ادامه دهد، کاربردهای مؤثر ناچیزی از جامعه پژوهشی سر می زند و جامعه پژوهشی به طور مؤثر از آن استفاده نمی کند.

ممکن است به صفحه های اینترنت به صورت مجزا با استفاده از ابزارهای پیوسته رجوع شود و بعضی از پژوهشگران به ایجاد آرشیوهای کوچک برای موضوعات پژوهشی خاص مبادرت ورزند ، اما پژوهش اینترنت تنها با اولویت تمرکز بر روی وب پویا ادامه خواهد داشت و در آینده نزدیک توجه اندکی به

1- Kurzweil

2 – reCAPTCHA : برنامه ای برای رقومی کردن کتب، نشریات و ... دسترسی در (http://www.google.com/recaptcha/learnmore

3- Von Ahn

4- Maurer

5- Mc Millen

Abraham –6

Blum –7

(Amazon's Mechanical Turk(Amazon Corp., Mechanical Turk at www.mturk.com –8

استفاده از وب قبلی برای پژوهش جدی ، گسترش خواهد یافت.

این سناریو با سناریوی آخر الزمان متفاوت است. در آن سناریو، فناوری آرشیوسازی وب با تغییرات فناورانه روی اینترنت همگام نخواهد بود. در این سناریو سرعت آرشیوسازی وب هماهنگ با فناوری تحویل وب تنظیم می شود.

اگر چه داده حفاظت شده به صورت اسطوره ای حفاظت شده برای استفاده در آینده نامطمئن باقی می ماند.

در مراحل نگارش این گزارش، مشخص شده است که به جای آرشیوهای وب مرجع، به طور روزافزونی کاربران و پژوهشگران با وب پویا همانند آرشیو برخورد می کنند. وب پویا، به رشد خود ادامه می دهد و برای بیشتر بخش ها، داده هایی که ناپدیدمی شوند، از نظر بسیاری به صورت یک مشکل ساده دیده می شود و مهم تر از آن، برای بخشی بزرگ تر، حجم عظیم دیگری از داده ها بر روی وب در هر زمان باقی می ماند.

تصویر ما از آرشیو، تصویری از اقلام فیزیکی همچون کاغذها و اسناد ذخیره شده در یک مکان فیزیکی است. با وجود این، این ماهیت وب نیست. پژوهشگران می توانند حجم زیادی از مواد را از منابع مختلف وب پویا برای نیل به اهداف پژوهشی خود ذخیره کنند. ما آرشیوها را به صورت چیزهایی که برای آیندگان حبس شده اند، درک می کنیم. در حالی که وب، به تنهایی پدیده ای در حال رشد و منبعی متنوع از انواع مواد است که به صورت بالقوه مورد توجه پژوهشگران است. به طوری که، آن را نه به عنوان یک آرشیو سنتی، بلکه به سادگی به صورت یک منبع داده می بینند.

این یک سناریوی بدبینانه است، اما به نظر می رسد که وزنی از شواهد را در جانب خود داشته باشد. در رایزنی هایی با محققان برجسته، به یک فقدان علاقه دیرپای بر پرسش از وب قبلی و درک اینترنت به عنوان یک توسعه تاریخی، رسیدیم. البته استثناهایی وجود دارد که بعداً در این گزارش شرح داده خواهد شد، اما قادر بوده ایم تا گرایش نهفته کار با آرشیوهای وب را کشف کنیم که انتظار ساده ای برای [ظهور] فناوری مناسب برای بیدار شدن آن وجود دارد.

ممکن است انتظار تغییر از گوشه و کنار برای آمادگی ایجاد یک گام تغییری در تصورات پژوهشگران، بر اساس جلوه های استفاده جدید یا فناوری تازه، وجود داشته باشد. اگر این امر رخ ندهد، بیم آن می رود که به جمع آوری گردو غبارهای رقومی ادامه دهیم.

اگر از این سناریو اجتناب شود، نیازمند نوعی جدیدی از آرشیویست هستیم که با پژوهشگران و عموم [کاربران]، در استخراج داده هایی که آن ها از وب پویا نیاز دارند در تعامل باشد و زمانی که این داده ها از وب پویا ناپدید شدند، قادر به ذخیره سازی مجدد آن ها به صورتی که قابل مشاهده و استفاده بوسیله ابزارهای وب پویا باشند، هستند. بسیار فراتر، از زمانی که رقومی سازی اسناد تاریخی، از آرشیوها حجم زیادی از مواد تاریخی قابل دسترسی بر روی وب در دهه گذشته ایجاد کرده است (مه یر 2001 (1) ،، تانر 2010 (2) ، تانر و دیگان 2011 (3) ، آرشیوهای وب به انتقال وب به محفظه ها (جعبه ها) نیاز ندارند، بلکه در عوض نیازمند برگشت به وب در زمانی که دارای محتواست، هستند.

Meyer -1

مرور آینده

در حین این که ما مسیر خود را به سوی آیندهٔ آرشیو وب مرور می کنیم، پرسش های متعددی در مورد چگونگی دیدگاه مان نسبت به آینده، می توانیم از خود بپرسیم.

برای مثال، آیا آرشیو آینده، باغ دیوار کشیده شده ای خواهد بود که از آسیب ها محفوظ و در امان است، اما با دسترسیِ محدود است؟ آیا فضایی کاملاً باز و قابل دسترسی برای تازه واردان خواهد بود؟ آیا مجموعه ای از مخازن است یا یک عرصهٔ ارتباط متقابل منفرد؟ یا اینکه یک شهر ارواح خالی از سکنه باز و مرتبط خواهد بود؟

بخشی از آرشیو ها برای محققان و دانشمندان و بخشی دیگر برای عموم است (1) . آیا این دو می توانند در فضایی سایبرنتیک با هم مرتبط باشند – تا اینکه دانشمندان به صورت دائمی، آن چه که عموم [شامل دانشمندان] به آن دسترسی دارند و از آن استفاده می کنند و آن را ایجاد می کنند، پایش نمایند (بنابراین درک از آگاهی جهانی یا دریافت جمعی را افزایش می دهند) در حالی که در زمان مشابه شکل گیری چنین فضایی با این سبک، برای گسترش، مجموعه سازی و تأثیرگذاری و لذت بخش بودن و کاربرد غنی همه دنیا، غنی سازی می شوند ؟

حتی در صورتی که منحصر به فرد بودن ، با شکست مواجه شود، اینترنت به طور فزاینده ای اجازهٔ ارتباطات در این سبک را که از نظر استفاده برای هوش جهانی خیلی کم قابل فهم می باشد، می دهد (شرودر و مه یر، 2009). در هوش جهانی – با خوراک ها و پیوندهای پویا – هوش های انفرادی از طریق ابزارهای ورودی و خروجی به یکدیگر مرتبط هستند. چگونه آرشیوهای وب ماهیت ارتباطات درونی هوش جهانی را به صورت غیر از اسناد به ظاهر جدا از هم منعکس می نماید؟

اینها و بسیاری از سؤالات دیگر ما را با حرکتی رو به جلو مواجه می کند. در بخش های آینده این سند، ما نگاهی خواهیم داشت به بعضی از فنون فهم وب پویا که می تواند الهام بخش جامعه آرشیوسازی وب باشد. سپس به چالش های حرکت رو به جلویی که وب آرشیوی را به صورت بالقوه برای تحقیق، ارزشمندتر می کند، اشاره خواهیم نمود.

یادگیری از وب پویا

باید پرسید که چرا در دنیای پژوهشی اینترنت، آرشیوهای وب به صورت شهروندان درجه دوم به نظر می رسند؟ در مقایسه با کسانی که در وب پویا مطالعه می کنند، به مراتب تعداد کمتری از پژوهشگران از آرشیوهای وب استفاده می کنند و تعداد کمی از ابزارهای غیر تخصصی برای آرشیوهای وب در حال ساخت هستند، به ویژه در مقایسه با ابزارهایی که برای مطالعه وب پویا ساخته می شوند.

چالش عمومی برجسته در این بخش این است که، جامعه آرشیوسازی وب به ایجاد ارتباط بین منابعی که تولید می کنند با ابزارهای لبه برش (2) نیاز دارند، که توسط متخصصان رایانه، پژوهشگران، توسعه

1- مـا در این جا با کاربرد های تجاری و دولتی آرشـیوی وب اسـناد که غالبا برای برآوردن نیازهای حقوقی طراحی می شونـد، سـر و کار نداریم، همچنان که این مورد فراتر از حیطه اختیار و تخصص ماست.

2- (نام نوعی فناوری) cutting edge

دهندگان مستقل، و هکرها برای مطالعه وب پویا گسترش یافته اند.

در حال حاضر، انواع ابزارهای توسعه یافته برای مطالعه وب پویا، روی هم رفته، به آسانی برای مطالعه داده های قابل دسترسی آرشیوهای وب به کار گرفته می شوند. این [امر] بزرگ ترین مانع برای فهم وب، نه تنها به عنوان یک تصویر لحظه ای بلکه به عنوان یک اکوسیستم در حال توسعه است.

مصورسازی

هر آرشیوی که ایجاد شده و مورد استفاده قرار می گیرد به [مرور] وسیع، غیر قابل نظر اجمالی و بدون نقشه یا دسترسی بصری و روش حسی برای مشاهده آرشیوها و چگونگی پیوندهای آن ها، خواهد شد. مصورسازی در اینجا یک راه حل اساسی است، اما انواع متعددی از آن ها در دسترس است.

چالش ها

ابزارهای بسیاری برای بررسی ارتباط بین کاربران رسانه های اجتماعی، ابزارهایی برای مرور پیوستار زمانی، ترکیب نقشه ها با ردپای کاربران، ابزارهای تصویری برای نمایش نحوه پیوند داده ها و نظیر آن وجود دارد؛ اما اینها به منظور کار کردن با آرشیوهای وب نیاز به متمرکز شدن دارند. مصور سازی اطلاعات حوزه پیشرفته پژوهشی است، اما بهترین نحوه مشاهده یک آرشیو (یا جست و جوی یکی از آن ها، یا دیدن ارتباطات درونی و ما بین آن ها) شامل رابط های حسی (بصری)، تغییرات دیداری تصاویر سه بُعدی و پویشگران زمانی، هنوز دست نیافتنی هستند؛ و برای مثال آیا احتمال شناسایی موضوع هایی درون مجموعه به وسیله روش های بازبینی تصویری یا سازماندهی تصویری یک سری از ارتباطات وجود دارد؟ به بیان دیگر ، ایجاد ابزارهای مصورسازی برای کمک به پژوهشگران ممکن است؟

مثال ها: Touch grahp (2)، (Apple Time Machine (1)

عکس

دهندگان مستقل، و هکرها برای مطالعه وب پویا گسترش یافته‌اند.

درحال حاضر، انواع ابزارهای توسعه یافته برای مطالعه وب پویا، روی هم رفته، به‌آسانی برای مطالعه داده‌های قابل دسترسی آرشیوهای وب به‌کار گرفته می‌شوند. این [امر] بزرگ‌ترین مانع برای فهم وب، نه تنها به‌عنوان یک تصویر لحظه‌ای بلکه به‌عنوان یک اکوسیستم در حال توسعه است.

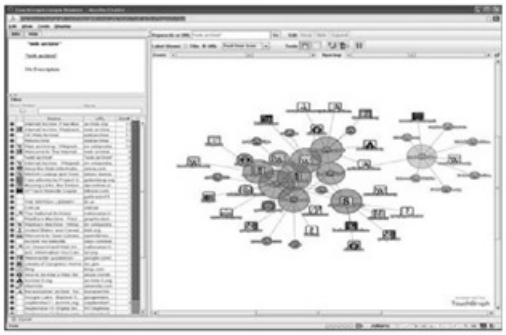
مصورسازی

هر آرشیوی که ایجاد شده و مورد استفاده قرار می‌گیرد به [مرور] وسیع، غیرقابل نظر اجمالی و بدون نقشه یا دسترسی بصری و روش حسی برای مشاهده آرشیوها و چگونگی پیوندهای آنها، خواهد شد. مصورسازی در اینجا یک راه‌حل اساسی است، اما انواع متعددی از آنها در دسترس است.

چالش‌ها

ابزارهای بسیاری برای بررسی ارتباط بین کاربران رسانه‌های اجتماعی، ابزارهایی برای مرور پیوستار زمانی، ترکیب نقشه‌ها با ردپای کاربران، ابزارهای تصویری برای نمایش نحوه پیوند داده‌ها و نظیر آن وجود دارد؛ اما اینها به‌منظور کار کردن با آرشیوهای وب نیاز به متمرکز شدن دارند. مصور سازی اطلاعات حوزه پیشرفته پژوهشی است، اما بهترین نحوه مشاهده یک آرشیو (یا جست‌وجوی یکی از آنها، یا دیدن ارتباطات درونی و ما بین آنها) شامل رابطهای حسی (بصری)، تغییرات دیداری، تصاویر سه بُعدی و پویشگران زمانی، هنوز دست نیافتنی هستند؛ و برای مثال، آیا احتمال شناسایی موضوع‌هایی درون مجموعه به‌وسیله روش‌های بازبینی تصویری یا سازماندهی تصویری یک سری از ارتباطات وجود دارد؟ به بیان دیگر ، ایجاد ابزارهای مصورسازی برای کمک به پژوهشگران ممکن است؟

مثال‌ها: Touch grahp[1]، Apple Time Machine[2]



تصویر ۱. Touchgraph، در اینجا، توانایی بررسی پیوندهای میان وبگاه‌ها با استفاده از رابط گرافیکی را ارائه می‌کند. داده‌ها از وب پویا ترسیم شده‌اند.

1. Touchgraph(http://www.touchgraph.com/)
2. Apple Time Machine (http://www.apple.com/macosx/what-is-macosx/time-machine.html)

تصویر 1 . Touchgraph، در اینجا، توانایی بررسی پیوند های میان وب گاه ها با استفاده از رابط گرافیکی را ارائه می کند. داده ها از وب پویا ترسیم شده اند.

ص: 117

(Apple Time Machine (http://www.apple.com/macosx/what-is-macosx/time-machine.html -1

(/Touchgraph(http://www.touchgraph.com -2

برنامه های کاربردی جست و جو همانند شکارچی

همـان طوری که اطلاعـات اینترنت به تکثیر و افزایش، هم در حجم و هم در تنوع انواع محتوا ادامه می دهنـد، جست و جوهای به مراتب پیچیده تری برای قابلیت اسـتخراج هر چیز با معنا و استفاده از این مجموعه عظیم مورد نیاز خواهد بود. جست و جو مانند جست و جوی تصویر و ویدئو به سوی تکلیفی پیچیده تر سوق می یابد.

چـالش: ایجـاد سـطح بلنـد پروازانه تری از برنامه ها با هزینه ای تقریباً ناچیز، به ویژه ابزارهایی که برای مجموعه ها می تواننـد به کار گرفته شوند. این ،امر ممکن است نیازمند طراحانی برای اسـتفاده جسورانه از موتورهای جست و جوی مبتنی بر فناوری ابر و زبان های جست و جوی بهتر با قابلیت انعطاف در پرسش سؤال های پیچیده از داده های موجود در آرشیوهای وب، باشد.

مثال ها: (در حال حاضـر آپاچی) یاهو! طرح زیربنایی 1) (PIG Latin(http://pig.apache.org/) برای پشتیبانی موردی تحلیل سری داده های خیلی بزرگ (تصویر 2)

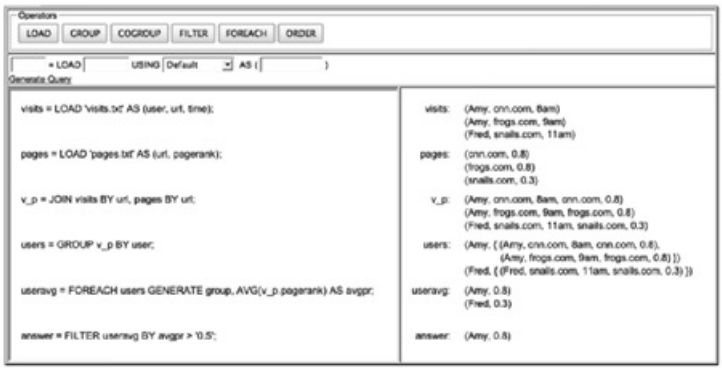
عکس

برنامه‌های کاربردی جست‌وجو همانند شکارچی

همان‌طـــوری کـــه اطلاعات اینترنت به تکثیر و افزایش، هم در حجم و هـــم در تنوع انواع محتوا ادامه می‌دهند، جســـت‌وجوهای به مراتب پیچیده‌تری برای قابلیت استخراج هر چیز با معنا و استفاده از این مجموعه عظیم مورد نیاز خواهد بود. جســـت‌وجو مانند جست‌وجوی تصویر و ویدئو به سوی تکلیفی پیچیده‌تر سوق می‌یابد.

چالش: ایجاد ســـطح بلند پروازانه تری از برنامه‌هابا هزینه‌ای تقریباً ناچیز، به‌ویژه ابزارهایی که برای مجموعه‌ها می‌توانند به‌کار گرفته شوند.این امر، ممکن است نیازمند طراحانی برای استفاده جسورانه از موتورهای جست‌وجوی مبتنی بر فناوری ابر و زبان‌های جست‌وجوی بهتر با قابلیت انعطاف در پرسش سؤال‌های پیچیده از داده‌های موجود در آرشیوهای وب، باشد.

مثال‌ها: (درحال حاضر آپاچی) یاهو! طرح زیربنایی PIG Latin[1] برای پشتیبانی موردی تحلیل سری داده‌های خیلی بزرگ (تصویر ۲)



تصویر ۲. نمایی از Pi٦ Pen نشان دهندهٔ برنامه‌ای است که افراد متقاضی مشاهده صفحه‌های با رتبه‌بندی بالا را می‌یابد. (اولستون[2]، رید[3]، سریواستاوا[4]، کومر[5] و تامکینز[6]، ۲۰۰۸)

تحلیل‌های شبکه اجتماعی

تحلیل شـــبکه اجتماعی (SNA) حوزهٔ قابل توجهی از علاقه و فعالیت پژوهشـــی در میان پژوهشگران اینترنت، جامعه شناسان، فیزیکدانها و بسیاری دیگر است. تنوع موضوع‌های مورد توجه بسیار گسترده

1. PIG Latin(http://pig.apache.org/)
2. Olston
3. Reed
4. Srivastava
5. Kummer
6. Tomkins

تصویر 2. نمایی از Pi6 Pen نشان دهنـده برنامه ای است که افراد متقاضـی مشاهده صـفحه های با رتبه بندی بالا را می یابد. (اولسـتون (2)، رید (3)، سریواستاوا، (4) کومر، (5) و تامکینز، 2008 (6))

تحلیل های شبکه اجتماعی

تحلیل شبکه اجتماعی (SNA) حوزهٔ قابل توجهی از علاقه و فعالیت پژوهشی در میان پژوهش گران اینترنت، جامعه شناسان، فیزیک دان

ها و بسیاری دیگر است. تنوع موضوع های مورد توجه بسیار گسترده

ص: 118

1- PIG Latin

2- Olston

3- Reed

4- Srivastava

5- Kummer

6- Tomkins

است که شامل فهم ارتباطات بین دوستان در شبکه های اجتماعی مانند فیس بوک (هوگان 2010 (1)) ، بررسی وابستگی های سیاسی مشارکت کنندگان در مباحث سیاسی (هیندمن (2) ، 2007) ، و کشف شبکه های توطئه(طرح) در ادبیات انگلیسی (مورتی 2011 ، 2005) (3) است.

ابزارهای جست و جوی مبتنی بر تحلیل های شبکه اجتماعی شامل NodeXL (7) , Voson (6) , Pajec (5) , UCINET (4) ، و بسیاری دیگر، در حال تکثیر و توسعه هستند.

اگر چه تعداد کمی از این ابزارها، هر چه که باشد، برای استفاده به وسیله آرشیوهای وب توانمند یا بهینه شده اند.

چالش: نخست، کار با طراحان اصلی ابزارهای تحلیل شبکه اجتماعی برای توانمند و بهینه سازی آن ها در کار با داده های آرشیوی وب.

همچنین، توسعهٔ روش های جدید ابداعی با احتمال یک بار در بعد زمان، که به داده های شبکه برای ردیابی مواردی همچون تکامل تدریجی شبکه های اجتماعی در طول زمان، افزوده می شود. به وسیله آرشیوسازی، نه تنها وضعیت سایت های شبکه های اجتماعی، بلکه زمانی که افراد پیوندها را ایجاد، نگهداری و حذف نموده، با دیگران ارتباط برقرار کرده، به گروه ها می پیوندند و گروه ها یا وبگاه ها را ترک می کنند.

ما باید به خاطر داشته باشیم که وب شبکه ای از پیوندهاست و تحلیل وب بینشی، کلی به ماهیت آن شبکه برای ما فراهم می کند.

مثال ها:

تحلیل های فیس بوک: ابزارهای بسیاری برای تحلیل تعاملات بین کاربران فیس بوک، جریان نفوذ و منحنی اجتماعی در دسترس هستند.

نمایش مثالی از فیس بوک (8) :

عکس

اســـت که شامل فهم ارتباطات بین دوستان در شـــبکه‌های اجتماعی مانند فیس بوک (هوگان[1]،۲۰۱۰) ، بررســـی وابستگی‌های سیاسی مشارکت کنندگان در مباحث سیاسی (هیندمن[2]،۲۰۰۷) ، و کشف شبکه های توطئه(طرح) در ادبیات انگلیسی(مورتی[3]،۲۰۰۵،۲۰۱۱) است.

ابزارهای جســـت‌وجوی مبتنی بر تحلیل‌های شبکه اجتماعی شـــامل،Voson[6] ،Pajec[5] ،UCINET[4] NodeXL[7]، و بسیاری دیگر، در حال تکثیر و توسعه هستند.

اگر چه تعداد کمی از این ابزارها، هر چه که باشد، برای استفاده به‌وسیله آرشیوهای وب توانمند یا بهینه شده‌اند.

چالش: نخست، کار با طراحان اصلی ابزارهای تحلیل شبکه اجتماعی برای توانمند و بهینه‌سازی آنها در کار با داده‌های آرشیوی وب.

همچنین، توسعهٔ روش‌های جدید ابداعی با احتمال یک‌بار در بعد زمان، که به داده‌های شبکه برای ردیابی مواردی همچون تکامل تدریجی شبکه‌های اجتماعی در طول زمان، افزوده می‌شود، به وسیله آرشیوسازی، نه تنها وضعیت سایت‌های شبکه‌های اجتماعی، بلکه زمانی که افراد پیوندها را ایجاد، نگهداری و حذف نموده، با دیگران ارتباط برقرار کرده، به گروه‌ها می‌پیوندند و گروه‌ها یا وبگاه‌ها را ترک می‌کنند.

ما باید به خاطر داشته باشیم که وب شبکه‌ای از پیوندهاست و تحلیل وب بینشی، کلی به ماهیت آن شبکه برای ما فراهم می‌کند.

مثال‌ها:

تحلیل‌های فیس بوک: ابزارهای بسیاری برای تحلیل تعاملات بین کاربران فیس بوک، جریان نفوذ و منحنی اجتماعی در دسترس هستند.

نمایش مثالی از فیس بوک[8]:

تصاویر نمودارهای اجتماعی فیس بوک:

1. Hogan
2. Hindman
3. Moretti
4. http://www.analytictech.com/ucinet/
5. http://pajek.imfm.si/doku.php
6. http://voson.anu.edu.au/
7. http://nodexl.codeplex.com/
8. http://www.facebook.com/insights/

تصاویر نمودارهای اجتماعی فیس بوک:

ص: 119

Hogan -1

Hindman -2

Moretti -3

/http://www.analytictech.com/ucinet -4

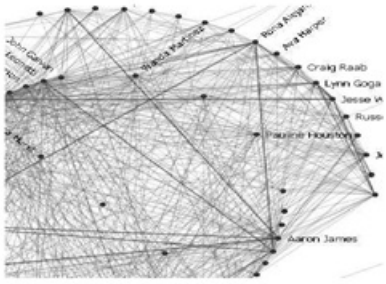
http://pajek.imfm.si/doku.php -5

/http://voson.anu.edu.au -6

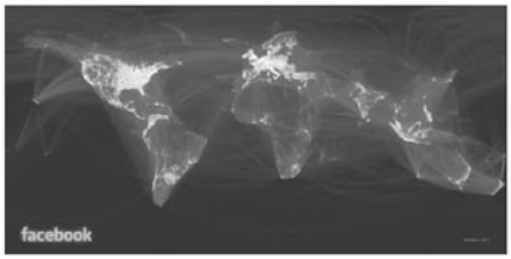
/http://nodexl.codeplex.com -7

http://www.facebook.com/insights/ -8

عکس

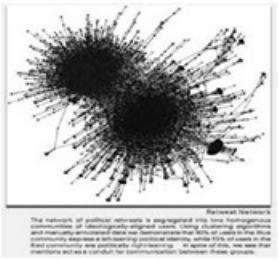


تصویر ۳. منبع:
http://infosthetics.com/archives/2008/03/facebook_social_network_graph.html

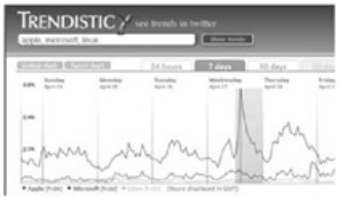


تصویر ٤. منبع :
http://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919

به همین ترتیب، طیف گسترده‌ای از تحلیل‌های توییتر مانند Twitalyzer [1]، Trendistic [2]، و دیگران وجود دارد.



تصویر ٦. Truthy (http://truthy.indiana.edu/)



تصویر ٥. trndistic

1. http://www.twitalyzer.com/
2. http://trendistic.com/

تصویر 3. منبع: http://infosthetics.com/archives/2008/03/facebook_social_network_graph.html

تصویر 4 منبع: http://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919

به همین ترتیب، طیف گسترده ای از تحلیل های توییتر مانند Twitalyzer (2) (1)، Trendistic)، و دیگران وجود دارد.

تصویر 5. trndistic

تصویر 6. Truthy (http://truthy.indiana.edu/)

1- http://www.twitalyzer.com/

2- http://trendistic.com/

عکس

Far Left	Moderate Left	Center	Moderate Right	Far Right
#healthcare	#aarp #women	#democrats #social	#rangel #waste	#912project #twisters
#judaism #hollywood	#citizensunited	#seniors #dnc	#saveamerica	#gop2112 #israel
#2010elections	#democratic	#budget #political	#american #gold	#foxnews #mediabias
#capitalism #recession	#bankster #energy	#goproud #christian	#repeal #mexico	#constitution
#security #dreamact	#sarahpalin	#media #nobel	#terrorism #gopleader	#patriots #rednov
#publicoption	#progressives		#palin12	#abortion
#topprogs	#stopbeck #iraq			

تصویر ۸. برچسب‌های نشانه‌گذاری شده به‌وسیله کاربران حوزه سیاست (کونوور[1]، راتکویچ[2]، فرانسیسکو[3]، و دیگران؛ ۲۰۱۱)

تصویر۷. توییت کردن و سیاست (کونوور،راتکویچ، گونکالوز[4]،فلامینی[5] و منسزر[6]،۲۰۱۱)

سنجش‌های دگرساز

اصطلاحی درحال پدیدارشدن برای راه‌های جدید اندازه‌گیری تأثیر علمی، فراتر از معیارهای کتابسنجی، وب‌سنجی و علم‌سنجی است.

ارتباطات بین ما و بین دانشمندان و درون گروه‌های علمی به‌طور روزافزونی بر روی وب درحال شکل‌گیری است. جامعه درحال ظهوری از پژوهشگران که مطالعه پژوهشی انجام می‌دهند، از توالی‌ها و پیوندهایی که به‌وسیله ابزارهایی چون توییتر، مندلی[7]، فرند فید و بسیاری دیگر برجای گذاشته شده است، برای درکی که بسیار سریع‌تر از تأثیرات سنتی می‌توانند توسعه یابند، استفاده می‌شوند.

فراتر از چیزی که می‌توان به‌عنوان سنجش دگرساز تصور کرد: چگونگی ردیابی مشارکت‌های غیرآکادمیک به سمت دانش است.

آیا می‌توان ابزارهای تحقیقاتی پژوهشگران را برای سایر حوزه‌های غیرتخصصی به‌کار برد؟ برای مثال، آیا می‌توانیم تأثیر مشارکت کنندگان انفرادی بر گروه‌های سرگرمی در طول زمان را با استفاده از معیارهای مشابه برای درک چگونگی تحول تأثیر پژوهشگر، ارزیابی کنیم؟

چالش: فعال‌سازی روش‌هایی به مراتب ساده تر برای مشخص کردن طیف زمانی مواد رقومی، بنابراین، تحلیل سنجش دگرساز می‌تواند به روش‌های کاملاً مشابه کتابسنجی انجام شود: در الگوی

1 . Conover
2. Ratkiewicz
3.Francisco
4. Goncalves
5. Flammini
6. Menczer
7. Mendeley

تصویر 8. برچسب هاي نشانه گذاري شده به وسیله کاربران حوزه سیاست (کونوور (1)، راتکویچ (2)، فرانسیسکو (3)، و دیگران؛ 2011)

سنجش های دگرساز

اصـطلاحی در حال پدیدار شدن برای راه های جدید اندازه گیری تأثیر علمی، فراتر از معیارهای کتاب سـنجی، وب سـنجی و علم سنجی است.

ارتباطـات بین مـا و بین دانشـمندان و درون گروه هـای علمی به طور روزافزونی بر روی وب در حـال شـکل گیری است. جـامعه در حال ظهوری از پژوهشـگران که مطالعه پژوهشـی انجام می دهند، از توالی ها و پیوندهایی که به وسیله ابزارهایی چون توییتر مندلی (4) ، فرند فید و بسیاری دیگر برجای گذاشته شده است، برای درکی که بسیار سریع تر از تأثیرات سنتی می توانند توسعه یابند، استفاده می شوند.

فراتر از چیزی که می توان به عنوان سنجش دگر ساز تصور کرد: چگونگی ردیابی مشارکت های غیرآکادمیک به سمت دانش است.

آیـا می توان ابزارهـای تحقیقـاتی پژوهشـگران را برای سـایر حوزه هـای غیر تخصصـی به کار برد؟ برای مثال، آیا می توانیم تأثیر مشارکت کنندگان انفرادی بر گروه های سـرگرمی در طول زمان را با استفاده از معیار های مشابه برای درک چگونگی تحول تأثیر پژوهش گر، ارزیابی کنیم؟

چـالش: فعـال سـازی روش هایی به مراتب ساده تر برای مشـخص کردن طیف زمانی مواد رقومی، ،بنابراین، تحلیل سـنجش دگر ساز می تواند به روش های کاملاً مشابه کتاب سنجی انجام شود: در الگوی

Conover -1

Ratkiewicz -2

Francisco -3

Mendeley -4

انتشارات رسـمی، هر نشـریه ای یک نویسنده و تاریخ نشر دارد و برای ردیابی توالی استنادها به یک اثر انفرادی استفاده می شود. نشریات، نظام آرشیوی دارند که آزمایش شده و به نحو قابل اعتماد و مناسبی پذیرفته شده اند: نشریه تخصص با وجود این، نشریات غیر رسمی وب، روش توسـعه یافته مناسب و مشابهی برای آرشـیو سازی سـهم ها (مشارکت ها) با دانش، به شـکلی که قابل اسـتناد باشند و در طی زمان دوباره جایگزین شوند، ندارند. این شکافی است که انتظار می رود رفع شود و آرشیوهای وب زمینه های بارزی برای شروع هستند.

همچنین، برای مشـارکت های غیر تخصصـی چه ابزارهایی برای تحلیل ویکی ها و موجودیت های مشترک اسـتفاده می شود که می تواند فهم این تغییرها را در دوره های زمانی توسعه دهد؟

مثال ها: 1) Data Cite) و 2) Reader Meter)

عکس

انتشـــارات رسمی، هر نشریه‌ای یک نویسنده و تاریخ نشر دارد و برای ردیابی توالی استنادها به یک اثر انفرادی استفاده می‌شود. نشریات، نظام آرشیوی دارند که آزمایش شده و به نحو قابل اعتماد و مناسبی پذیرفته شده‌اند: نشـــریه تخصص با وجود این، نشریات غیر رسمی وب، روش توسعه یافته مناسب و مشابهی برای آرشیو سازی سهم‌ها (مشارکت‌ها) با دانش، به شکلی که قابل استناد باشند و در طی زمان دوباره جایگزین شوند، ندارند. این شکافی است که انتظار می‌رود رفع شود و آرشیوهای وب زمینه‌های بارزی برای شروع هستند.

همچنین، برای مشـــارکت‌های غیرتخصصی، چه ابزارهایی برای تحلیل ویکی‌ها و موجودیت‌های مشترک استفاده می‌شود که می‌تواند فهم این تغییرها را در دوره‌های زمانی توسعه دهد؟

مثال‌ها: Reader Meter[1] و Data Cite[2]

تصویر۲. reader meter نوعی از ابزار Alt metric برای درک نحوه خواندن یک نویسنده، براساس آمارهای مندلی (http://www.mendeley.com/) منبع: http://readermeter.org

وب‌نوشت (حاشیه نگاری) اجتماعی

کاربران مایل‌اند که بتوانند پیوندها و برچسب‌هایشـــان[3] و همچنین نظرها و حاشیه‌نویسی‌هایشان را بر روی منابع نشـــر دهند. برای پژوهشـــگران، فهم چگونگی توسعه این جوامع در طول زمان و نگهداری خودشان سؤال مهمی است. برای مثال Reddit بیش از ۸ میلیون خواننده انحصاری و یک میلیون بازدید صفحه در هر ماه دارد (جسرا، ۲۰۱۱).

چالش‌ها: گســـترده‌ای راکه آرشـــیوها می‌توانند نه فقط وبگاه‌ها و مجموعه‌هایشان ، بلکه پیوندها و

1. ReaderMeter
2. Data Cite
3.bookmark

تصـویر 2 . reader meter نـــــوعی از ابزار Alt metric برای درک نحـوه خوانـــدن یـــک نویســـنده بر اســـاس آمارهـای منـــدلی (http://www.mendeley.com/) منبع: http://readermeter.org

وب نوشت (حاشیه نگاری) اجتماعی

کاربران مایل اند که بتوانند پیوند ها و برچسب های شان (3) و همچنین نظر ها و حاشیه نویسی های شان را بر روی منابع نشر دهند. برای

پژوهشگران فهم چگونگی توسعه این جوامع در طول زمان و نگهداری خودشان سؤال مهمی است. برای مثال Reddit بیش از 8 میلیون خواننده انحصاری و یک میلیون بازدید صفحه در هر ماه دارد (جسرا ، 2011).

چالش ها: گستره ای را که آرشیوها می توانند نه فقط وب گاه ها و مجموعه های شان ، بلکه پیوند ها و

Data Cite -1

ReaderMeter -2

bookmark -3

حاشیه نویسی های آن صفحه ها و مجموعه ها را ذخیره نمایند، در نظر بگیرید. توانایی پاسخ به این سؤال که «چگونه افراد این منابع را برای یکدیگر نشانه گذاری می نمایند، و چگونه در طول زمان تغییر می کند؟» و بررسی استفاده از فناوری های موجود ترکیب و تطبیق دادن ابزارهای حاشیه نویسی اجتماعی موجود. گامی بیشتر بردارید، آیا می توان اجتماعی از پیوندها و وب نوشت ها برای مجموعه های آرشیوی با استفاده از ابزارهای مشابه مورد تأکید قرار داد تا معلوم شود افراد چگونه موارد موجود روی وب پویا را برای یکدیگر نشانه گذاری می کنند؟

مثال ها: (1 Delicious)، روایت بر مبنای بوک مارکلت (2) مثال: (3 Madcow)

معماران جدید

برای استفاده از داده های وب به منظور درک اتکای دنیا بر منابعی چون (4 API)، و داده های پیوند شده، فعالیتی تعریف شده است تا داده ها را برای استفاده و هدف گذاری مجدد و ترکیب [آماده سازی]، قابل نماید. دلیل مهمی که چرا وب پویا نسبت به وب آرشیوی بسیار فعالانه تر مورد جست و جو قرار می گیرد، این است که طراحان ابزار به داده های وب پویا یا از طریق خزش گرهای وبگاه ها به طور مستقیم و یا از طریق APIها به گوگل / یاهو ،توییتر، فیسبوک و مانند آن می توانند دسترسی یابند. حتی با وجود محدودیت بعضی از این APIها دلیل اصلی، پژوهش هایی هستند که با استفاده از وب پویا رونق گرفته اند.

API روش قدرتمندی برای ساخت نرم افزار های کاربردی جدید است که بر روی داده ها طراحی شده و منابع چندگانه داده را ترکیب می نماید.

پژوهشگری به ما گفت: من در مورد استفاده از برچسب نشانه گذاری اصلی در توییتر تحقیق می کنم، و محدودیت شان را در استفاده از API در حالتی که بیشترین آشفتگی را در حین توییت کردن دارند می یابم، من می خواهم به آن ها در حالی که هنوز برخط و در دسترس هستند دسترسی داشته باشم، اگر چه استقرار آن ها ، یا به عبارت دیگر اجرای گزارش ها یا انباشت آن ها کاملاً مشکل است. برای مثال آن ها Twapper keeper را محدود به خدمت قابل دسترسی کردند که اجازه آماده سازی گزارش هایی برای کار در مورد این برچسب ها که از آن ها شناخت داشتم را می داد.

سؤال بعدی این است که چگونه API ها می توانند آرشیو شوند و چه زمانی محدود یا متوقف می شوند؟ چگونه منابعی که تاکنون آرشیو شده اند از API تأثیر می گیرند؟ آیا روش هایی برای حفاظت محتوا همراه با حفظ و احترام به حقوق مالکان و دوره های مجوز وجود دارد؟

چالش: چگونه داده های وب قدیمی از طریق API ها باز شده و پیوند داده می شوند، تا افراد هوشمند خارج از آن بتوانند برای کاربرد راه های جدید تولید دانش آن ها را ترکیب کنند [؛ البته] به وسیله ارائه ابزارهایی به افراد، برای تعریف روال های کاری به صورتی انعطاف پذیرتر از اینکه وادار به استفاده از

1- Delicious(http://www.delicious.com/)

2- bookmarklet

3- MadCow(http://www.web-notes.com/)

4- API: Application Program Interface (رابط یا میانجی برنامه های کاربردی)

ابزارهایی یک منظوره یکپارچه بشونـد. یک رهیافت، پیاده سازی کارکردهای تحلیلی ماننـد خدمات وب خواهد بود که می تواند با گردش کار قانونی موتور ترکیب شود. این رهیافت به صورت گسترده ای در bioinformatic که با مشکلات مشابهی در خصوص یکپارچه سازی داده ها از مخازن متعدد روبه روست استفاده می شود.

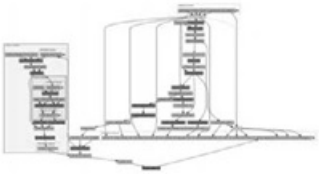
مثال: گردش کار قانون – مصوب موتور Taverna (1) برای ترکیب [تلفیق] خدمات وب پیشنهادی توسط پردازش توزیع شده و نظام های ذخیره سازی اسـتفاده می شود، سـپس گردش کارها می تواننـد به اشتراک در آمده، دوباره اسـتفاده و هدفگذاری شوند. myexperience (2) مثالی از یک مخزن قابل اشتراک گردش کارهای علمی:

عکس

ابزارهایی یک منظوره یکپارچه بشوند. یک رهیافت، پیاده‌سازی کارکردهای تحلیلی مانند خدمات وب خواهد بود که می‌تواند با گردش کار قانونی موتور ترکیب شـــود. این رهیافت به‌صورت گسترده‌ای در bioinformatic که با مشکلات مشابهی درخصوص یکپارچه سازی داده‌ها از مخازن متعدد روبه‌روست، استفاده می‌شود.

مثـــال: گردش کار قانون ـ مصوب موتور Taverna[1] برای ترکیب [تلفیق] خدمات وب پیشـــنهادی توسط پردازش توزیع شـــده و نظام‌های ذخیره‌سازی استفاده می‌شود، سپس گردش کارها می‌توانند به اشـــتراک در آمده، دوباره استفاده و هدفگذاری شوند. myexperience[2] مثالی از یک مخزن قابل اشتراک گردش کارهای علمی:

تصویر ۱۰: منبع: کارول گلوبال و دیگران [3]



تصویر ۱۱. گردش کارهای تاورنا

ماشین‌های اجتماعی

تیم برنزلی و همکارانش در مورد اینکه وب در حال تبدیل شـــدن به یک ماشین اجتماعی است، بحث کرده‌اند. به این معنی که فقط یک مخزن اطلاعات نیست، بلکه زیرساختی برای همکاری در رفع مشکل اجرای وظایف موجودیت‌های انسانی است که به‌آسانی نمی‌توانند توسط ماشین انجام شوند. دانشمندان علوم اجتماعی علاقه‌مند به درک تعامل بین فناوری و علوم اجتماعی، می‌خواهند بدانند که چگونه افراد و فناوری برای حل وظایف پیچیده‌ای که هر یک به تنهایی از عهده رفع آن بر نمی‌آیند، با یکدیگر همکاری می‌کنند.

چالش: چگونه می‌توانیم تجربه و تعامل بین کاربران و ماشین‌های اجتماعی بر روی وب را ذخیره و درک کنیم؟ همه امور اجتماعی در مورد تعامل‌هاست. اگر نتوانیم تعاملات را درک کنیم، هرگز نمی‌توانیم آنچه را که درباره ماشین اجتماعی است درک کنیم.

مثال: Amazon's Mechanical Turk[4] مکانیســـمی برای توزیع مشـــکلات بین انسان‌های خبره و جمع‌آوری راه‌حل‌هاست. گردآوری منابع، می‌تواند برای رفع مشکلات سخت به‌کار رود. به‌عنوان مثال

1. Taverna (http://www.taverna.org.uk/)
2. http://www.myexperiment.org/
3.Carol Global (http://nar.oxfordjournals.org/content/38/suppl_2/W677.full)
4. Amazon›s Mechanical Turk (https://www.mturk.com/mturk/welcome)

تصویر 11. گردش کارهای تاورنا

تصویر 10: منبع: کارول گلوبال و دیگران (3)

ماشین های اجتماعی

تیم برنزلی و همکارانش در مورد اینکه وب در حال تبدیل شدن به یک ماشین اجتماعی است، بحث کرده اند. به این معنی که فقط یک مخزن اطلاعات نیست، بلکه زیرساختی برای همکاری در رفع مشکل اجرای وظایف موجودیت های انسانی است که به آسانی نمی توانند توسط ماشین انجام شوند. دانشمندان علوم اجتماعی علاقه مند به درک تعامل بین فناوری و علوم اجتماعی، می خواهند بدانند که چگونه افراد و فناوری برای حل وظایف پیچیده ای که هر یک به تنهایی از عهده رفع آن بر نمی آیند، با یکدیگر همکاری می کنند.

چالش: چگونه می توانیم تجربه و تعامل بین کاربران و ماشین های اجتماعی بر روی وب را ذخیره و درک کنیم؟ همه امور اجتماعی در مورد تعامل هاست. اگر نتوانیم تعاملات را درک کنیم، هرگز نمی توانیم آن چه را که درباره ماشین اجتماعی است درک کنیم.

مثال : (4 Amazon's Mechanical Turk) مکانیسمی برای توزیع مشکلات بین انسان های خبره و جمع آوری راه حل هاست گردآوری منابع می تواند برای رفع مشکلات سخت به کار رود. به عنوان مثال

1- (/Taverna (http://www.taverna.org.uk

2- /http://www.myexperiment.org

3- (Carol Global (http://nar.oxfordjournals.org/content/38/suppl_2/W677.full

4- (Amazon›s Mechanical Turk (https://www.mturk.com/mturk/welcome

(1 CAPTCHA) برای شناسایی نویسه نوری کمک کننده انسان است. افراد چگونه با این ابزارها و از طریق اینها با یکدیگر تعامل می کنند؟

برای مثال، در مورد «بازهای هدف دار» همچون (2 Foldit)، چالش نه در آرشیوسازی وبگاه ها و نه فقط در بازی بلکه در چگونگی بازی افراد است. کاربران چگونه با بازی تعامل دارند؟ از تحقیق در مورد چگونگی بازی بازیکنان در تعامل با فضای پیوسته، دروسی را می توان طراحی نمود (برای مثال: ویلیامز (3)، یی (4) و کاپلان (5)، 2008).

شبکه های نقشه برداری

به طور روزافزونی جغرافی دانان از داده های اینترنت برای دریافت اطلاعات مکان ها، جریان ها، جهات و ثروت، فقر و تغییر شکل محتوا، و تأثیر در طول زمان و فضا استفاده می کنند.

چالش: استخراج خودکار اطلاعات جغرافیایی از پیوندهای درونی و بیرونی درون یک مجموعه که می تواند نقشه برداری شود. این چالش، در حال حاضر در وب پویا وجود دارد و حتی در زمان افزایش پیچیدگی تغییرات در طول زمان، بیشتر هم می شود. در حال حاضر، بخش عمده ای از اطلاعات که می تواند با استفاده از روش های دو بعدی نمایان شود به داده های سه بعدی و چهار بعدی (همانند اسلایدر (6) ها) نیاز خواهند داشت. تا اطلاعات جغرافیایی را که در طول زمان تغییر می یابند، ایجاد کند برای مثال، درک تأثیر جغرافیایی درون و ما بین دانشگاه ها، دولت ها، و شرکت ها در طول زمان از نظر تئوری محتمل است، اما به استخراج اطلاعات جغرافیایی از داده های ساختار نیافته وب نیاز دارد.

مثال: (7 Floating Sheep)

1- CAPTCHA: برنامه ای کامپیوتری است برای اجرای آزمون های نرم افزاری که فقط انسان می تواند به آن پاسخ دهد. (دسترسی در http://www.google.com/recaptcha/captcha)

2- Foldit

3- Williams

4- Yee

5- Caplan

6- Slider : رابط های کاربر گرافیکی برای ارتباط افراد با نرم افزار در ابزارهایی نظیر موبایل (دسترسی در :http://www.wisegeek.com/in-computing-what-is-a-slider.htm

7- Floating Sheep (http://www.floatingsheep.org)

عکس

تصویر۱۳. نقشه فلوتینگ شیپ «نقشه جغرافیایی مذهبی گوگل»

علم وب

علم وب تلاشی است که توسط محققان برای مطالعه وب به‌عنوان یک «مصنوع اطلاعاتی»، فهم چگونگی رشد و تکامل آن و کیفیت توسعه جوامع آن، انجام می‌گیرد.

چالش‌ها: نیاز به ابزارهای قدرتمندی برای تحلیل «نمودار وب» به‌عنوان شیء ریاضی. مکان شناسی آن چیســـت؟ چگونه گروه‌ها تکامل می‌یابند ؟ چه نوع رتبه‌بندی قانونی بـــه‌کار می‌رود؟ آیا واقعاً وب به‌وسیله قانونی قوی مدیریت می‌شود؟ چگونه اطلاعات روی وب منتشر می‌شوند؟

وب تنها یک فضای اطلاعاتی نیست، بلکه مجموعه‌ای پیچیده و خانواده‌ی از روابط درونی فضاهای فرعی اســـت که محتوای اطلاعاتی آن در بعضی مواقع توســـط جوامع مجزا تعیین می‌شـــوند. چگونه اطلاعات، بین این صفحه‌ها به اشتراک گذاشته و منتشر می‌شوند؟

برای پاسخ به این سؤال، به توسعه ابزارهایی نیاز داریم که قادر به ردیابی تکامل و انتقال مفاهیم در طول زمان و فضاهای مختلف باشند (برای مثال: بین وبلاگ‌نویسی و رسانه «mainstream»).

مثال: Media Cloud [1] و Recorded future [2]

1. Media Cloud (http://cyber.law.harvard.edu/research/mediacloud)
2. Recorded future (https://www.recordedfuture.com/)

تصویر 13. نقشه فلوتینگ شیپ «نقشه جغزافیایی مذهبی گوگل»

علم وب

اشاره

علم وب تلاشی است که توسط محققان برای مطالعه وب به عنوان یک «مصنوع اطلاعاتی»، فهم چگونگی رشد و تکامل آن و کیفیت توسعه جوامع آن، انجام می گیرد.

چالش ها: نیاز به ابزارهای قدرتمندی برای تحلیل «نمودار وب» به عنوان شیء ریاضی مکان شناسی آن چیست؟ چگونه گروه ها تکامل می یابند؟ چه نوع رتبه بندی قانونی به کار می رود؟ آیا واقعاً وب به وسیله قانونی قوی مدیریت می شود؟ چگونه اطلاعات روی وب منتشر می شوند؟

وب تنها یک فضای اطلاعاتی نیست، بلکه مجموعه ای پیچیده و خانواده ی از روابط درونی فضاهای فرعی است که محتوای اطلاعاتی آن در بعضی مواقع توسط جوامع مجزا تعیین می شوند. چگونه اطلاعات، بین این صفحه ها به اشتراک گذاشته و منتشر می شوند؟

برای پاسخ به این سؤال، به توسعه ابزارهایی نیاز داریم که قادر به ردیابی تکامل و انتقال مفاهیم در طول زمان و فضاهای مختلف باشند (برای مثال: بین وبلاگ نویسی و رسانه «mainstream»).

مثال: Media Cloud (1) و Recorded future (2)

1- Media Cloud (http://cyber.law.harvard.edu/research/mediacloud)

2- Recorded future (https://www.recordedfuture.com/)

عکس

تصویر ۱۴. Rcorded Futur سیر اخبار را در طی زمان را دنبال می کند.

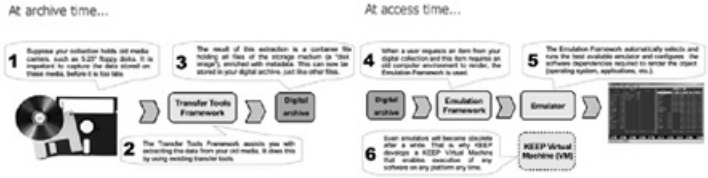
تصویر ۱۵. Media Cloud سیر اخبار جغرافیایی را دنبال میکند.

درک تجربه به‌جای محتوا

به‌طور فزاینده، دانشمندان به سمت درک اهمیت چگونگی استفاده افراد از محتوای وب و نه فقط محتوا به‌خودی خود، سوق پیدا کرده‌اند. این امر، به حساب وضعیت تجربه وب و محتوای قابل اجرا گذاشته می‌شود: تجربه درباره اینکه کدام طرح، کدام مرورگر یا افزونه‌ها[1] یا رمزگذارهای دوطرفه مورد استفاده قرار می‌گیرد و به‌طور فزاینده به موقعیت مکانی کاربران وابسته است.

چالش‌ها: برای درک تجربه‌ها به توانایی تکرار تجربه نیازمندیم، طرح‌های زیربنایی، سیستم‌های عملیاتی، مرورگرها، و به همین ترتیب تغییر تجربه وب.

مثال‌ها: Browsershots[2] و KEEP[3] (نگهداری نمونه‌سازی محیط‌های قابل انتقال)



تصویر ۱۶. KEEP (نگهداری نمونه سازی محیط‌های قابل انتقال)

تحلیل وب معنایی و مجموعه داده‌های پیوند شده

مجموعه داده‌های پیوند شده به‌سرعت با حداقل ۲۸/۵ بیلیون پیوند سه تایی در مجموعه‌های شناخته شده در

1. plugins
2. Browsershots (http://browsershots.org)
3. KEEP (http://www.keep-project.eu)

تصویر 14. Rcorded Futur سیر اخبار را در طی زمان را دنبال می کند.

تصویر 15. Media Cloud سیر اخبار جغرافیایی را دنبال می کند.

درک تجربه به جای محتوا

به طور فزاینده دانشـمندان به سمت درک اهمیت چگونگی استفاده افراد از محتوای وب و نه فقط محتوا به خودی خود، سوق پیدا کرده اند. این امر، به حساب وضعیت تجربه وب و محتوای قابل اجرا گذاشته می شود: تجربه درباره اینکه کدام طرح کدام مرورگر یا افزونه ها (1) یا رمزگذارهای دوطرفه مورد استفاده قرار می گیرد و به طور فزاینده به موقعیت مکانی کاربران وابسته است.

چالش ها: برای درک تجربه ها به توانایی تکرار تجربه نیازمنـدیم، طرح های زیربنایی، سیسـتم های عملیاتی، مرورگرها، و به همین ترتیب تغییر تجربه وب.

مثال ها: (2 Browsershots) و (3 KEEP) (نگهداری نمونه سازی محیط های قابل انتقال)

تصویر 16. KEEP (نگهداری نمونه سازی محیط های قابل انتقال)

تحلیل وب معنایی و مجموعه داده های پیوند شده

مجموعه داده های پیوند شده به سرعت با حداقل 28/5 بیلیون پیوند سه تایی در مجموعه های شناخته شده در

1- plugins

2- Browsershots (http://browsershots.org)

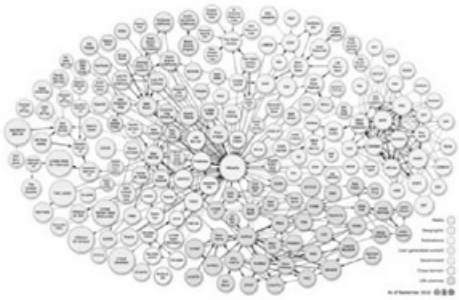
3- KEEP (http://www.keep-project.eu)

حال رشـد هسـتند. ابزارها توسط وب معنایی یا جامعه داده های پیونـد شـده توسـعه می یابنـد، به طوری که می توانند به سـرعت مشـکل مـدیریت فراداده را آسـان کرده و به این فراداده هـا اجازه دهنـد که در مقیاس بزرگ قابل جست و جو بوده و همچنین برای یکپارچه سازی داده ها در مجموعه به روشی بسیار پیچیده تر استفاده شوند.

عکس

۱۲۸ مدیریت منابع اطلاعاتی وب

حال رشد هستند. ابزارها توسط وب معنایی یا جامعه داده‌های پیوند شده توسعه می‌یابند، به‌طوری‌که می‌توانند به سرعت مشکل مدیریت فراداده را آسان کرده و به این فراداده‌ها اجازه دهند که در مقیاس بزرگ قابل جست‌وجو بوده و همچنین برای یکپارچه سازی داده‌ها در مجموعه به روشی بسیار پیچیده‌تر استفاده شوند.



تصویر ۱۷. منبع: نمودار ابرگونه داده های باز پیوند خورده
توسط ریچارد کی گانیاک[1] و انجا ینتس[2] /http://lod-cloud.net

مثال‌ها: Sig. ma / Sindic [3] جست‌وجوی مبتنی بر آر دی اف [4]

تصویر۱۸. Sig. ma / Sindic جست‌وجوی مبتنی بر RDF.
منبع : http://sig.ma/search?q=Tim/۲۰Berners/۲۰Lee

1. Richard Cyganiak
2. Anja Jentzsch
3. http://sig.ma/
4. Resource Description Framework (RDF)

تصویر 17. منبع: نمودار ابر گونه داده های باز پیوند خورده توسط ریچارد کی گانیاک (1) و انجا ینتس (2) http://lod-cloud.net

مثال ها: (3 Sindic / Sig.ma) جست و جوی مبتنی بر آر دی اف (4)

تصویر 18. Sindic / Sig. ma جست و جوی مبتنی بر RDF

منبع: http://sig.ma/search?q=Tim%20Berners%20Lee

ص: 128

1- Richard Cyganiak

2- Anja Jentzsch

3- http://sig.ma/

4- Resource Description Framework (RDF)

چالش های فعلی و آینده

اشاره

اما گام های پیش رو برای جامعه آرشیوی وب چیست؟

بسیاری از مواردی که محققان [به آن ها] نیاز دارند، به وضوح قابل مشاهده هستند، ولی این به معنی در دسترس بودن آن ها نیست. در گزارش های قبلی GISC که با همکاران دیگر نوشتیم و در قسمت های پیشین توضیح دادیم (دوگرتی و دیگران 2010؛ توماس و دیگران، 2010) توجیه های متعددی پیرامون موضوع اصلی ارائه دادیم:

ساخت جامعه، ساخت ابزار، و منابع و طراحی تجربه ها (دو گرتی و دیگران، 2010، ص27-29). قصد نداریم که فهرستی کامل از پیشنهادهای گزارش های قبلی را در اینجا مطرح کنیم. 22 پیشنهاد در گزارش دو گرتی و همکارانش و بیش از 20 پیشنهاد در گزارش توماس و همکارانش وجود دارد. بنابراین، خواننده را به مرور این گزارش ها همانند این گزارش توصیه می کنیم با این حال می توانیم برای نیل به هدفمان بعضی از موضوع های مهمی را که محققان به آن ها علاقمند هستند، برجسته کرده و چالش هایی را شناسایی کنیم که آرشیوها برای کمک به آرشیوهای وب در تبدیل آن ها به بخشی از ابزارهای استاندارد برای محققان رشته های متنوع، با آن ها مواجه هستند.

این بخش از گزارش، موضوع ها و سؤال هایی را بر می شمارد که گروه های مختلف محققان می خواهند بپرسند یا از آرشیو وب - آرشیو - در - جعبه - می خواهند بپرسند. ما آن دسته را که چالش ها و راه حل های بالقوه ای دارند شناسایی کرده ایم بعضی از این راه حل ها، به طور ویژه راه حل هایی کوتاه مدت هستند که می توانند در سطح مؤسسه ها انجام شوند. بسیاری از رهیافت های نه چندان گسترده به نگاه وسیع تری در سطح ملی منطقه ای یا بین المللی توسط سازمان هایی چون IIPC نیاز دارند.

ما از آرشیو - در - جعبه یاد کردیم، زیرا در میان بعضی افراد این تفکر وجود دارد که وب آرشیوی فراتر از روزهای اولیه خودش در ایجاد صفحه های قابل دسترسی برای تحلیل های آینده، (تفکر تعامل سنتی (1) برای پایگاه Wayback Machin که به کاربر اجازه می داد به طور عمده به صفحه های منفرد آرشیو دسترسی داشته و آن ها را ببیند) در حال حرکت به سمت ایجاد مجموعه های قابل دسترسی به عنوان ابزارهای تحقیقی، است. برای مثال، زمانی که با دامنهٔ «the UK government. Gov.uk» از 2011 - 2020 مواجه هستیم، یک پژوهش گر چه تصوری می تواند از توانایی انجام آن داشته باشد؟ چه سؤال هایی می تواند از یک مجموعه پرسیده شود؟ برای نمونه، محتوای کامل وب در مورد بانکدارهای اصلی وال استریت و سایت هایی که به طور مستقیم با آن پیوند دارند، در صورتی که آرشیو، دوره ای را پوشش دهد که طی آن بحران های بانکی توسعه یافته اند. به عبارت دیگر، به جای تحلیل یک وبگاه منفرد در سطحی خرد یا تحلیل همه وب در سطح کلان، با مجموعه های متمرکز وب در سطح میانی چه کار می توانیم بکنیم؟ قسمت عمده ای از پژوهش علوم اجتماعی در فضای برخط، در تعاملات سطوح میانی مشاهده می شود. آیا می توانیم برای فهم این که چگونه تغییر وب، واقعیت اجتماعی را منعکس و تقویت می کند یا تغییر می دهد، به صورت یکسان عمل کنیم؟

1- Way Back Machin(http://classic-web.archive.org)/

پاره ای از اقدامات برای پشتیبانی آرشیوهای وب موجود غیر ممکن خواهد بود – ممکن است برای انجام آن داده یا محتوا گردآوری نشده و تقریباً از بین رفته باشد. با این حال، رو به سوی تا آینده چه تغییراتی را می توانیم امروز و در سال های آینده برای آرشیوهای وب، ایجاد کنیم محققان در سالهای 2015 ، 2020 یا 2050 قادر به طراحی منابعی باشند که در حال حاضر برای پاسخ به این سؤال ها گردآوری می کنیم. محققان آینده، چه چیزی را از ما طلب می کنند که حالا در سال 2011، و در آینده، انجام نمی دهیم؟ مؤسسه های خاص چه می توانند انجام دهند؟ اگر IIPC به صورت جمعی در بهره داری از قدرت آرشیوهای چندگانه، عمل نماید چه اتفاق بهتر و مؤثرتری ممکن است رخ دهد؟

وب مجتمع: زندگی آرشیو وب

سؤال: چرا آرشیوهای وب به آرشیو شدن نیاز دارند؟ چرا آن ها نمی توانند با وب پویا یکپارچه گردند و به طور شفاف برای عموم و محققان قابل دسترسی باشند؟ امکان تصور وبی که با وب پویای فعلی بر روی صفحه ای قابل دسترسی به صورت منبعی پیش فرض از داده ها و اطلاعات لایه بندی شده باشد، وجود دارد. با این حال، این سطح می تواند بر روی لایه های زیرین وب قدیمی ایجاد شود که به آسانی برای علاقه مندان با حرکت نزولی یک لایه یا تعداد بیشتری از لایه ها به سمت پایین، قابل دسترسی است. اگر ده ها هزار ابزار قابل دسترسی برای پژوهش وب پویا بتواند برای این لایه های زیرین با استفاده از مکانیسم های ساده به کار گرفته شود احتمال کشف کاربردهایی برای اطلاعات و داده های موجود در وب گذشته توسط محققان افزایش می یابد.

احتمالاً این بزرگترین و بلند پروازانه ترین چالش در این گزارش است، زیرا نیاز به تغییر زیادی در زیر ساخت وب دارد. در عین حال، به معنی احتمال ضعیف وقوع آن است، اما نتایج مورد انتظار آن بسیار خواهد بود فراتر از ارزش پژوهشی که وب قدیمی به عنوان لایه های زیرین وب فعلی، دارد، به صورت بالقوه نیز باعث تغییر ساختاری در نحوه نگرش کاربران به وب خواهد بود. وب فعلی از نظر بسیاری برای اشاعه پیوندهای از دست رفته، از دست دادن اطلاعات، گم شدن صفحه ها، تغییر نشانی ها و تغییر اطلاعاتی که روی ویرایش های قبلی بدون هیچ امکانی برای مشاهده یا برگرداندن به ویرایش های قبلی، بازنویسی می شوند، غیر قابل اطمینان است. اگر ساختار اینترنت در برگشت زمان به عقب، به یکی از لایه های چندگانه تبدیل شود، به نحوی که حفره های موجود در لایه بالایی باعث ایجاد حفره در وب نشود، و به جای آن لایه پایین تر آشکار شود امکان اینکه وب به صورت منبعی با ثبات قابل اعتماد و مقاوم به از دست رفتن اطلاعات باشد، وجود خواهد داشت.

مشکل پیوندهای از دست رفته که تحلیل پیوند (1) نامیده می شود، مشکل دیرپایی برای کاربران وب پویاست. مشکل، زمانی پیچیده می شود که توجه بر وب آرشیوی متمرکز شود که روی هم رفته شناسه– گرهای ثابتی برای ویرایش های آرشیوی صفحه ها وب ندارد. تلاش های متعددی صورت گرفته است،

1- linkrot

به عنوان مثال (1) /(Web Cite (http://www.webcitation.org به نویسندگان امکان آرشیو نسخه ای از صفحه وب و ایجاد پیوند محافظت شده یا تجزیه کننده DOI را می دهد. (2) (Dead URL (http://deadurl.com رهیافت متفاوتی با اتکا بر آرشیو اینترنت و مخزن گوگل، در میان منابع دیگر برگزیده است تا سعی کند نسخه های پیوندهای از دست رفته را بیابد. اگر چه، تلاش های این چنینی بوسیله اکثریت وسیعی از محققان که عمدتاً از وجود آن آگاه نیستند، بدون استفاده باقی می ماند. فراتر از عامل مشکل ساز، نتیجه ناخواسته و غیر منتظره دیگری نیز وجود دارد. عادت محققان به اینکه تا حد امکان نشانی های متعددی را در آثار علمی شان درج کنند. این امر اثرات متعددی دارد:

نخست: زمانی که سعی در ارزیابی تأثیر منابع برخط با استفاده از فنونی چون وب سنجی دارند، فقدان پیوندها باعث کاهش تأثیر منابع شان می شود.

دوم: خوانندگان را وادار به تلاشی جدی تر برای پیگیری منابع اطلاعاتی می کند، به طوری که سعی می کنند نه تنها منبع صحیح استناد، بلکه منبع استناد شده ای را که ممکن است ویرایش صفحه هایش می کنند به طور قابل توجهی تغییر کرده باشد نیز کشف کنند. اگر آرشیوهای وب تبدیل به منبعی قابل اتکا برای استناد پیوسته اطلاعات بشوند، این امر باعث افزایش پژوهش می شود و وجهه (نیمرخ) آرشیوهای وب را به صورت عمومی تر رشد خواهد داد.

در این وب مجتمع، دانش هویدا در وب، رشد و تکامل می یابد، اما به همان روش مشابه وب فعلی، به دور انداخته نمی شود. وب مجتمع، با استفاده از موتورهای جست و جویی نظیر گوگل، قابل خزش، زدودن پیوند شدن و تحلیل و جست وجو خواهد شد. پیوندها از بین نخواهند رفت، بلکه به سوی تبدیل شدن به و موادی هدایت می شوند. که مدت زیادی از وجودشان به صورت فعال بر روی وب فعلی نمی گذرد.

چالش بلند مدت: دو چالش در این سؤال نهفته است، که هر دوی آن ها مستلزم مشارکت کنندگان و فعالان زیادی خواهد بود. نخست: ما باید دوباره در مورد اینکه چگونه اینترنت را ببینیم و مهندسی کنیم فکر کنیم، در حال حرکت از موجودیتی تک لایه با پیوندهای جانبی بسیار به سوی موجودیتی چند لایه با پیوندهای جانبی فعلی، بلکه با پیوندهای فعلی به مواد قدیمی و پیوندهای قدیمی به مواد قدیمی یا فعلی [باشیم]. این چالش کم اهمیتی نیست و متقاعد کردن بسیاری از نقش آفرینان که در ساختن حال آینده اینترنت، سرمایه گذاری کرده اند، برای انتخاب، مشکل خواهد بود. با این حال، نتیجه آن منجر به زیر ساختی خواهد شد که وب گذشته را بسیار قابل دسترس تر برای ارجاع و جست وجو خواهد ساخت. چالش بزرگ بعدی این خواهد بود که آرشیویست های وب نیاز به تعریف مجدد نقش شان خواهند داشت، در حقیقت نه برای آرشیویست بودن به شکل کاملاً سنتی اش، بلکه به صورت متخصصانی که می توانند به پژوهشگران در درک روندها و منابع اینترنت در طول زمان کمک کند و در ابزارهایی که برای دسترسی و دستکاری لایه های این اینترنت چند لایه به آن ها نیاز دارند، برای راهنمایی پژوهشگران و عموم در پاسخ به پرسش های پیش بینی نشده در مورد رشد وب همانگونه که به صورت جدید توسعه یافته اند خبره هستند.

1- Web Site

2- Dead URL

وب در حال تغییر

سؤال: محققان چگونه به تغییر رویـدادها در جهان پاسخ می دهند و یا حوادث جاری را پایش می کنند؟ به طور فزاینده، رویدادهای محلی و جهانی خارج از وب اتفاق می افتد. این حوادث ممکن اس---ت رویدادهای مهم بین المللی و جالب توجه باشـند. ممکن است مانند رویـدادهای سیاسـی اخیر در آفریقـای شـمالی و خاور میانه یا زلزله هایی در هاییتی ژاپن و سایر مکان ها رویـدادهای محلی یا منطقه ای مهمی باشـند. ممکن است کوچک اما گسترده، یا رویدادهای جاری باشند که در اولویت توجه گروهی کوچک یا حتی یک محقق باشند. این امر، به طور بالقوه انواع بینش ها را در مورد ماهیت و چیستی اطلاعاتی که مردم به اشتراک می گذارند، موضوع ها و حوادثی که به طور برجسـته توسعه می یابند، و چگونگی پاسخ افراد و سازمان ها و دولت ها به بحران ها و در طی زمان، چگونگی استحاله و افول حوادث در گفت و گوهای عامه مردم، به بار می آورد. این مورد ساده ترین چالش ولی واضح ترین آن هاست.

علاوه بر این، تعـداد زیادی از محققان در این حوزه با هم کار می کننـد، پدیـده برداشت (1) موضوع چندین گفتگو در نشـست IIPC در سال 2011 بود، توسط سخنرانانی که مثال های زیادی از پدیده برداشت در ارتباط با انقلابهای 2011 از سراسر دنیای عرب تا نشت نفت در آب های عمیق، تا المپیک سال 2012 لندن را بیان می داشتند.

یکی از موضوع هـای محوری کار با آرشـیوهای وب، این است که ما نیاز به حرکت از فهم وب به صورت مجموعه ای از داده های انتخاب شده داریم و به جای آن باید آن را به صورت یک شـبکه در حال تغییر و تحول ببینیم که به دوره های زمانی و رهیافت های طولانی نیاز دارد در این زمینه، تلاـش هـایی انجـام گرفته است. برای نمونه، پروژه اروپایی تحلیل های طولی آرشـیوهای وب (2) در حال ایجاد یک رصـد خانه مجازی وب برای انجام تحلیل های طولی است.

چالش فوری: ایجاد سازوکارهایی برای محققان به منظور پیشنهاد سریع گرانولیته (3) افزایشی و دامنه مناسبی برای استفاده در سایت های آرشـیوی و موضوع های در حال تغییر، در حال حاضر، ممکن است یک پژوهشگر ماهر، ابزارهایی را برای تکرار اجرای خزش، راه اندازی کند، اما پژوهشگری با مهارت کمتر از نظر فنی بدون پشتیبانی قوی سازمانی، یادگیری نزولی خواهد داشت. زمانی که پدیده ای به سرعت متغیر، در حال توسعه است، پژوهشگر متخصص علاقه مند به آن پدیده که هیچ تجربه ای در وب آرشیوی ندارد، باید راهی برای گردآوری داده برای تحلیل آن، قبل از دست رفتن داشته باشد. سازمان های ماهر در راه اندازی ابزارهایی برای گردآوری داده های موجود بر روی این موقعیت های توسـعه یافته، در سـطح مناسبی از گرانولیته، می تواننـد راه هایی برای محققان یا دیگران فراهم کنند که وبگاه ها، موضوع ها، کلید واژه ها، مانند آن را برای پاسخگویی سریع به رویدادهای وب برگزینند.

1- Harvesting

2- the European Longitudinal Analytics of Web Archives project(http://www.lawa-roject.eu/)

3 – granularity ([گرانولیته ، دانه دانه بودن] اصـطلاحی در مورد انـدازه واحـد بـا مفهوم ،اسـتاندارد برای یـک حـالت عملیـاتی خاص. اصـطلاح مزبور در عناوینی چون وضوح صـفحه نمایش معیار جست وجو و مرتب سازی پایگاه داده ها سـطوح پردازش داده ها و مقـدار زمانی که ریز پردازنده به یک عملیات اختصاص می دهد به کار می رود) (منبع: اصطلاحنامه تخصصی مایکروسافت).

عکس

تصویر۱۹. روند موضوعات توییتر ، ۲۹ آوریل ۲۰۱۱

چالش در حال توسعه: اســـتفاده از ابزارهایی نظیر خوراک‌های .R.S.S برای به‌کارگیری نشانه ها، بیانگراین است که تغییرات صفحه‌های وب نیاز به آرشیو شدن دارند.

در ارتباط با توسعه رویدادها، امکان پایش مواردی نظیر خوراک‌های .R.S.S یا نرم‌افزارهای کاربردی اخیراً توسعه یافته، برای داشتن سیستم‌هایی که به فعالیت تکثیر به وسیله افزایش تکرار ذخیره صفحه وب، پاسخ داده یا بوسیله آگاه‌سازی متصدیان انسانی حوزه های درحال توسعه بالقوه مورد علاقه، وجود دارد.

چالش طولانی مدت: ساخت الگوریتم‌هایی که از روندهای فعالیت برخط (نظیر روندهای گوگل، یا روند موضوع‌های توییتر) به‌منظور راه‌اندازی گرانولیته آرشیوسازی افزایشی برای صفحه‌های وب مرتبط با آن موضوع‌های استفاده می‌کنند.

این امر مستلزم مهارت و پختگی بیشتری برای آرشیوهایی است که به‌طور مناسبی برای استفاده مجدد و به اشـــتراک گذاری و استانداردسازی ایجاد شده و استمرار دارند. پژوهشگران علاقه‌مند به استفاده از چنین آرشیوهای گردآوری شده الگوریتم گونه‌ای (دارای الگوریتمی) ، می‌خواهند منطق حذف یا افزودن را بدانند و قادر به فهم ماهیت و محتوای مجموعه باشند. سؤال محوری که پرسیده خواهد شد این است که چگونه مجموعه‌ها با اکوسیستمی از منابع پژوهشی قابل استفاده و قابل دسترسی، متناسب شوند؟

کاربردهای آرشیوها و وبگاه‌ها

سؤال: چگونـــه افراد آرشیوهای وب را بسیار مهم‌تر از وبگاه‌ها به‌کار می‌گیرند؟ در حال حاضر، احتمال زیادی وجود دارد که وبگاه‌های بر روی وب را در مقاطع مشـــخص زمانی، با اســـتفاده از آرشـــیوهای وب و زیر ســـاخت آرشیوی وب، مشاهده نمود. برای پژوهشـــگران دانشگاهی و صنعت تحلیل‌های

تصویر 19 روند موضوعات توییتر ، 29 آوریل 2011

چالش در حال توسـعه: اسـتفاده از ابزار هایی نظیر خوراک های R.S.S . برای به کـارگیری نشانه ها بیان گر این است که تغییرات صـفحه های وب نیاز به آرشیو شدن دارند.

در ارتبـاط بـا توسـعه رویـدادها، امکان پایش مواردی نظیر خوراک های .R.S.S یـا نرم افزارهای کاربردی اخیراً توسـعه یافته برای داشـتن سیسـتم هایی که به فعالیت تکثیر به وسیله افزایش تکرار ذخیره صفحه ،وب پاسخ داده یا بوسیله آگاه سازی متصدیان انسانی حوزه های در حال توسعه بالقوه مورد علاقه، وجود دارد.

چالش طولانی مـدت: ساخت الگوریتم هایی که از رونـدهای فعالیت برخط (نظیر رونـدهای گوگل یا روند موضوع های توییتر) به منظور راه اندازی گرانولیته آرشیو سازی افزایشی برای صفحه های وب مرتبط با آن موضوع های استفاده می کنند.

این امر مسـتلزم مهارت و پختگی بیش تری برای آرشـیو هایی است که به طور مناسبی برای استفاده مجدد و به اشتراک گذاری و استاندارد سازی ایجاد شده و استمرار دارند پژوهش گران علاقه مند به استفاده از چنین آرشیوهای گردآوری شده الگوریتم گونه ای دارای الگوریتمی ، می خواهنـد منطق حـذف یا افزودن را بدانند و قادر به فهم ماهیت و محتوای مجموعه باشـند. سؤال محوری که پرسـیده خواهد شد این است که چگونه مجموعه ها با اکوسیستمی از منابع پژوهشی قابل استفاده و قابل دسترسی متناسب شوند؟

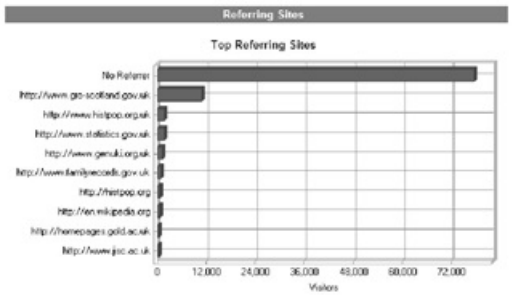
کاربرد های آرشیو ها و وب گاه ها

سؤال چگونه افراد آرشـیوهای وب را بسـیار مهم تر از وب گاه ها به کار می گیرنـد؟ در حال حاضـر، احتمال زیادی وجود دارد که وب گاه های بر روی وب را در مقاطع مشـخص ،زمانی با اسـتفاده از آرشیو های وب و زیر ساخت آرشیوی ،وب مشاهده نمود برای پژوهش گران دانشگاهی و صنعت تحلیل های

سرویس دهنـده ثبت وقـایع و تحلیـل های تحلیل گران تکنیکی پر تکرار برای ارزیابی کاربردها تأثیر و الگوهای ترافیک در وب پویاست با وجود ،این این فنون برای وب آرشـیوی امکان پذیر نیست زیرا داده ها برای درک اسـتفاده های وب قدیمی و آرشـیو های وب به سادگی در دسترس نیستند.

عکس

سـرویس‌دهنده ثبت وقایع و تحلیل‌های تحلیل‌گران، تکنیکی پر تکرار برای ارزیابی کاربردها، تأثیر و الگوهای ترافیک در وب پویاسـت. با وجود این، این فنون برای وب آرشـیوی امکان‌پذیر نیست. زیرا داده‌ها برای درک استفاده‌های وب قدیمی و آرشیوهای وب به سادگی در دسترس نیستند.



تصویر. ۲۰نمونه ای از فایل داده ثبت وقایع برای سایت histpop. org منبع: مه یر و دیگران، ۲۰۰۹

چالش فوری: آرشـیو سـرویس دهنده ثبت وقایع سایت‌های وب آرشـیوی، که محققان بتوانند نحوه اسـتفاده از آرشیوهای وب را مطالعه نمایند. این راه حلی سـاده و سرراست و در دسترس برای مؤسسه‌هایی است که آرشیوهای وب را می سازند. سرویس دهنده ثبت وقایع آرشیوهای وب می تواند برای پژوهشگران علاقمند به فهم چگونگی مرور آرشیوهای وب توسط کاربران و چگونگی دسترسی آنها به منابع و اینکه چه بخش‌هایی از آرشـیو بیشترین تکرار استفاده را دارند، ذخیره، نگهداری و قابل دسترسی شود. این اطلاعات به‌طور عمده مورد توجه جامعه آرشیوی وب خواهد بود. اما این گام نخست است.

چالش بلند مدت: تلاش بلند پروازانه تری اسـت، اما علاقه بالقوهٔ بسیار گسترده‌تری خواهد بود: راه‌اندازی زیر سـاختی برای امکان آرشیوسازی سـرویس‌دهنده‌های ثبت وقایع و تحلیل‌های مرتبط و پیوند با وبگاه‌ها، به‌طوری که پژوهشگران نه فقط آنچه که بر روی وب موجودبوده، بلکه نحوه استفاده از آن را نیز بتوانند ببینند. این هدفی بسـیار بلند پروازانه‌تر است، زیرا سرویس دهنده‌های ثبت وقایع و اعتبارهای تحلیل‌ها فقط به‌صورت داخلی برای سـرور و مدیران اعتبار در وضعیتی حفاظت شده قابل مشـاهده هستند. تجربیات عمومی مدیران سـرور برای ذخیره سرویس‌دهنده ثبت وقایع در بلند مدت ضروری نیستند، به‌طوری که آنها به‌طور عادی برای ذخیره سازی فضا و دوری از انباشتگی و آشفتگی سرور، حذف و بازنویسی می‌شوند. اگر چه این داده ها، به‌طور بالقوه برای پژوهشگرانی ارزشمند است که نه فقط می‌خواهنداز وضعیت یک سـایت آگاه شوند ، بلکه می‌خواهند از چگونگی و مقدار استفاده ازآن، ترافیک منابع و دیگر حقایقی که از ثبت وقایع و داده‌های تحلیلی می‌تواند گردآوری شود، آگاهی یابند. راه‌حل‌های اجتماعی شـامل ایجاد مکانیسم‌هایی برای مدیران سرور برای مشارکت در ثبت وقایع

تصویر. 20 نمونه ای از فایل داده ثبت وقایع برای سایت histpop. org منبع: مه یر و دیگران، 2009

چالش فوری: آرشیو سرویس دهنده ثبت وقایع سایت های وب آرشیوی، که محققان بتوانند نحوه استفاده از آرشیوهای وب را مطالعه نمایند این راه حلی ساده و سر راست و در دسترس برای مؤسسه هایی است که آرشیو های وب را می سازند. سرویس دهنده ثبت وقایع آرشیوهای وب می تواند رای پژوهش گران علاقمند به فهم چگونگی مرور آرشیوهای وب توسط کاربران و چگونگی دسترسی آن ها به منابع و اینکه چه بخش هایی از آرشیو بیش ترین تکرار استفاده را دارند، ذخیره، نگهداری و قابل دسترسی شود. این اطلاعات به طور عمده مورد توجه جامعه آرشیوی وب خواهد بود. اما این گام نخست است.

چالش بلند مدت: تلاش بلند پروازانه تری ،است اما علاقه بالقوه بسیار گسترده تری خواهد بود راه اندازی زیر ساختی برای امکان آرشیو سازی سرویس دهنده های ثبت وقایع و تحلیل های مرتبط و پیوند با وب گاه ها به طوری که پژوهش گران نه فقط آن چه که بر روی وب موجود بوده، بلکه نحوه استفاده از آن را نیز بتوانند ببینند این هدفی بسیار بلند پروازانه تر است، زیرا سرویس دهنده های ثبت وقایع و اعتبارهای تحلیل ها فقط به صورت داخلی برای سرور و مدیران اعتبار در وضعیتی حفاظت شده قابل مشاهده هستند. تجربیات عمومی مدیران سرور برای ذخیره سرویس دهنده ثبت وقایع در بلند مدت ضروری نیستند به طوری که آن ها به طور عادی برای ذخیره سازی فضا و دوری از انباشتگی و آشفتگی سرور، حذف و بازنویسی می شوند. اگر چه این داده ها به طور بالقوه برای پژوهش گرانی ارزشمند است که نه فقط می خواهند از وضعیت یک سایت آگاه شوند، بلکه می خواهند از چگونگی و مقدار استفاده از آن، ترافیک منابع و دیگر حقایقی که از ثبت وقایع و داده های تحلیلی می تواند گردآوری شود، آگاهی یابند. راه حل های اجتماعی شامل ایجاد مکانیسم هایی برای مدیران سرور برای مشارکت در ثبت وقایع

است که می تواند با وبگاه های آرشیو ارتباط داشته باشد و برای ارائه کنندگان تحلیل هایی همچون گوگل به منظور ارائه گزینه ای به منظور مشارکت سایت های تحلیلی با آرشیوهای وب و تا حد امکان تعیین دوره محدودیت قبل از این که داده ها انتشار یابند.

چالش بلند پروازانه طراحی نظام هایی برای آرشیو نه تنها سرویس دهنده های ثبت وقایع، بلکه خود ترافیک وب در یک سبک ناشناخته حفاظت شده این راه حل حتی بسیار بلند پروازانه تر است؛ زیرا سازو کارهای تحلیل ترافیک ،وب روی هم رفته عموماً در دسترس نیستند. نگرانی های نهفته ای از مواردی نظیر بازرسی بسته های عمیق که به تحلیلگران دربارهٔ ترافیک روان ،وب می گوید وجود دارد.

پرسش محوری برای سؤال کردن این است که چه زمان مزایای فهم نحوه رفتار افراد در وب بر خطرات غلبه مینماید ،بنابراین توجه به این امر که روش هایی وجود دارند که این داده ها می توانند برای تحلیل های بعدی آرشیو و ذخیره شوند در زمانی که خطرات برای افراد و سازمان ها به حد کافی به واسطه گذر زمان و گمنام کردن داده ها کاهش می یابد ارزشمند است.

متخصص وب

سؤال: آیا امکان شناسایی مجموعه هایی که اندازه و شکل وب تاریخی را در رابطه با حوزه های تخصصی مورد توجه می سنجند، وجود دارد؟ فرض می کنیم که بسیاری از گروه ها و تنالگان ها وب گاه هایی را در طول دوره های زمانی ایجاد خواهند کرد چگونه امکان دارد که شناسایی تداوم حضور آن ها در وب و جمع آوری بدنه مناسبی از سایت های پژوهشی، برای آن ها کوچک و بی اهمیت باشد؟

مثال های متعددی را می توان بر شمرد گروه های علاقه مند به سرگرمی موضوع های تخصصی دانشگاهی میراث صنایع دستی نظیر دیداری و شنیداری سایت های گروه های سیاسی، و مانند آن. مطالعه این موارد می تواند موجب ساخت ابر مجموعه ها شود مجموعه مجموعه ها، آرشیو – در – جعبه چه راهنمایی در زمان ساختن مجموعه های آرشیوی از منابع متنوع وب برای افزایش ارزش این مجموعه ها مورد نیاز است؟ یک ابر مجموعه تا چه اندازه می تواند با دیگری مقایسه و پیوند داده شود و حتی به بخش هایی از ابر مجموعه های بزرگ تر تبدیل شوند؟

متخصص وب، ضرورت مقیاس کوچک تری از داده های انتخاب شده بر اساس رویداد ها یا موضوع هایی هم چنان ،انتقادی و موازی با تجربیات و انتظارات موجود پژوهشی را تشخیص می دهد. مجموعه قوانینی از این دست به صورتی قابل مشاهده و جست و جو بوسیله افراد یا تیمی باقی میماند که این رشته های داده را از یک چشم انداز ویژه رشته یا موضوع پژوهشی مهمی در دنیای آکادمیک، ایجاد و تحلیل می کنند به ویژه در زمینه هایی که باوری دیرینه وجود دارد که پژوهش گران، خالقان مدیران و تحلیل گران مجموعه قوانین خود شان باشند با این حال حتی این مجموعه های تخصصی به طور بالقوه در زمانی که به طور استاندارد برای پاسخ به سؤال های محققان دوباره ترکیب و دسته بندی می شوند، می توانند ارزش آفرین باشند

این ،امر مسئله مقیاس پذیری را در مسیر پرسش های مرتبط با آرشیو های وب متخصص پرورش

می دهـد. آیا انـدازه بحرانی در طول دوره های پوشـش و دامنه مجموعه وب آرشـیوی برای اینکه در طول زمان مورد اسـتفاده دیگر پژوهش گران باشـد وجود دارد؟ چگونه راهبرد هـا و سـیاست هـای خزنـدگی وب مؤسـسه هـای مختلف (همچون برداشت هـای عمـده و حوزه رویـدادی یـا انتخابی) انتظارات پژوهش گران را برآورده کرده، و چه راهبرد های متنوعی به مناسب ترین وضع برای پشتیبانی، اسـتفاده می شوند؟

بعضی از مثال های ویژه حوزه مجموعه های تخصصی در ابزارهای علوم اجتماعی برای ترکیب بانک جهانی داده با تحلیل های موجود در ولف رام آلفا (1) در ارتباط با سـلامتی یا دیگر اطلاعات جمعیتی می توانند توسعه یابند. این مستلزم استفاده از ابزاری (همچون ولف رام آلفا) دارد که به طور مداوم با داده های جدید روزآمد می شود به همان صورت الگوریتم ها و ابزار های مصور سازی آن، بستر تحلیلی داده های پویا را ارائه می کننـد در علوم ،طبیعی پژوهش گران به گرفتن اطلاعات آب و هوا (دما) علاقه منـد خواهنـد بود و این اطلاعات جمع آوری شده را با طبیعت شـناسان غیر حرفه ای (مانند مشاهده کنندگان پرندگان) در گروه های مختلف (با اسـتفاده از ابزارهایی چون گوگل پلاس یا فیس بوک) در سراسـر دنیا، به اشتراک گذاشـته و به گروه هایی می پیوندنـد که چگونگی تغییر الگو های مهاجرت پرندگان با آب و هوا یا آن چه مهاجرت پرندگان درباره تغییرات آب و هوا می تواند بگوید را تحلیل می نمایند. این امر نیاز به آشـنایی با علم ،شهری جامعه ای برخـط و محیـط گرایی دارد در علوم انسـانی به عنوان عقایـد تاریخی ممکن است مصاحبه های آلن مک فارلن (2) با متفکران معاصر اصلی که در سخنرانی های iTunesu (3) آکسفورد نمایش داده می شود مقایسه شود تا الگو های چگونگی تفکر اجتماعی متفکران مهم در مقابل آن هایی که در iTunesu مشهور هستند شفاف شود. (به عبارت دیگر مقایسه افکار اصلی در مقابل با افکار مردمی).

چالش فوری: ارائه راهنمایی برای انواع متفاوتی از ابر مجموعه ها به طوری که عناصر استاندارد در آرشیو وب وارد شوند و در نتیجه دامنه و انسجام آن ها تضمین و تعیین شود.

چالش توسـعه یافته: ارائه ابزارها و زیر ساخت های سازمانی برای این که پژوهش گران بتواننـد بر تردیـدها غلبه یابند تا ابر مجموعه هایی را بسازنـد که حـداقل به صورت بـالقوه توسط یک پژوهش گر منفرد مورد اسـتفاده قرار گیرد. اسـتانداردهایی تعریف شودتا ابر مجموعه های آرشیوی توسط یکدیگر قابل استفاده باشند.

چالش بلند مدت: تشویق به پدید آمدن سازمان هایی که ابر مجموعه ها را به طور مفید و گسترده تشویق و حمایت می کنند.

وب دیداری

سؤال: اگر بخواهم از تصاویر برای فهم چگونگی تغییرات دنیا اسـتفاده کنم آیا می توانم برای فهم این فرآیندها به صورت دیداری تصاویر را از آرشیوهای وب استخراج کنم؟ برای مثال: آیا این امکان وجود دارد که با استخراج تصاویر تغییر یافته در طول زمان از مکان های یکسان بر روی صفحه هایی نظیر

1- Wolfram Alpha

2- Alan MacFarlane

3- iTunes U

(1 Flicker) تحلیل های تصویری انجام گیرد؟ تصویر برداری مجدد، تجربه ای است که طی آن مکانی که قبلاً از آن عکس برداری شده بازدیـد کرده و برای مسـتند کردن تداوم و تغییرات در طول زمان، عکس جدیدی از آن گرفته می شود. یکی از نخستین پروژه ها برای انجام این کار بازدید مجدد از 1200 سایت در آمریکای غربی و تصویر برداری از صد سال قبل به وسیله پژوهشگران دولتی بوده است. (کلت، (2) منچستر (3) و وربورگ 1984. (4)).

حالا، تصور کنید 100 سـال بعـد از حـال را که قادر باشـیم نه تنها عکس های یکسانی را از مکانی مشـخص مقایسه کنیم بلکه بتوانیم از آرشیو های وب سری کامل عکس های مستند شده ای را از دنیای در حال تغییر و ثابت در طول زمان استخراج کنیم.

عکس

Flicker[1] ، تحلیل های تصویری انجام گیرد؟ تصویر برداری مجدد، تجربه‌ای است که طی آن مکانی که قبلاً از آن عکس برداری شـــده، بازدید کرده و برای مستند کردن تداوم و تغییرات در طول زمان، عکس جدیدی از آن گرفته می‌شود . یکی از نخستین پروژه‌ها برای انجام این کار، بازدید مجدد از ۱۲۰۰ سایت در آمریکای غربی و تصویربرداری از صد سال قبل به‌وسیله پژوهشگران دولتی بوده است.(کلت[2]،منچستر[3]و وربورگ[4] ، ۱۹۸٤)

حالا، تصور کنید ۱۰۰ ســـال بعد از حال را که قادر باشـــیم نه تنها عکس‌های یکسانی را از مکانی مشخص مقایسه کنیم، بلکه بتوانیم از آرشیوهای وب سری کامل عکس‌های مستند شده‌ای را از دنیای در حال تغییر و ثابت در طول زمان استخراج کنیم.

تصویر ۲۱. تصویر قصر باکینگهام که از ۲۲۰ تصویر مختلف گردآوری شده‌است.
منبع:http://photosynth.net/view.aspx?cid=34e49d3e-2d1e-4118- bbad-d2f5d74ce340

چالش فوری: تضمین اینکه، تصاویری که بیشتردر صفحه‌های وب آرشیو شده، از دست می‌روند، در اولویت حفاظت قرار گیرند.

چالش توسعه یافته: ساخت فناوری‌هایی همانند photosynth [5]، که قادر به اتصال تعداد زیادی از تصاویر برای نمایش پانوراما‌ی یک مکان یا شی برای کار با اطلاعات زمانی به‌منظور جمع آوری مناظر مشابه در طول زمان باشند.

چالش بلند مدت: ایجاد یک آرشیو از تصاویر جهان، شامل اطلاعات بسیار از زمان و مکان که تا حد امکان برگرفته از EXIF[6] و داده های صفحه‌های وب باشـــند، به‌طوری که تصاویر بتوانند برای پژوهش استفاده شوند. ابزارهایی برای جای دادن، استخراج، ترکیب و تصاویر مورد نیاز است.

1. Flicker
2. Klett
3. Manchester
4. Verburg
5. http://photosynth.net/Background.aspx.
6. EXIF

تصویر 21. تصویر قصر باکینگهام که از 220 تصویر مختلف گردآوری شده است.

منبع: http://photosynth.net/view.aspx?cid=34e49d3e-2d1e-4118- bbad-d2f5d74ce

چالش فوری: تضمین، این که تصاویری که بیش تر در صفحه های وب آرشیو شده، از دست می روند، در اولویت حفاظت قرار گیرند.

چالش توسـعه یافته ساخت فناوری هایی همانند (5)http://photosynth.net/Background.aspx. که قادر به اتصال تعـداد زیادی از

تصاویر برای نمایش پانوراما‌ی یک مکان یا شی برای کار با اطلاعات زمانی به منظور جمع آوری مناظر مشابه در طول زمان باشند.

چالش بلند مدت: ایجاد یک آرشیو از تصاویر جهان شامل اطلاعات بسیار از زمان و مکان که تا حد امکان برگرفته از EXIF (6) و داده های صفحه های وب باشند به طوری که تصاویر بتوانند برای پژوهش استفاده شوند. ابزارهایی برای جای دادن استخراج ترکیب و تصاویر مورد نیاز است.

1- Flicker

2- Klett

3- Manchester

4- Verburg

5- photosynth

6- EXIF

وب همان گونه که بود

سؤال: چگونه من می توانم وب را همان گونه که بود ببینم؟ اگر من بخواهم وب را همان گونه که، در یکم ژانویه 2011 بود مرور کنم و قادر باشم تا روی صـفحه ها ،تصاویر پیوندها و دیگر محتوای آن همان گونه که در آن روز ظاهر شده بود کلیک کنم چگونه این کار را می توانم انجام دهم؟

عکس

وب همان‌گونه که بود

سؤال: چگونه من می‌توانم وب را همانگونه که بود ببینم؟ اگر من بخواهم وب را همانگونه که، در یکم ژانویه ۲۰۱۱، بود مرور کنم و قادر باشم تا روی صفحه‌ها، تصاویر، پیوندها و دیگر محتوای آن همانگونه که در آن روز ظاهر شده بود، کلیک کنم، چگونه این کار را می‌توانم انجام دهم؟

تصویر۲۲-ویرایش بتای پاسخ WayBack Machine

(منبع http://replay.web.archive.org/20041010185532/http://netpreserve.org/about/index.php)

ویرایش بتای فعلی نسخه پاسخ [پایگاه] Wayback Machin چنین کارکردی را نوید می‌دهد (گشت و گذار در وب همانگونه که بود، نسـخه بتا، که در حال حاضر در سـایت تبلیغ می‌شود)، اما IIPC یا آرشیوهای انفرادی چه تلاش‌های دیگری می‌توانند برای امکان تکرارپذیری وب، انجام دهند؟

چالش: گسـترش و افزایش تلاش‌هایی برای ایجاد وب فعلی به‌صورت تکرارپذیردر آینده که نیاز به گردآوری، ذخیره‌سـازی و اشاعه مجدد وب فعلی همانگونه که در گذشته بود، خواهد داشت. سؤال محوری برای پرسـش در اینجااین است که[وب] چگونه می‌تواند فراتر از این که منبعی ارجاعی صرفاً برای رفع کنجکاوی یا مراجعه گاه به گاه باشـد، مورد اسـتفاده قرار گیرد. چه نوع نیاز بکر یا سـؤال پژوهشـی غیر متصوری بر جست‌وجوهای دستی و گشت و گذار در وب گذشته (قدیمی) اتکا دارد؟ آیا مورخان آینده ، همانگونه که در دنیای امروز خواندن اخبار و انتشـارات و دوران گذشـته فانی وب مورد علاقه بوده‌است ، به آن علاقه‌مند هستند؟ آیا آنها می‌خواهند از نرم‌افزارهای کاربردی استفاده کنند یا صرفاً از منابع مرجع؟ به عبارت دیگر، بزرگ‌ترین سـؤال، ساختن موارد استفاده برای وب تکرارپذیر و سپس سـاختن رابطهایی که این موارد استفاده را پشـتیبانی می‌کند، می‌باشد. انجام این امر مستلزم مشورت متخصصان آرشیو سازی وب با متخصصان حوزه هایی شامل مورخان و دیگر افرادی است که به بازسازی مجدد گذشته علاقه‌مند هستند.

تصویر 22- ویرایش بتای پاسخ WayBack Machine

(منبع http://replay.web.archive.org/20041010185532/http://netpreserve.org/about/index.php)

ویرایش بتای فعلی نسخه پاسخ [پایگاه] Wayback Machin چنین کارکردی را نوید می دهد (گشت و گذار در وب همان گونه که ،بود نسـخه بتا که در حال حاضر در سایت تبلیغ می شود) اما IIPC یا آرشـیو های انفرادی چه تلاش های دیگری می توانند برای امکان تکرار پذیری وب، انجام دهند؟

چالش: گسترش و افزایش تلاش هایی برای ایجاد وب فعلی به صورت تکرار پذیر در آینده که نیاز به گردآوری، ذخیره سازی و اشاعه مجدد وب فعلی همان گونه که در گذشـته بود، خواهد داشت. سؤال محوری برای پرسـش در این جا این است که [وب] چگونه می تواند فراتر از این که منبعی ارجاعی صـرفاً برای رفع کنجکاوی یا مراجعه گاه به گاه باشـد مورد اسـتفاده قرار گیرد چه نوع نیاز بکر یا سؤال پژوهشـی غیر متصوری بر جست و جوهای دستی و گشت و گـذار در وب گذشـته (قـدیمی) اتکا دارد؟ آیا مورخان آینده ، همان گونه که در دنیای امروز خوانـدن اخبار و انتشارات و دوران گذشـته فانی وب مورد علاقه بوده است ، به آن علاقه مند هسـتند؟ آیا آن ها می خواهند از نرم افزارهای کاربردی اسـتفاده یا صـرفاً از منابع مرجع؟ به عبارت دیگر بزرگ ترین سؤال ساختن موارد اسـتفاده برای وب تکرار پـذیر و سـپس ساختن رابطه ای که این موارد اسـتفاده را پشتیبانی می کند می باشد. انجام این امر مسـتلزم مشورت متخصـصان آرشـیو سازی وب با متخصصان حوزه هایی شامل مورخان و دیگر افرادی است که به بازسازی مجدد گذشته علاقه مند هستند

ص: 138

ساختار وب

سؤال: وب چگونه مقایسه شده و در طول زمان تغییر می کند؟

افزایش تلاش هایی برای فهم وب به صورت یک سیستم مستلزم قابلیت های افزایش به مقیاس بزرگ است که قادر خواهد بود الگوها و روندها را در طی زمان دست بی اندازد (تغییر دهد). برای انجام این امر، ما نیاز داریم که بپرسیم چه رهیافت هایی برای توسعهٔ روش های تحلیلی معتبر در دسترس هستند؟ چگونه ما می توانیم فرضیات ساخته شده درباره داده های وب را به صورت یک سری داده اعتبار سنجی کنیم؟ چه ابزارهای آماری برای مجموعه های وب آرشیوی می تواند به کار گرفته شود و چه ابزارهایی نیاز به توسعه دارند؟

حتی آمارهای ساده نیز برای استخراج درباره وب بی اهمیت نیستند برای نمونه، چه تعداد از وب گاه ها به طور سالانه (سراسر جهان در یک کشور مشخص روی موضوع مشخص) برای x تعداد سال گذشته بوده اند؟

در مجموعه آرشیوی (آرشیو در جعبه) تاریخ ایجاد صفحه ها چیست؟ صفحه های آن به چه زبان هایی هستند؟ آیا در زمان ایجاد صفحه ها روندهایی وجود داشته است؟ آیا خوشه ای است؟ آیا یک فرایند ساختمانی ثابت بوده است؟ آیا موضوع های مشخصی بیش تر از دیگران پیوند داده شده اند؟ آیا برخی از انواع مجموعه ها احتمال کم تر یا بیش تری برای پیوند به منابع خارجی دارند؟ آیا وب گاه ها می توانند به رده هایی که می توانیم با استفاده از تحلیل های خوشه ای کشف کنیم تقسیم بندی شوند؟

آیا ما می توانیم سایت ها را به وسیلهٔ آمار هایی همچون اندازه میانگین وب گاه ها در رده های مختلف، میانگین تعداد پیوند ها میزان داده های غیر متنی (عکس ها، تصاویر و مانند آن) سن محتوا در سن محتوا در فاصله بین روز آمدسازی ها دفعات روز آمدسازی، نوع رابط (ثابت در مقابل دینامیک به طور مثال)، مقایسه کنیم؟

چگونه این آمارها به ما در فهم ساختار مجموعه ها و وب کمک می کنند؟

چالش: ایجاد ابزار ها و روش هایی برای استفاده از وب به عنوان یک رشته داده عظیم به جای مجموعه ای از اسناد در حال حاضر در صورت وجود داده ها اگر شخص بخواهد بداند چه چیزی باید به طور بسیار اساسی دربارهٔ اندازه و ساختار وب فعلی یا وب به صورت گذشته اش مورد سؤال باشد داده ها برای پژوهش گران قابل دسترسی نیستند. بنابراین ابزارهایی باید ایجاد شوند که آمارگیری روی وب یا روی صفحه های یک مجموعه آرشیوی را انجام دهند.

ایده ها چگونه تکثیر می شوند

سؤال: ایده ها چگونه روی اینترنت کشش یافته و تکثیر می شوند؟ یکی از جنبه های قابل توجه اینترنت توانایی شگفت انگیز آن برای پشتیبانی انتقال از الگوی رفتاری ایده هایی که رشد می یابند و گسترش فرهنگی است اگر علاقه مند هستیم به چگونگی ویروسی شدن ویدئو ها یا چگونگی گسترش شوخی ها یا چگونگی وارد شدن یک بیت از اطلاعات یا اطلاعات غلط به آگاهی عمومی، چه ابزارهایی به ما کمک خواهد نمود؟ چگونه ابزارهایی را با توانایی امکان این که یک آرشیو ساخته شود نه بر اساس

جغرافیایی فیزیکی یا مجازی بلکه بر اساس حرکت از یک ایـده می سازیم؟ هر کس می توانـد تصور کند که قادر به تعیین یک ایده باشد و برای دنبال نمودن آن ایده به صورتی که در طول زمان توسعه می یابد، حرکت کند

هم چنین چه زمینه گسترده ای در اطراف محتوایی که ما در آرشیو می بینیم، وجود دارد؟ برای مثال: مردم روی وب در زمان پیدایش آن چه چیزی را جست و جو می کردنـد؟ گوگل زیتگیست (1) و گوگل ترندز (2) در مورد چیزهایی که مردم درباره آن جست وجو می کردنـد به مـا می گوینـد. چه چیزهایی دیگری را می توانیم برای فهم زمینه جمع آوری کنیم؟ برای مثال- در تویتر اشاره شـده است - به فهم زمینه محتوا بوسیله مشاهدهٔ چیزهایی که با یکدیگر آمده اند ، کمک می کند.

برای نمـونه واتسون آی بی ام (3)، نظام اختصاصـی است که از مواد آرشـیوی زیادی برای ساخت موتور دیپ کیو ای (4) خودش استفاده می کنـد که به آن برای برنده شدن در بحران 2011 کمک می کند. چگونه این نوع از ابزار به طور گسترده تری برای جست وجوی پیشـرفته در دسترس خواهند شد؟

چالش: بعـد زمان وب همان گونه که ایجاد شده نیاز دارد که حفاظت شده قابل اسـتخراج و قابل تحلیل .باشد گرانولیته بهتری نیاز است تا ببینـد ایده ها از کجا آغاز شده، چگونه گسترش می یابند و این که چه فعالیت هایی سـرعت تکثیر آن ها را کاهش یا افزایش می دهد؟ ایده ها به صورت موجودی زنده در دنیا پدید می آیند و تنها پس از این که نگهداری ،شوند برای برگشت به اصلیت شان مورد توجه قرار خواهند گرفت با این حال بدون گرانولیته و عمق مناسب آرشیوی ایده اصلی ممکن است در زمانی که شخص به جست و جوی آن فکر می کند، از دست رفته باشد.

وب غیر قانونی

سؤال: چگونه وب برای پشتیبانی و توانایی فعالیت های غیر قانونی اسـتفاده می شود و چگونه در طول زمان تغییر می کنـد؟ نوعی محتوا که به صورت برخط تکثیر می شود و کم تر مورد توجه دانشـمندان است، مواد غیر قانونی وب است این طیف گستره ای از محتوای جنسـی تا اطلاعـاتی دربـاره اسـتفاده از مواد ،مخـدر قمـار منابع گروه های تنـدرو محتوای مرتبط با تروریسم و دیگر منابعی است که یا غیرقانونی یا دارای مشـکلات اجتماعی هسـتند سؤال این جاست که چه کسـی باید محتوا های غیر قانونی یا قانونی ولی کم تر از نظر اجتماعی پذیرفته شده وب را آرشیو کند؟ چگونه بدون شکستن قانون می توانیم این کار را انجام دهیم؟ و چگونه می توان برای محققان بدون در خطر افتادن هم پژوهش گر و هم مؤسسه هایی که دسترسی را ایجاد ساخته اند قابل دسترسی شود؟

دانسـتن این که چه فعالیت های غیر قانونی ای از نظر حجم و عمومیت یافتن در حال رشد هستند کدام یک در طول زمان رنگ می بازند و چه فعالیت های غیرقانونی غیر منتظره ای پدید می آیند نه فقط

1- Google Zeitgeist (http://www.google.com/press/zeitgeist 2010/)

2- Google Trends (http://www.google.com/trends)

3- IBM›s Watson

4- Deep QA (http://www.research.ibm.com/deepqa/deepqa.shtml)

برای پژوهش گران، بلکه برای سیاست گذاران عمومی، متخصصان حوزه سلامت که احتیاج به پیگیری نتایج رفتارهای خطر آفرین دارند خبرگان سلامت عمومی، متخصصان و مؤسسه های حمایت اجتماعی و کسانی که مسئول حفظ رفاه قشر آسیب پذیر هستند مفید می باشد

چالش: بزرگ ترین چالش در این جا این است که حتی اگر منابع غیر قانونی بر روی اینترنت رواج داشته باشند سازمان های اندکی هستند که خارج از اعمال فشار قانونی مایل به پذیرش خطر مشارکت در جمع آوری داده های مربوط به این منابع .هستند تابوهای فرهنگی و خطرات قانونی در ارتباط با دسترسی و ذخیره سازی منابع غیر قانونی حتی برای اهداف مثبت همانند پژوهش فهم جنبه های جامعه مدرن بیشتر پژوهش گران وب و آرشیویست های وب را از چنین منابعی بر حذر می دارد.

به نظر می رسد مهم ترین سازوکاری که می تواند در این جا وجود داشته باشد نظامی با حفاظت قانونی خواهد بود که به طور مناسب و تا حد امکان افراد و سازمان ها را برای آرشیو و جست وجوی داده های غیرقانونی قابل دسترسی اینترنت بدون در معرض خطر قرار گرفتن سازمان ها یا پژوهش گران استفاده کننده از این مجموعه ها تأیید کند

رد پای رقومی

سؤال: چگونه می توانیم (و باید) رد پای رقومی یک شخص را آرشیو کنیم؟ فعالیت ها و اقدامات یک شخص به طور پیوسته مورد توجه بالقوه است به ویژه اگر شخص (مشهور باشد یا بشود) قبلاً در این مورد بحث شده است که (گارفینکل (1) و کاکس 2009 (2) فهرست آثار زندگی یک شخص برای آرشیویست ها یک وظیفه خواهد بود آرشیو وب یک شخص می تواند شامل صفحه های وب پروفایل های شبکه اجتماعی و پست هایش، ارتباطاتش، انتشارات، و منابع دیگر در مورد زندگی رقومی او باشد.

چالش: چالش محوری، فهمیدن چگونگی ساختن ابزارهایی است که به افراد اجازه می دهد تا به صورت دستی مشخص کنند که چگونه به صورت خودکار آثار رقومی خود را جمع آوری کنند. آیا ابزارها به طور خودکار ردپای رقومی را تا حد امکان بر مبنای انتخاب ، جمع آوری می کنند؟ چگونه می توانیم نظام هایی را ایجاد کنیم که نه فقط یادآوری بلکه امکان فراموشی به وسیله حذف های بعدی توسط افراد (و فراموشی) بخش ها یا همه رد پاها را داشته باشد به صورتی که برخی دانشمندان از آن به عنوان یک حق اساسی ذکر کرده اند مه یر (3) شونبرگر 2009 (4)؟ پیشرفت ها به ویژه در این حوزه مشکل است به دلیل این که مسائل بسیار زیاد حقوقی و خصوصی آن نیاز به بررسی خواهد داشت.

1- Garfinkel

2- Cox

3- Meyer

4- Schonberger

وب داده

سؤال: چگونه داده ها از آرشیو های وب دوباره استخراج می شوند؟ در سال های اخیر، رشد چشم گیری در وب مبتنی بر داده (در مقابل وب اسنادی) صورت گرفته است مسائل متعددی مطرح شده است، مانند این که چگونه داده ها در کنار اسناد آرشیو بشوند؟ چه نوع ابزار پاک کننده داده ها برای کار با داده ها مورد نیاز خواهد بود؟

برای مثال به فرآیندهای بین حوزه های علمی یا صنعتی برای فراهم آوری دانش مانند طراحی ،هوافضا، کشف مواد مخدر و غیره، فکر کنید زمانی که یک هواپیما دچار سانحه می شود و بازرسان می خواهند محاسبات اولیه مهندسی را دوباره ارزیابی کنند چگونه می توانیم توانایی درک داده ها از یک زنجیره تأمین طراحی مهندسی هفتاد ساله را حفظ کنیم؟ طراحی رقومی ،بود دانش توسط 100 نفر همکار تدارک می شد که بعضی از آن ها در فواصل زمانی از حوزه کار خارج شده اند و همه آن توسط مجموعه کاملاً متفاوتی از افراد به کار گرفته می شد در این مورد چرخه زندگی دانش بسیار طولانی تر از چرخه زندگی کسب و کار است و آرشیوها نقش اساسی در حفظ این اطلاعات ایفا می نمایند.

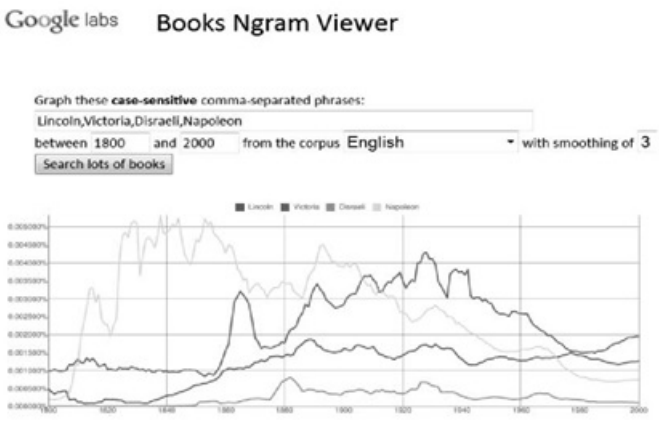
سؤال های مرتبط این است که چگونه می توانیم اطلاعات اختصاصی را آرشیو و تحلیل کنیم؟ بعضی از داده ها اختصاصی هستند و لازم است برای تعهد به موافقت نامه های حقوقی نگهداری شوند. با وجود ،این اگر نهادهای قابل ،اعتماد داده های خام را ذخیره کنند و فقط به ابزارهای تحلیلی اجازه دسترسی به آن ها را بدهند تحلیل های بسیاری بدون دسترسی به داده های خام می توانند انجام گیرند. سپس نتایج گردآوری و تلخیص شده می توانند بدون افشای داده های حفاظت شده در دسترس پژوهش گر قرار گیرند. ابزارهایی برای سرویس دهی به عنوان الگویی که امکان دسترسی به بخش هایی از داده های خام را می دهند به جای اینکه داده های خام را بارگذاری کنید وجود دارند همانند الیکسر، (1) برنامه ای برای دسترسی به اطلاعات نجوم مثال ،دیگر مشاهده گر n-gram کتاب های گوگل (2) که به کاربران امکان تحلیل داده بدون دسترسی به داده خام را می دهد.

اگر یکبار به داده هایی از مخازن داده های قابل ،تحلیل دسترسی داشته باشید، امکانات افزوده فزونی خواهند یافت. همانند پیوند داده ها از طریق ،ابزارها ایجاد کتابخانه هایی متشکل از اجزایی که به پژوهش گران اجازه می دهند که داده ها را از راه های گوناگون تحلیل کنند و امکاناتی برای ترکیب و تطابق آن ها را به روش های جدید فراهم می کنند. اگر در ایجاد این سری داده ها و ابزارها، دقت شود، الگو ها و همبستگی هایی که قبلا کشف نشده اند، افزایش می یابد.

1- Elixer

2- Google Books Ngram Viewer

عکس



تصویر ۲۳- مشاهدهگر n-gram کتاب‌های گوگل

چالش: برای بخش‌هایی از اینترنت که حاوی داده‌ها به جای اسناد هستند، تغییر نگرش از اینکه آرشیوهای وب برای نگهداری اسناد هستند به سمت مشاهده آنها به‌صورت روشی برای آرشیوسازی داده‌هایی است که این اسناد در بردارند. این امر مستلزم مدل‌های جدیدی برای ذخیره‌سازی و استخراج داده‌هایی است که از مدل‌های متمایل به تحلیل‌های داده‌های داده‌های ساختار یافته، به جای تحلیل‌های داده‌های ساختار نیافته، پیروی می‌کنند.

وب‌های ملی: چه ارزشی در ایجاد یک وب ملی زمانی‌که وب پدیده‌ای فراتر از مرزهاست، وجود دارد؟ تلاش‌هایی در حوزه آرشیو- در جعبه- برای اجرا در سطوح ملی، با توجه به محدودیت-های بودجه‌ای و قانونی، اساسی است. در حال حاضر ، در بریتانیا ، کتابخانه بریتانیا، خودش را برای پیش بینی قوانینی تأثیرگذار بر روی آرشیوسازی فضای وب بریتانیا به‌صورت کتابخانه واسپاری برای همه انتشارات بریتانیا آماده می‌کند. در می ۲۰۱۱، وزارت پژوهش و اختراع دانمارک، تصمیم به ایجاد زیرساخت‌های پژوهشی ملی در حوزه‌های پژوهشی مختلف(علوم طبیعی، علوم انسانی) که متمرکز بر استفاده از ابزارهای تحلیلی منابع وب آرشیوی خواهد بود.

چالش فوری: عدم شفافیت نحوه استفاده پژوهشگران از آرشیوهای ملی. بسیاری از آنها هنوز در مراحل طراحی هستند. ما در مورد یکی از مهم‌ترین مواردی که باید انجام شود، یعنی تعامل دامنه پژوهشگران با تخصص‌هایی نه فقط در پژوهش اینترنت، بلکه در زمینه‌هایی نظیر جامعه‌شناسی، علوم سیاسی، و دیگر حوزه‌های علوم اجتماعی، فیزیک و سایر علوم، هنر و علوم انسانی، بحث خواهیم کرد. همان‌طورکه این زیرساخت‌ها به‌منظور انعکاس نیازهای پژوهشگران ملی در مجموعه‌های ایجادشده، طراحی می‌شوند. این فرآیندی زمان‌بر است و تعامل خبرگان هر حوزه می‌تواند مشکل باشد. با وجود

تصویر 23- مشاهده گر n-gram کتاب های گوگل

چالش: برای بخش هایی از اینترنت که حاوی داده ها به جای اسناد هستند، تغییر نگرش از این که آرشیو های وب برای نگهداری اسناد هستند به سمت مشاهده آن ها به صورت روشی برای آرشیوسازی داده هایی است که این اسناد در بردارند. این امر مستلزم مدل های جدیدی برای ذخیره سازی و استخراج داده هایی است که از مدل های متمایل به تحلیل های دادهای داده های ساختار ،یافته به جای

تحلیل های داده های ساختار نیافته پیروی می کنند

وب های ملی: چه ارزشی در ایجاد یک وب ملی زمانی که وب پدیده ای فراتر از مرزهاست وجود دارد؟ تلاش هایی در حوزه آرشیو – در جعبه–برای اجرا در سـطوح ،ملی با توجه به محدودیت های بودجه ای و ،قانونی اساسی است در حال حاضر، در بریتانیا ، کتابخانه بریتانیا خودش را برای پیش بینی قوانینی تأثیر گذار بر روی آرشـیوسازی فضای وب بریتانیا به صورت کتابخانه واسپاری برای همه انتشارات بریتانیا آماده می.کند در می، 2011 وزارت پژوهش و اختراع دانمارک، تصـمیم به ایجاد زیر ساخت های پژوهشی ملی در حوزه های پژوهشی مختلف (علوم طبیعی علوم انسانی) که متمرکز بر استفاده از ابزارهای تحلیلی منابع وب آرشیوی خواهد بود.

چالش فوری: عدم شـفافیت نحوه اسـتفاده پژوهش گران از آرشـیو های ملی. بسـیاری از آن ها هنوز در مراحل طراحی هستند ما در مورد یکی از مهم ترین مواردی که بایـد انجام شود، یعنی تعامل دامنه پژوهشـگران با تخصـص هایی نه فقط در پژوهش ،اینترنت بلکه در زمینه هـایی نظیر جـامعه شناسـی، علوم سیاسـی و دیگر حوزه هـای علوم اجتماعی فیزیک و سایر علوم هنر و علوم انسانی، بحث خواهیم کرد. همان طور که این زیرساخت ها به منظور انعکاس نیازهای پژوهشـگران ملی در مجموعه های ایجاد شده، طراحی می شوند. این فرآیندی زمانبر است و تعامل خبرگان هر حوزه می تواند مشکل باشد. با وجود

این شکست اجرای آن احتمال کسب کاربرد گسترده زیر ساخت های جدید را کاهش می دهد.

نتایج: مسیر پیش رو

آن چه گفته شد، فقط تعداد معدودی از مواردی است که گروه کوچک ما توانست به آن فکر کند و موارد بسیاری هنوز وجود دارد. بعضی از مباحث عمومی است زیرا فنون گام به گام ویژه ای برای مشخص نمودن مجموعه وب یا روندهای تحلیلی از اینکه چگونه محتوای وب در طول زمان تغییر می کند نیاز به منابعی از پروژه تحقیقاتی خاص، یک تیم از متخصصان حوزه و مجموعه های مرتبط برای آزمون این ،روش ها خواهد داشت ،بنابراین ما گوی جادویی نخواهیم داشت با این حال تعدادی از چالش های عمومی را نشان داده ایم که پژوهشگر علاقه مند به کار در حوزه آرشیوهای وب با آن مواجه است یکی از موارد اصلی که در طول زمان حاصل می شود و دوباره در مصاحبه ها و بحث ها وجود دارد، فقدان پایداری و رابطه ای کاربر پسند برای ایجاد آرشیو های وب و یکبار ساخته شده به منظور دسترسی و تحلیل داده های موجود در آن هاست.

در دراز مدت، ما امیدوار به نتایج دیگری هستیم که ممکن است از این تلاش ها حاصل شود. برای مثال، می توانیم گروه کاری پست هاگ (1) را تصور کنیم که برای توسعه کارگاه بحث و ایده های مورد استفاده احتمالی در آینده و تمرکز بر توسعه ابزاری ایجاد شد. این گروه نه تنها برای اعضای IIPC، بلکه برای انواع جوامع پژوهش گرانی که با جامعه آرشیوی وب در تعامل نبوده و بیش تر تمایل به استفاده از وب آرشیوی را دارند آگاهی رسانی خواهد کرد. مثال ها شامل پژوهش گران اینترنت (مانند اعضای 2) AOIR) دانشمندان اطلاعات (مانند 3) IFIP) و 4) ASIS T)) و طیف فهرست ها و انجمن های علاقه مند به علوم انسانی رقومی هستند. ما قویاً توصیه می کنیم که IPC نمایندگانی را به نشست های سالانه سازمان هایی از این نوع و پانل ها رهنمون می سازد و کارگاه هایی را برای تعامل پژوهشگران با امکانات آرشیوهای وب سازماندهی کند. ما چند ایده را به صورت برجسته مطرح ساخته ایم اما آن جوامع موارد بیش تری را رانند آماده نمایند. برای آمدن آن ها به IIPC منتظر نمانید. IPC باید به سمت آن ها برود.

ایده دیگر برای آینده فعالیت، نوعی برنامه مشترک کدنویسی نرم افزاری (5) است که برنامه نویسان رایانه ای هکرها با همدیگر و با پژوهشگران برای 2 الی 3 روز گردهم آمده و به ایجاد دسترسی به داده های آرشیوی وب توجه و همکاری می کنند آن ها باید بتوانند به گروه ها و وظایف مرتبط با کشف رهیافت های خلاقانه و ابتکاری توجه کنند تا بتوانند داده ها و ابزارهای موجود را برای ایجاد سریع ابزارها و رابطه ای جدید ، به کار گیرند پژوهش گران انتخاب خواهند شد به دلیل این که آن ها سؤالاتی دارند

1- (گروهی کاری در لاهه) post-Hague

2- AoIR)

3- IFIP

4- ISAST

5 – Hackathon: برنامه ای است که طی آن تعدادی از برنامه نویسان به صورت مشترک و فشرده در یک دوره کوتاه زمانی اقدام به کدنویسی برای برنامه نرم افزاری خاص می نمایند منبع قابل دسترسی در: http://www.techopedia.com/definition/23193/hackathon

که تمایل به پاسخ گویی به آن ها دارند و برنامه نویسان رایانه ای همه مهارت شان را برای کمک به آن ها در رسیدن (نیل یا نزدیک شدن) به اهـداف پژوهشـی شان به کار می برنـد دوباره برنامه نویسان رایانه ای که با داده های وب پویا کار می کنند، مهارت هایی برای اجرای طرح های خلاقانه با ابزارها دارند که نتایج بسیار بیشتری از انتظار برای درخواست از آرشیو های وب به سمت آن ها گسیل خواهد کرد.

آیا ما به نیروانا خواهیم رسید یا محکوم به آخر الزمان خواهیم ،شد یا با منحصر به فردی (فنی) جایگزین خواهیم شد یا اینکه با آرشیو های غبار آلوده مواجه خواهیم شـد؟ هیچ راهی برای دانسـتن آن .نداریم با این حال ما در نقطه ای هستیم که این سؤال را ایجاب می کند امروز برای اطمینان از این که در آینده به چه چیزی می توانیم دسترسی داشته باشیم چه گام هایی می توانیم برداریم؟ تجمیع تصمیمات به ندرت قابل توجه ، ساده ،نبود اما بخشـی از تلاشی است برای تضمین این که آرشیوهای وب قوی پایدار قابل دسترسی ، ارزشمند و بالاتر از همه توسط پژوهش گران آینده قابل استفاده باشند.

منابع

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., Menczer, F. (2011). . 1 Political Polarization on Twitter. Paper presented at the ICWSM: International Conference on Weblogs and .Social Media 2011, Barcelona

Conover, M. D., Ratkiewicz, J., Gonçalves, B., Flammini, A., Menczer, F. (2011). The Echo Chamber. . 2 Paper presented at the Journal of Information Technology Politics Conference 2011: The Future of .Computational Social Science, Seattle

Dougherty, M., Meyer, E. T., Madsen, C., Van den Heuvel, C., Thomas, A., Wyatt, S. (2010). Researcher .3 Engagement with Web Archives: State of the Art. Report. London: JISC. Retrieved from ./http://ssrn.com/abstract=1714997 and http://ie-repository.jisc. ac.uk/544

Garfinkel, S. Cox, D. (2009, 9-11 February). Finding and Archiving the Internet Footprint. Paper . 4 presented at the First Digital Lives Research Conference: Personal Digital Archives for the 21st Century, .London

.(Gazan, R. (2008). Social annotations in digital library collections. D-Lib Magazine, 14(11/12 .5

Hindman, M. (2007). "Open-source politics" Reconsidered: Emerging Patterns in Online Political . 6 Participation. In V. Mayer-SchÖnberger D. Lazer (Eds.), Governance and information technology: From .electronic government to information government (pp. 183-207). Cambridge: The MIT Press

Hogan, B. (2010). Analyzing Facebook Networks. In D. Hansen, M. Smith B. Schneiderman (Eds.), . 7
Analyzing Social Media Networks with NodeXL. New York, NY: Morgan Kaufman

Jasra, M. (2011, 3 February). Reddit Surpasses 1 Billion Monthly Page Views Retrieved 30 April, 2011, .8
from http://www.webanalyticsworld. net/ 2011/ 02/ reddit-surpasses- 1-billion-monthly-page.html.
((Archived by WebCite® at http://www.webcitation. org/5yKdMBKNc

Kay, A. (1995). The Best Way to Predict the Future is to Invent it. Mathematical Social Sciences, 30, . 9
.326-326

Klett, M., Manchester, E., Verburg, J. (1984). Second View: The Rephotographic Survey Project. . 10
.Albuquerque: University of New Mexico Press

Kling, R., McKim, G., King, A. (2003). A Bit More to IT: Scholarly Communication Forums as Socio- .11
Technical Interaction Networks. Journal of the American Society for Information Science and Technology,
.54(1), 46-67

.Kurzweil, R. (2005). The Singularity is Near: When Humans Transcend Biology. New York: Viking .12

Mayer-SchÖnberger, V. (2009). Delete: the virtue of forgetting in the digital age. Princeton, NJ: . 13
.Princeton Univ Press

Meyer, E. T. (2006). Socio-technical Interaction Networks: A discussion of the strengths, weaknesses .14
and future of Kling's STIN model. In J. Berleur, M. I. Numinem J. Impagliazzo (Eds.), IFIP International
Federation for Information Processing, Volume 223, Social Informatics: An Information Society for All? In
.Remembrance of Rob Kling (pp. 37-48). Boston: Springer

Meyer, E. T. (2011). Splashes and Ripples: Synthesizing the Evidence on the Impact of Digital . 15
.Resources. Report. London: JISC. Retrieved from http://ssrn.com/ abstract=1846535

Meyer, E. T., Eccles, K., Thelwall, M., Madsen, C. (2009). Final Report to JISC on the Usage and . 16
Impact Study of JISC-funded Phase 1 Digitisation Projects the Toolkit for the Impact of Digitised Scholarly
Resources (TIDSR). Retrieved from http://microsites. oii.ox.ac.uk/tidsr/system/files/TIDSR-Final Report-
20July2009.pdf

.Moretti, F. (2005). Graphs, Maps, Trees: Abstract models for a literary history. London: Verso Books .17

.Moretti, F. (2011). Network Theory, Plot Analysis. New Left Review, 68, 80–102 .18

Olston, C., Reed, B., Srivastava, U., Kumar, R., Tomkins, A. (2008). Pig Latin: A Not- So-Foreign . 19
Language for Data Processing. Paper presented at the ACM SIGMOD'08 Conference, Vancouver, BC,
.Canada

Schroeder, R. (2011). Being There Together: Social Interaction in Shared Virtual Environments. New . 20
.York, NY: Oxford University Press USA

Schroeder, R. Meyer, E. T. (2009). An Emerging Global Brain: How the Internet is Revolutionising . 21
Scientific Research. Britain in 2009 (Economic Social Research Council Annual Magazine), 113.27

Tanner, S. (2010). Inspiring Research, Inspiring Scholarship. Report. London: JISC. Retrieved from . 22
http://www.jisc.ac.uk/media/documents/programmes/digitisation/12pag
.efinaldocumentbenefitssynthesis.pdf

Tanner, S. Deegan, M. (2011). Inspiring Research, Inspiring Scholarship: The value and benefits of . 23
digitised resources for learning, teaching, research and enjoyment. Report. London: JISC. Retrieved from
http://www.kdcs.kcl.ac.uk/fileadmin/documents/ Inspiring-Research-Inspiring-Scholarship-2011-Simon
.Tanner.pdf

Thomas, A., Meyer, E. T., Dougherty, M., Van den Heuvel, C., Madsen, C., Wyatt, S. (2010). . 24
Researcher Engagement with Web Archives: Challenges and Opportunities for Investment. Report. London:
./JISC. Retrieved from http://ssrn.com/abstract=1715000 and http://ie-repository.jisc.ac.uk/543

van den Heuvel, C. (2009). MAPS: Manuscript Map Annotation and Presentation System: Linking . 25
formal ontologies with social tagging to (re-) construct relationships between manuscript maps and
contextual documents. Digital Humanities 2009 (University of Maryland, Maryland Institute for
.Technology in the Humanities (MITH) Abstracts), 138- 141

Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M. (2008). reCAPTCHA: Human-Based .26
.Character Recognition via Web Security Measures. Science, 321(5895), 1465-1468

Williams, D., Yee, N., Caplan, S. E. (2008). Who plays, how much, and why? Debunking the . 27
stereotypical gamer profile. Journal of Computer-Mediated Communication, 13(4), 993-1018. doi:
10.1111/j.1083-6101.2008.00428.x

فصل دوم: تجارب جهانی و مسائل بومی در آرشیو سازی وب

اشاره

ص: 149

چکیده

کتابخانه ملی ،استرالیا مؤسسه راهبر آرشیو و حفاظت رقومی در استرالیاست. آرشیو پاندورا (1)، که بیش از 10 سال مخزن آرشیو منابع وب استرالیا بوده است، سیستمی کامل و در حال پیشرفت و توسعه مداوم است. پانداس PANDAS، سیستم مدیریت ،آرشیوی که آرشیو را پشتیبانی می کند از سال 2007 ، در حال سومین بازبینی خود .می باشد فعالیت های دیگر آرشیو وب از جمله گردآوری سالانه دامنه استرالیا و استفاده از آرشیو – آی تی (2) با همکاری آرشیو اینترنت اداره می شوند.

این مقاله به بررسی وضعیت کنونی آرشیو وب در استرالیا می پردازد و نشان می دهد که چگونه کتابخانه ها خدمات خود را با ورود فزاینده مواد برخط در مجموعه های شان وفق می دهند سالیان متمادی تصور می شد که با آرشیو کردن فقط می توانیم نمونه ای کوچک اما روایت گر از اینترنت را ضبط کنیم. امروزه شکاف بین آن چه موجود است و آن چه می تواند آرشیو شود در حال کاهش است. درعین حال هر چه آرشیو ها بیشتر می شوند و توانایی ما در آرشیو کردن افزایش می یابد بیش تر با فناوری های جدید و برنامه های کاربردی وب 2/0 درگیر خواهیم بود به عنوان مثال در انتخابات فدرال سال 2007 که تعداد وسیعی از سایت های متعامل نظیر MySpace Kevin و YouTube آرا را آرشیو کردند، معلوم شد که آرشیوکننده های استرالیایی وب به سازگاری ادامه می دهند و با چالش های جدید روبه رو می شوند.

PANDORA -1

Archive-It -2

آرشیو وب در دنیای وب 2/0 شعبه آرشیو وب و حفاظت رقومی کتابخانه ملی استرالیا

اشاره

*آرشیو وب در دنیای وب 2/0 شعبه آرشیو وب و حفاظت رقومی کتابخانه ملی استرالیا (1)

ادگار کروک (2) | ترجمه: مرجان هادی زاده (3)

مقدمه

مقاله حاضر آرشیو وب در استرالیا را مورد بحث و بررسی قرار می دهد. از آن جا که کتابخانه ملی استرالیا نقش اصلی را در این زمینه دارد و همچنین با توجه به این که نویسنده خود در آن جا شاغل است این مقاله می تواند تأثیر زیادی بر مسائل این مؤسسه داشته باشد. لازم به ذکر است که در استرالیا پروژه های آرشیو وب دیگری نیز وجود دارند نظیر پروژه Our Digital Island در منطقه تاسمانی (http://odi.statelibrary.tas.gov.au/) و پروژه خدمات Territory Stories (http://www.territorystories.nt.gov.au/) در منطقه شمالی.

اخیراً کتابخانه ملی ،استرالیا برای اجرای آرشیو وب از سه روش شناسی مختلف استفاده می کند. آرشیو انتخابی در پاندورا (آرشیو وب استرالیا)، طی قراردادی با آرشیو ،اینترنت گردآوری کل دامنه و بهره برداری از خدمات آرشیو - آی تی را با همکاری یکدیگر انجام داده اند. با این روش به سمت ایجاد مجموعه کامل و جامع نشریات برخط استرالیا حرکت می کنیم البته با افزایش چالش های نوین فناوری کتابخانه مجبور است برای تداوم این امر مهم روش هایی برای تعدیل طرح های آینده اتخاذ کند؛ نظیر محدوده مجموعه ها و گسترش همکاری های جدید

1- Web Archiving in a WEB 2/0 World

2- Edgar Gruk

3- کارشناس ارشد سازمان اسناد و کتابخانه ملی ایران

سه روش شناسی آرشیو

آرشیو پاندورا (http://pandora.nla.gov.au)، انتشارات وب استرالیا را از سال 1996 بایگانی کرده است؛ و در همان زمان به عنوان مخزن مورد قبول در میان تمام ملت ها معرفی شد. اصلی ترین موفقیت پاندورا ساخت شبکه ای مشتمل بر 9 گروه آرشیو استرالیایی بوده است مانند تمام کتابخانه های دولتی قاره استرالیا و AIATSIS (مؤسسه استرالیایی مطالعات بومیان و جزیره نشینان تنگه تارس)، آرشیو ملی فیلم و صدا و یادمان جنگ استرالیا، بنیانگذاری پانداس – سیستم گردش کار آرشیو که سیستمی است برای شناسایی مداوم مطالب آرشیو شده و ارتباط با طیف وسیعی از سازمان های نمایه سازی و چکیده نویسی– و همچنین تعداد بی شماری از ناشران استرالیایی؛ به گونه ای که در اول جولای 2008 ، آرشیو، در برگیرنده 307، 19 عنوان شامل 531,140,080 فایل بالغ بر 2/2 ترابایت داده بوده است.

آرشیو پاندورا موضوع های زیر را آرشیو کرده است:

- انتشارات برخط منتخب استرالیا در سطح جهانی نظیر مجله های الکترونیکی، انتشارات دولتی و وب گاه های مهم پژوهشی و یا فرهنگی؛

- خط مشی ها، شیوه ها و دستورالعمل های منتخب برای مجموعه و تأمین دسترسی بلند مدت به مواد موجود در آرشیو؛

- رویکرد ملی مشترک برای آرشیو و حفاظت بلند مدت انتشارات پیوسته استرالیا ، شامل مشارکت کتابخانه های دولتی و دیگر مؤسسه های فرهنگی؛

- سیستم آرشیو رقومی (پانداس) برای جمع آوری و بارگذاری ساده و مؤثر انتشارات در آرشیو، ذخیره اطلاعات در مورد آن ها و مدیریت دسترسی عمومی به آن ها؛

- طرحی برای نام گذاری مستمر تمام اشیای موجود در آرشیو و ارائه خدمات با دقت به آن ها؛

- تنظیم مقررات همکاری با مؤسسه های نمایه سازی و چکیده ،نویسی در آرشیو پاندورا، به منظور استناد و دسترسی دائمی به مواد بایگانی شده و شناسه های ماندگار برای انتشارات نمایه سازی و چکیده نویسی در آن ها؛ و

- محتوا که در اول جولای 2008 بالغ بر 307، 19 عنوان مشتمل بر 531,120,080 ،فایل به میزان 2/2 ترابایت داده می باشد

گستره وسیعی از انتشارات وجود دارد نیمی از آن ها از وب گاه های دولت فدرال و ایالتی آرشیو شده اند و نیمی دیگر بازتابی از تنوع کامل فرهنگ و پژوهش استرالیاست نوع انتشارات درون آرشیو می تواند یک تک مدرک PDF یا یک وبگاه کامل سازمانی شامل هزاران فایل باشد در آرشیو ، وبنوشت، پادکست، و فیلم نیز وجود دارند.

جمع آوری گزینشی هدایت شده انتشارات ،وب بر اساس ارزش بلند مدت فرهنگی و پژوهشی ،آن ها بدان معناست که تنها بخش کوچکی از دامنه استرالیا آرشیو شده است کتابخانه، از سال 2005 ، این مطلب را دریافت و از آن زمان با آرشیو اینترنت (http://www.archive.org) قرارداد بست که خزش های هدایت شده سالانه را به منظور جمع آوری هر ماده ممکنی از دامنه استرالیا، انجام دهد. این

خزش ها هر سال حـدود یک ماه صورت می گیرد و مقـدار داده هایی که جمع آوری می کنـد به انـدازه ای است که محتوای پاندورا را کـم جلوه می دهنـد؛ به عنوان مثال در سال 2007 از کل گردآوری دامنه 18 ترابایت داده در یک ماه برداشت شـد در حالی که پاندورا در 11 سال 2 ترابایت داده جمع آوری کرده است. انتظار می رود جمع آوری سال 2008 یک بیلیون فایل باشد.

عکس

خزش‌ها، هر سال حدود یک ماه صورت می‌گیرد و مقدار داده‌هایی که جمع‌آوری می‌کند به‌اندازه‌ای است که محتوای پاندورا را کم جلوه می‌دهند؛ به‌عنوان مثال، در سال ۲۰۰۷، از کل گردآوری دامنه ۱۸ ترابایت داده در یک ماه برداشت شد، در حالیکه، پاندورا در ۱۱ سال ۲ ترابایت داده جمع آوری کرده است. انتظار می‌رود جمع‌آوری سال ۲۰۰۸ یک بیلیون فایل باشد.

گردآوری دامنه وب استرالیا : تجزیه و تحلیل کمی مقدماتی آرشیو داده ، NLA ، ۲۰۰۸ –

تاریخ گردآوری دامنه	۲۰۰۵	۲۰۰۶	۲۰۰۷
مدارک (فایل‌های) تکی خزش شده	۱۸۵۵٤۹٦٦۲	۵۹٦۲۳۸۹۹۰	۵۱٦۰٦٤۸۲۰
کل مدارک (فایل‌های) خزش شده	۱۸۹۸۲٤۱۱۹	٦۲۱٦٦٤۸۷٦	۵۲۳۵۱۰۹٤۵
میزبان‌ها	۸۱۱۵۲۳	۱۲٦۰۵۵۳	۱۲٤۷٦۱٤
اندازه داده‌های خام	٦،٦۹ ترابایت	۱۹،۰٤ ترابایت	۱۸،٤۷ ترابایت
اندازه فایل‌های ARC فشرده شده	٤،۵۲ ترابایت	۱۰،٤۸ ترابایت	۱۰،۱۸ ترابایت

HTTP://PANDORA.NLA.GOV.AU/DOCUMENTS/AUSCRAWLS.PDFKOERBIN, P.

گردآوری‌هایی که با استفاده از HERITRIX هدایت شده‌اند (http://crawler.archive.org/) بزرگ است، ولی کاملاً جامع نیستند، زیرا فقط یک ماه در هر سال گردآوری می‌شوند (و مقادیر زیادی می‌توانند در این فاصله زمانی در اینترنت بیایند و بروند)، آنها از قوانین robots.txt پیروی می‌کنند، و اگرچه HERITRIX عملکردی قوی دارد، وبگاه‌هایی وجود دارد که از لحاظ فنی خزش به آنها مشکل است. نقص‌ها هرچه باشند، گردآوری در اندازه قابل توجهی انجام می‌گیرد و بنابراین مقادیر زیاد داده گردآوری شده باعث می‌شود هرگونه تلاشی در جهت ارزشیابی کیفی وبگاه‌های شخصی مشکل گردد. بنابراین، به‌طور معکوس، پاندورا، جایی است که می‌توان مشکلات را درون هر عنوان شناسایی کرد و تحلیل کیفی را انجام داد– در صورت لزوم صفحه‌ها را گردآوری کرد و تثبیت کرد – که در اینجا ممکن نیست. یکی دیگر از اشکال‌ها این است که بر خلاف پاندورا –که در آن اجازه ناشر برای آرشیو کسب می‌شود– در اینجا چنین امکانی وجود ندارد. بنابراین، با توجه به قانون کپی‌رایت (حق تألیف) فعلی استرالیا که طبق آن نشریات برخط شامل واسپاری قانونی نیستند، در حال حاضر، قادر نیستیم آنچه را که حفظ شده به عموم نمایش دهیم. البته،این عبارت به‌این معنی نیست که هیچ کاری در این آرشیو انجام نمی‌شود، چرا که دانشگاهیان در حال کار بر روی این داده‌ها هستند و به‌نظر می‌رسد تحقیقات آنها

گردآوری دامنه وب استرالیا : تجزیه و تحلیل کمی مقدماتی آرشیو داده ،2008 ، NLA –

گردآوری هایی که با استفاده از HERITRIX هدایت شده اند (http://crawler.archive.org) بزرگ است، ولی کاملاً جامع نیستند زیرا فقط یک ماه در هر سال گردآوری می شوند (و مقادیر زیادی را می توانند در این فاصله زمانی در اینترنت بیایند و بروند آن ها از قوانین robots.txt پیروی می کنند، و اگر چه HERITRIX عملکردی قوی دارد وب گاه هایی وجود دارد که از لحاظ فنی خزش به آن ها مشکل است. نقص ها هر چه باشند گردآوری در اندازه قابل توجهی انجام می گیرد و بنابراین مقادیر زیاد داده گردآوری شده باعث می شود هر گونه تلاشی در جهت ارزشیابی کیفی وب گاه های شخصی مشکل گردد. ،بنابراین به طور معکوس ،پاندورا جایی است که می توان مشکلات را درون هر عنوان شناسایی کرد و تحلیل کیفی را انجام داد– در صورت لزوم صفحه ها را گردآوری کرد و تثبیت کرد – که در این جا ممکن .نیست یکی دیگر از اشکال ها این است که بر خلاف پاندورا–که در آن اجازه ناشر برای آرشیو کسب می شود– در اینجا چنین امکانی وجود ندارد ،بنابراین با توجه به قانون کپی رایت (حق تألیف) فعلی استرالیا که طبق آن نشریات برخط شامل واسپاری قانونی نیستند در حال حاضر، قادر نیستیم آن چه را که حفظ شده به عموم نمایش دهیم البته این عبارت به این معنی نیست که هیچ کاری در این آرشیو انجام نمی شود، چرا که دانشگاهیان در حال کار بر روی این داده ها هستند و به نظر می رسد تحقیقات آن ها

ارزشمند باشد آرشیو- آی تی سرویس آرشیو وب های میزبانی شده توسط آرشیو اینترنت است. اولین (و تا کنون تنها) سازمانی که در استرالیا از این سرویس استفاده کرده و می کند بخش مجموعه های آسیایی کتابخانه ملی (http://www.nla.gov.au/asian/asianwebarchive html) است.

از آرشیو-آی تی برای جمع آوری مجموعه ای از وب گاه های خارج از کشور استفاده می شود که رویدادهای خاص اجتماعی و سیاسی را ضبط می کنند؛ زیرا انتظار نمی رود که هیچ سازمان دیگری در منطقه این نقش را به انجام برساند گزینه میزبانی نیز از آن رو انتخاب شد که به نظر می رسید می تواند راهی سریع و آسان برای جمع آوری و نگهداری مجموعه ها باشد و به مهارت های فنی نیاز ندارد و میزان زیادی از وقت کارکنان را نمی گیرد. در حالی که خیلی زود معلوم شد فقط قسمتی از این واقعیت دارد چرا که برای ساخت موفقیت آمیز مجموعه ها زمانی بسیار زیادتر از آن چه تصور می شد صرف شد گزینش وب گاه ها برای ،خزش فعالیتی است که اغلب با سو تفاهم همراه است و می تواند به طور شگفت آوری زمان زیادی بطلبد.

در حالی که استفاده از آرشیو آی تی دارای این مزیت است که دیگر در مورد میزبانی و حفظ محتوای جمع آوری شده نگرانی وجود نخواهد داشت. البته برخی مشکلات عمده در مورد عدم کنترل مواد آرشیوی و نمایش توابع وجود دارد. آرشیو آی تی اجازه می دهد تا برخی URL های هسته انتخابی جمع آوری شوند؛ با وجود این، اگر فایل های جمع آوری شده از بین بروند چه جمع آوری بشوند چه جمع آوری نشوند، هیچ راهی برای اصلاح خرابی یا از دست دادن محتوا وجود ندارد که بتوانید با سیستم خود انجام دهید به همین ترتیب، کنترل واقعی یا مالکیت فرآیند نمایش نیز وجود ندارد؛ به طوری که به عنوان مثال یک پیوند به URL هسته که جمع آوری نشده ، هم چنان درون یک مجموعه ظاهر می.شود یکی دیگر از اشکال ها این است که اگر شما اشتراک سالیان خدمات خود را قطع کنید مجموعه های شما به مخزن محتوای آرشیو اینترنت عمومی باز می گردد. با وجود این مسائل کتابخانه در نظر دارد به آرشیو با استفاده از این روش ادامه دهد.

سایت های آرشیو شده با استفاده از آرشیوآی تی برای مجموعه های آسیا - کتابخانه ملی استرالیا

Papua New Guinea Government and Research Websites

(http://archive-it.org/collections/1039)

وب گاه های انتخابی مؤسسه های دولتی و پژوهشی مهم پاپوآ گینه نو که از سال 2008 آرشیو شده اند. برخی سایت های آرشیوی که در زمان ضبط کردن جاری نبودند.

(http://archive-it.org/collections/918)

(http://archive-it.org/collections/1040)

وب گاه انتخابی بین المللی بین الدولی مرتبط با انتخاب عمومی سال 2007 تایلند

وب گاه های منتخب دولتی، احزاب سیاسی و رسانه های مرتبط با انتخابات ملی کامبوج سال 2008

(http://archive-it.org/collections/937)

(http://archive-it.org/collections/1054)

وب گاه های منتخب بین المللی مرتبط با شورش راهبان برمه سپتامبر- اکتبر 2007

وب گاه های انتخابی دولتی و غیر دولتی جمهوری دموکراتیک خلق لائوس که از سال 2008 آرشیو شده اند.

ص: 154

(http://archive-it.org/collections/920)

وب گاه های انتخابی بین المللی بین دولتی و پاپوآگینه نو مرتبط با انتخابات پارلمانی پاپوآگینه نو در سال 2007 . شامل تصاویری از وبگاه کمیسیون انتخابات پاپوآگینه نو ، دادخواست انتخابات پاپوآگینه ،نو گزارشی از گروه ارزشیابی انتخابات انجمن جزایر مشترک المنافع اقیانوس آرام و بیانیه رسانه ای شفافیت بین المللی (پاپوآگینه نو).

(http://archive-it.org/collections/912)

وب گاه های منتخب بین المللی بین دولتی مرتبط با انتخابات ریاست جمهوری و پارلمانی تیمور شرقی در سال 2007 شامل آموزش رأی دهندگان و مواد تبلیغاتی سیاسی

وب گاه های منتخب اندونزیایی

جمع آوری فایل ها

(http://archive-it.org/collections/664)

وقتی پاندورا پا به عرصه وجود گذاشت به علت ناتوانی اولیه نرم افزارهای جمع آوری وب، بسیاری از وب گاه ها نمی توانستند جمع آوری شوند وب گاه های با پایه HTML می توانستند جمع آوری شوند، ولی سایت هایی با چنین قالب های ساده ای در ابتدا بسیار مشکل به وجود آوردند تا یک مشکل حل می گردید مشکل دیگری پیدا می شد بنابراین برای آرشیویست های وب پیشرفت شگفت انگیز اینترنت جاوا ،اسکریپت اپلت ها، شیوه نامه های آبشاری Shockwave flash ، و ده هزار فایل دیگر و انواع قالب ها فقط مشکلاتی پس از دیگری بود در حالی که مشکلات اغلب انواع فایل ها بر طرف شده اند، محتوای چندرسانه ای یک مسئله باقی مانده است پیش از این در مورد فایل های پخش (Real Player) و در حال حاضر، درباره پادکست ها مجبوریم نه با پیچیدگی خود فایل ها که با پیچیدگی سیستم های تحویل آن ها مقابله کنیم در حال حاضر این امر در مورد فیلم ها این مسئله لاینحل مانده است.

مجموعه انتخابات فدرالی سال 2007 بزرگ ترین تلاش تاکنون بوده است کتابخانه ملی مسئول آرشیو تمام منابع ملی مرتبط با انتخابات بود از جمله وب گاه های احزاب ،گروهای لابی وب گاه های برخی نامزدها، وب نوشت ها، فیلم ها و وب گاه های رسانه ای کتابخانه های ایالتی مسئول جمع آوری وب گاه های نامزدها، احزاب، و رسانه های محلی در ایالت خود بودند در مجموع بیش از 350 وبگاه توسط کتابخانه ملی و همکارانش آرشیو شده است، بسیاری از این سایت ها چندین بار جمع آوری شدند تا محتوای در حال تغییر را ضبط کنند بزرگ ترین چالش در آرشیو کردن این انتخابات تعداد زیاد فیلم ها بود؛ البته مشکلی در خود آن ها نبود، بلکه مشکل در مکانیزم تحویل و فناوری های تعبیه شده در آن ها به منظور استفاده مفید کاربران بود.

رویکردهای متفاوتی برای آرشیو کردن فیلم ها بسته به ماهیت وب گاه ها به کار گرفته شده است. برای وب گاه های عمومی که در آن فیلم های مجزا در یک صفحه وب موجود است فایل های فیلم را جداگانه با استفاده از ابزارهای متفاوت و رایگان موجود بارگیری کردیم (گردآورندگان وب به طور کلی نمی توانند فیلم ها را به طور خودکار جمع آوری کنند)؛ و سپس با استفاده از مبدل های فایل فایل ها را از flv. به چیزی کاربر پسند تر مانند قالب Mpeg تغییر دادیم جایی که تعدادی فیلم به یک صفحه وب پیوند داشت، فیلم ها در قالب اصلی

خود شان باقی ماندند و یک پخش کننده flv. در فایل های جمع آوری شده نصب گردید، به طوری که فایل ها می توانند به راحتی وب گاه زنده ارائه شوند وقتی که وب گاه انتخابات یوتیوب

ص: 155

(http://nla.gov.au/nla.arc-76644 شامل بیش از 700 فیلم) را جمع آوری کردیم، بـا مراجعه به مهارت های فنی در بخش فناوری اطلاعـات توانستیم برای فیلم ها URL هـا را از سایت جاری اسـتخراج کنیم آن ها را بارگیری کنیم و تغییرات ضـروری را برای گردآوری صفحات وب انجام دهیم.

هیـچ یک از این فرآینـد ها سـریع نبود و همه به مقدار نسـبتاً خوبی از مهارت های فنی نیاز داشـتند، از جمله کد گذاری مجدد صـفحات آرشـیو شـده با تغییراتی که بایـد انجام می دادیم تا فیلم ها درون آرشـیو قابل پخش باشـند. همچنین به خاطر انتخابات حفاظت از اسـناد برخـط در دولت پیشـین نیز به ما سپرده شـد. با پیش بینی این امر طرح هایی (همان گونه که برای انتخابات پیشـین داشتیم) ساخته بودیم و تمام وب گاه های وزارتی دولت را آرشـیو کرده بودیم؛ در حالی که آن ها درست قبل از تاریخ انتخابات در حالت مستحفظ بودند این کار صورت گرفت زیرا گمان می رفت با یک تغییر احتمالی در دولت حـذف کلی محتوا از اینترنت وجود داشـته باشـد همان گونه که در حوزه های قضایی دیگر اتفاق افتاده .بود. از آن جا که انتشارات وب برخی بخش های دولتی به ویژه بخش هایی که دارای اسناد تغییر یافته بودند، از دید عموم حذف شده ،بود این دور اندیشی اجر نهاده شد.

دستورالعمل های جمع آوری

با توجه به گردآوری دامنه استرالیا توسط کتابخانه ملی و آرشـیو انتخابی پاندورا ، می توان گفت که حجم قابل توجهی از نشـریات استرالیا و یا وب گاه ها آرشـیو شـده انـد اگر چه نمی توانیم بگوییم تا چه حد این مجموعه جامع و کامل است با این حال می دانیم که شـکاف های بزرگی باقی مانده است.

ما، وب گاه های تولیدی توسط استرالیایی های خارج از دامنه استرالیا را به طور جامع آرشـیو نمی کنیم (اگر چه امیدواریم آرشـیو اینترنت بسـیاری از آن ها را به دست خواهد آورد). خارج از نشریات مرسوم ، اما تا اندازه ای دارای اهمیت بیشتر ، ما هم چنین میزان بسیار زیادی از محتوای خلاقانه ای را که توسط افراد تولید شده و در وب گاه ها ، وب نوشت ها ، دنیای مجازی و سایت های شبکه های اجتماعی فیلم ، عکس و هنر میزبانی شده اند، جمع آوری نکرده ایم.

تلاش هایی برای جمع آوری برخی از این محتواها وجود دارد اما پروژه های هـدایت شـده کوچکی هسـتند. یکی از آن ها مجموعه رقص استرالیـایی است که به دنبـال گردآوری نمونه هـای از رقص استرالیـایی به محض قرار گرفتن بر روی وب گـاه هـای مختلف و سـایت های میزبان فیلم می باشد.

اگر چه کتابخانه ملی با Flickr قرارداد بسته است و از MySpace و YouTube اجازه آرشیو کردن دریافت کرده ایم کتابخانه فقط بخش کوچکی از این منـابع را آرشـیو یـا گردآوری کرده است. هم چنین، دیگر منابع برخط هر جایی که فعالیت های استرالیایی وجود دارد نظیر فضاهای مجازی (Second Life غیره) و شـبکه های اجتماعی (Facebook, Bebo و غیره) نیز آرشـیو نشـده اند. نخستین دلیل این که معمولاً محتوا دارای حق نشـر و حفظ حریم خصوصی است که آرشیو کردن آن را مجاز نمی کند و یا به دلیل ماهیت منابع که آن را خارج از اینترنت عمومی قرار می دهد.

ما به طور جداگانه به عنوان کتابدار نیز می توانیم تفاوتی نسبت به حفظ میراث برخط خودمان به وجود آوریم؛

البته بـا پـذیرش این مسـئولیت که می توانیم درون سازمـان خود یـا سازمـان مادر از حفظ و پشتیبانی آن چه در وب گاه سازمان قرار گرفته اطمینان حاصل نمائیم این ،جنبش به ویژه به طور قابل توجهی در دولت و دانشگاه ها در مورد حرکت نشریات از چاپی به برخط ادامه می یابد آن چه ما پیدا کرده ایم این است که نباید مطمئن باشـیم که نشـریات برخط در وبگاه ناشـران در بلند مدت (و یا حتی کوتاه مدت) در دسترس باقی بمانند دانشگاه ها در حال حاضر، مجبور به ایجاد مخازن رقومی برای خروجی فکری خود هستند و در این راه نشریات آن ها دسترس پذیر باقی می مانند. در حالی که ممکن است انتظار داشـته باشـیم که وب گاه های دولتی دسترسـی عمومی خود را حفظ نمایند؛ تجربه نشان می دهد که این موضوع همیشه صادق نیست بنابراین اگر نشـریه ای برای مجموعه و یا کاربران شما ارزشمند است، عاقلانه این است که تلاشی در جهت حفظ دسترسی بلند مدت آن انجام شود.

دستور عمل های آینده

تا زمانی که چشم انداز در حال رشد اینترنت وجود دارد فناوری های جدید و شکاف هایی که تازه شناسایی شده اند در گردآوری ما وجود دارند که پرداختن به آن ها ضـروری است بنابراین، به نظر نمی رسد آرشـیو وب به منطقه ای تبدیل شود که در آن شیوه ها و یا پروتکل های توسـعه مجموعه کاملاً تأسـیس و یا برقرار گردد. ما همیشه به شـناسایی و جمع آوری مطالب نیاز داریم نه این که منتظر بمانیم تا آن ها به سوی ما بیایند. وب بسـیار پویاست و فناوری بسـیار متغیر تعداد ناشران آن قدر وسیع است که بعید به نظر می رسد ما قادر به ایجاد سبکی مقیاس پذیر – نظیر سبک سیستم سپرده فیزیکی که نیاکان ما برای مواد چاپی انجام داده اند باشیم. هنگامی که کتابخانه ملی برای اولین بار شروع به آرشـیو کردن وب کرد، ابزار و نهادهای بسـیار کمی در خارج وجود داشت که ما می توانستیم برای یادگیری با آن ها کار کنیم. در نتیجه، باید مکانیسم ها و ابزار ها را خود مان اختراع می کردیم برای این ،منظور کتابخانه ملی در درون خودش سیستم مدیریت آرشیوی را – PANDAS – ساخت این سیستم در حال حاضر، در سومین و طبق تصور ما آخرین مرحله خود است چرا که کتابخانه دیگر نمی تواند به طور مسـتقل چنین سـرمایه گذاری را برای توسـعه متحمل شود. با این حال در حال حاضـر که آرشیو وب در طیف وسیعی از نهاد های بین المللی به طور عملی ایجاد شـده ،است شـرکای زیادی وجود دارنـد که می توان این بار را با آن ها تقسـیم کرد. کنسرسـیوم بین المللی حفاظت از اینترنت (1) که اعضای آن متشکل از تمام کتابخانه های ملی راهبر آرشیو وب از جمله کتابخانه ملی استرالیا و دیگر نهاد های مرتبط می باشد این توسـعه را هدایت و راهبری می کند با این ،روش کتابخانه می تواند به نقش راهبری خود در آرشیو ،وب از طریق تطابق با ابزار های در حال توسعه و سازگاری با کتابخانه ها و مؤسسه های همکار ادامه دهد.

شرح حال مختصری از پدید آورنده

ادگار کروک از 1999، در کتابخانه ملی استرالیا کار می کند. او از سال 2000 بر روی پاندورا: آرشیو وب استرالیا کار می کند. قبل از آن در کند. قبل از آن در کتابخانه عمومی ATC کار می کرده است.

IIPC -1

چکیده

همان گونه که در بسیاری از متون نیز مطرح است زبان فارسی به واسطه ویژگی های خاص خود چه از نظر رسم الخط و چه از نظر صرف و معنـا با چالش های منحصـر به فردی در زمینه ذخیره و بازیابی اطلاعات رو به روست در مقالات مختلف به این مشـکلات در قالب ها و سطوح مختلف بـدون توجه به نوع شناسـی این چالش ها اشاره شـده است. به عبارت بهتر کمتر مقاله ای در زبان فارسـی در حوزه اطلاع رسانی از دیدی مبتنی بر علم زبان شناسـی مدرن به بحث دسـته بندی این مشکلات و پاسخ دهی به آن ها پرداخته است. نوشتار حاضر بر آن است تا عمده ترین چالش های مطرح این زمینه را در سه گروه رسم الخط مسائل صـرفی و مسائل معنایی مورد اشاره قرار داده و پس از بیان نمونه هایی در پیونـد با هر یک از این چالش ها و ارائه ساختواره ای درختی از انواع مسائل مطرح در این سه گروه با نگاهی به پژوهش های صورت گرفته در زمینه زبان فارسی و عربی به ارائه راهکار هایی جهت حل این مسائل بپردازد.

کلیدواژه ها: چالش های صرفی زبان فارسی؛ چالش های رسم الخط زبان فارسی؛ چالش های معنایی زبان فارسی؛ بازیابی اطلاعات

آسیب شناسی زبان و خط فارسی در بازیابی اطلاعات: نگاهی به موتور های کاوش و پایگاه های برخط

اشاره

شعله ارسطو پور (1) | فاطمه احمدی نسب (2)

درآمدی بر مشخصه های زبان فارسی

از دیدگاه دستور زبان زایشی (3)، زبان یا دانش زبانی قوه مستقلی از قوای ذهنی به شمار می رود که از دیگر قوای شناختی انسان مستقل بوده و خود نیز دارای بخش های مستقل نحو، معنا، واژگان و بخش واجی است به عبارت دیگر این رویکرد یک رویکرد حوزه ای (4) به زبان است که بین دانش زبانی و کاربرد آن تمایز قائل است (دبیر مقدم 1383، 18-19). در همین راستا در این نوشتار نیز منظور از زبان فارسی دانش زبانی فارسی زبانان ایرانی و هم چنین تبلور این دانش زبانی در قالب گونه نوشتاری فارسی رایج در ایران یعنی خط فارسی است. زبان فارسی یکی از زبان های هندو اروپایی و از شاخه زبان های ایرانی جنوب غربی است که زبان رسمی ایران و تاجیکستان و یکی از دو زبان رسمی افغانستان است (کامری 1990، 13-16) این در حالی است که تفاوت هایی میان این انشقاق ها وجود دارد مثلا فارسی تاجیکی و فارسی رسمی ایرانی از لحاظ دستوری یکسان هستند

1- دکترای علم اطلاعات و دانش شناسی مرکز منطقه ای اطلاع رسانی علوم و فناوری arastoopoor@ricest.ac.ir 2

2- دکترای زبان شناسی همگانی مرکز منطقه ای اطلاع رسانی علوم و فناوری ahmadinasabricest.ac.ir

3- Generative grammar

4- Modular

و فقـط مقـداری تفـاوت های واژگانی دارنـد، اما از لحاظ نظام نوشـتاری کاملاً متفاوت هسـتند چرا که خط فارسـی ایرانی از الفبای عربی اقتباس شده است و خط فارسـی تاجیکی الفبای سـیریلیکی دارد. از دیدگاه رده شناسـی (1) ، زبـان فارسـی از برخی ویژگی های سه رده زبانی زبان های تحلیلی (2)، زبان های تصریفی (3) و زبان های پیوندی (4) برخوردارست. این تمایز سه گانه بر اساس ساختار واژگانی زبـان تعریف می شود و از آن جا که ذخیره و بازیابی اطلاعات و نمایه سازی غالباً بر اساس واژه صورت می گیرد به نظر می رسـد که توجه به این تمـایز از اهمیت ویژه ای برخوردار باشـد. زبان تحلیلی زبانی است که واژه های آن بر اساس نقش (فاعلی، مفعولی و...) هیـچ گونه تغییری نکند و صـرفاً جایگاه آن در جمله نشان دهنده نقش واژه باشد کریسـتال (2008 ، 256) در زبان های ،تصـریفی واژه غالباً از چند تکواژ (5) تشـکیل می شود که نشان دهنـده روابط دسـتوری است در این زبان ها جایگاه واژه در جمله نشان دهنده نقش واژه نیست بلکه تصریف و حالت (6) واژه نقش واژه را به تصویر می کشـد (همان، 244) و بالاخره زبان پیونـدی زبانی است که واژه های آن غالباً از بیش از یک تکواژ تشـکیل شده اند و هر کدام از تکواژها نشان دهنده یک نقش یا رابطه دسـتوری است (همان، 16). البته لازم به ذکر است که تمایز زبان تصـریفی و پیوندی در آن است که تقطیع و تعیین تکواژهای واژه در آن امکان پذیر است و به راحتی می توان مرز آن ها را تعیین نمود اما در زبان های تصـریفی واژه غالباً صـرف شده و تکواژها نه به صورت منظم و به ترتیب بلکه به شـکل تجمعی در یک صورت کلمه (7) ظاهر می شونـد صورت های مفرد و جمع واژگان روایت /روایات عاقبت/ عواقب مـدرسه/ مدارس و ده ها مثال دیگر که البته وام واژه هایی (8) از زبان عربی هستند نشان دهنده وجود تصریف در زبان فارسی .است از دیگر سو صد ها مثال مانند واژه دانشگاهی (دان + ِش + گاه +ی) نشان گر وجود ویژگی زبان پیوندی در فارسـی است اگر چه تحقیق جامعی در تمایل بیش تر زبان فارسـی به هر یک از این سـه رده صورت نگرفته اما به نظر می رسد که زبان فارسـی به زبان تحلیلی نزدیک تر بوده و از ترتیب واژه برای نمایش نقش های دسـتوری بیش از شـیوه هـای دیگر بهره می برد. از بحث های مرتبط با بحث توالی اجزای واژگانی می توان به بحث هسـته (9) واژه مرکب اشاره نمود در زبان فارسـی واژه مرکب از نظر معنایی و بر اساس هسته واژگانی خود به چهار دسته برون ،مرکز درون مرکز (هسته آغازی، هسته پایانی) دو سویه و متوازن تقسیم می شود. در واژه مرکب برون مرکز هیچ کدام از اجزای واژه هسته را تشکیل نمی دهند برای مثال «آب سیاه» نام یک بیمـاری است و نه آب است و نه سـیاهرنگ در واژه مرکب درون مرکز یکی از اجزای واژه هسـته است و به دو نوع هسـته –آغازی و هسـته پایانی

1- Typology

2- Isolating/analytic

3- Inflectional

4- Agglutinating

5- Morpheme

6- Case

7- Word form

8- Loan word

9- Head

تقسیم می شود. برای مثال «موش خرما» هسته آغازی و «مویرگ» هسته پایانی است. در واژه مرکب دوسویه هر دو جزء به یک مرجع واحد اشاره می نماید مانند «سرباز - معلم» و بالاخره اینکه در واژه مرکب متوازن هر دو جزء به یک اندازه در ساخت کلمه مرکب نقش دارند؛ مانند «دخل و خرج» شقاقی (123، 1386 -124).

نکات پیش گفته تنها بخشی از مهم ترین ویژگی های زبان فارسی است که به صورت بالقوه قابلیت ایجاد بزرگ ترین چالش ها را در بحث ذخیره و بازیابی اطلاعات دارد عمده ترین دلیل چنین امری آن است که ماشین قابلیت تشخیص ترتیب و نقش دستوری را به خودی خود ندارد و همین امر منجر به بازیابی منابعی می شود که بعضاً مورد نیاز کاربر نبوده و عملاً پاسخی به نیاز وی نخواهند بود. بنابراین بی شک بعد زبان شناختی فرایند ذخیره و بازیابی اطلاعات یکی از مهم ترین و در عین حال چالش برانگیز ترین مسئله طراحی و کاربرد پذیری پایگاه ها است بازیابی اطلاعات با مسائل زبانی گره خورده و نمی توان بدون توجه به زبان مبدأ و ویژگی های آن نظام های بازیابی اطلاعات را در جهت افزایش جامعیت و مانعیت از یک سو و بهبود ربط از دیگر سو ارتقا بخشید. این نوشتار بر آن است تا از سه بعد رسم الخط صرف و معنا به بررسی مشکلات و چالش های زبان فارسی پرداخته و به صورت همزمان به نمونه هایی عینی از این مشکلات و نتیجه عدم توجه به آن ها در سطح وب و برخی از پایگاه های مقالات فارسی اشاره کرده و به آسیب شناسی زبان و خط فارسی در رابطه با این حوزه بپردازد.

پیشینه پژوهش

تلاش در جهت اصلاح و تقویت زبان و خط فارسی را می توان به تأسیس فرهنگستان اول نسبت داد. در این نوشتار به منظور پرهیز از طولانی شدن بحث صرفاً به برخی از تلاش های اخیر اشاره شده است. به عنوان ،نمونه آشوری (1375) برای اصلاح خط فارسی پیشنهاد می دهد که صورت های صرفی زمان حال فعل بودن یعنی «اَم»، «ای»، «است»، «ایم»، «اید»، «اند» و ضمایر متصل «اَم»، «ات»، «اَش»، «مان»، «تان»، «شان» جدا از کلمه های قبل از خود نوشته شوند معصومی همدانی (1381) در مقاله ای تحت عنوان خط فارسی و رایانه یکسان سازی رسم الخط فارسی را برای استفاده از خطایاب امکان جستجوی واژه در متن و تهیه نمایه ضروری می داند وی راه حل را در تهیه یک دستورالعمل واحد برای خط فارسی دانسته و با اشاره به دستورالعمل فرهنگستان می نویسد: «متأسفانه میزان آزادیی که این دستور خط به استفاده کنندگان داده به قدری است که می توان گفت کاربرد آن در مواردی بر تشتت موجود خواهد افزود». اسلامی (1381) اعمال اصلاحاتی را در خط فارسی ضروری می داند؛ اصلاحاتی از قبیل وضع یک نشانه اصلی برای نمایش کسره اضافه استفاده از نشانه های متفاوت برای نمایشی نکره و ی اسم-ساز و صفت ساز سرهم نویسی کلمات غیر بسیط، قرار دادن علامت های جداگانه برای واژه بست (1) های ربطی فعل بودن و ضمایر ملکی، و بالاخره نشان

1- Clitic

دادن اسـامی خاص محقق زاده و زارعیان (1383) بـا الگوگیری از زبان های لاتین چپ نویسـی کاهش حروف از طریق جدانویسـی قائل شـدن به دو شـکل کوچک و بزرگ حروف را در مورد زبان فارسـی پیشـنهاد می نمایند. علاوه بر این نگارش حروفی که خوانده شده ولی نوشـته نمی شونـد، ایجاد علامتی برای کسـره اضافه و عـدم تمایز بین «آ» و «ا» را نیز در جهت سازگار ساختن زبان فارسـی با محیط های رایانه ای ضروری می دانند اسلامی (1386) نقطه دار بودن برخی از نویسه های فارسـی را یکی از ایرادهای خط فارسی می داند؛ چرا که علاـوه بر دشوار نمودن املای ،کلمات این ایراد اساسـی را دارد که اسـتفاده از برنامه نشانه خوان نوری (1) را برای متون فارسی دشوار می نماید. صفار مقدم (1386) فاصـله گذاری را در ترکیبات فارسـی ضروری دانسته و مفصلاً به مبحث فاصله گذاری درون کلمه ای و برون کلمه ای در انواع کلمات فارسـی پرداخته است. گل تاجی و بـذرگر (1389) با اسـتفاده از یک سیاهه 17 کلید واژه ای مشکلات ریخت شناسـی خـط فارسـی را در سه پایگاه اطلاعاتی مرکز منطقه ای اطلاع رسانی علوم و فناوری پژوهشـگاه اطلاعات و مـدارک علمی ایران و جهاد دانشگاهی مقایسه کرده و نتیجه می گیرند که هیچ کدام از پایگاه های مذکور به ور جامع به این مسائل توجه نداشته اند.

همان طور که از مرور آثار پیشـین پیداست برخی از پژوهش ها به ضـرورت اصلاح خط فارسی با توجه به محیط های رایانه ای و برخی به آموزش زبان فارسـی توجه داشـته اند که برخی از پیشـنهادات مطرح شده در این راسـتا افراطی بوده و در صورت اجرا چهره خط فارسی را کاملاً دگرگون می نمایند برای مثال چپ نویسـی اسـتفاده از حروف بزرگ و کوچک و یا نگارش صداهایی که خوانده می شود. خط دارای ماهیتی محافظه کارانه بوده و به راحتی تغییر را نمی پـذیرد. علاوه بر این تغییر آن به راحتی امکان پـذیر نیست و به بسـیج امکانات مادی و معنوی عظیمی نیاز دارد به نظر می رسـد که در راسـتای حفظ گنجینه زبان فارسـی باید خط فارسـی را به همین صورت حفظ کرد و تنها با اعمال برخی اصلاحات و پیروی از دستور العملی واحد در جهت یکسان سازی خط فارسی از یک سو و کسب توانمندی های بیش تر در حوزه پردازش زبان طبیعی از دیگر سو به سمت رفع مشکلات بازیابی اطلاعات در سطح وب حرکت کرد.

رسم الخط فارسی و بازیابی اطلاعات

در مورد مسائل مرتبط به خط فارسـی به طور مطلق و همچنین ارتباط آن با بازیابی اطلاعات آثار متعـددی به نگارش درآمـده است که هر کدام از آن ها به گوشه ای از این مشکل عمده توجه داشته است در این نوشتار تنها به دو مسئله فاصله گذاری و نگارش «الف» پرداخته می شود یکی از راه حل های ارائه شـده برای یکسان سازی خط فارسـی و بهبود بازیابی اطلاعات فارسـی در وب پیشـنهاد عدم تمایز صورت هـای مختلف الف آغـازی است یعنی «آ» و «ا» به یک صورت یعنی «ا» نوشـته شود (اسـلامی، 1381). اما این راهکار نمی توانـد مفید باشـد و به ابهـام معنایی منجر می شود. برای مثال در پایگاه RICeST بین «آ» و «ا» تمایزی وجود نـدارد و باعث می شود که جسـتجوی کلید واژه های آسم

1- OCR =Optical character reader

(بیماری تنفسی) و اسم (مقوله دستوری) در موتور جستجوی جامع به نتایج کاملاً یکسان منجر شود و حال آن که منطقاً این دو کلیدواژه باید به بازیابی نتایج کاملاً متفاوتی منجر شود. به عبارت بهتر، قائل شدن این تمایز سطح ربط نتایج بازیابی شده به میزان چشم گیری افزایش می یابد. مثلاً در پایگاه مگ ایران این تمایز در نظر گرفته شده است و جستجوی کلیدواژه ها به نتایج متفاوت و مرتبط به کلیدواژه موردنظر می انجامد. (1)

چالش عمده دیگر در پیوند با رسم الخط فارسی متوجه نحوه فاصله گذاری است. همان طور که می دانیم اکثر نویسه (2) های فارسی با توجه به جایگاه نویسه در واژه به صورت منفصل یا متصل نگاشته می شوند و این خود امر آموزش و همچنین ذخیره و بازیابی اطلاعات را دشوار می سازد. فرهنگستان زبان و ادب فارسی دستورالعملی را تحت عنوان دستور خط فارسی تدوین کرده است تا به یکسان سازی و بهبود نگارش فارسی منجر شود یکی از مباحث این دستورالعمل به فاصله گذاری مربوط است. طبق این دستورالعمل در خط فارسی از دو فاصله یعنی نیم فاصله (درون کلمه) و فاصله کامل (برون کلمه) استفاده می شود رعایت فاصله کامل و نیم فاصله در پرهیز از بدخوانی بسیار راه گشاست (فرهنگستان زبان و ادب فارسی 1389 10). برای مثال در مورد نام و نام خانوادگی ایرانیان تنها فاصله گذاری دقیق است که باعث خوانایی و رفع ابهام می شود. زنجیره «علی رضا خانی» دو خوانش محتمل «علیرضا خانی» و «علیِ رضاخانی» دارد که تنها فاصله گذاری صحیح می تواند خوانش درست را تعیین نماید صفار مقدم 1386، 125) در بازیابی اطلاعات فاصله گذاری مخصوصاً در مورد واژه های مشتق و مرکب نقش تعیین کننده ای دارد چرا که پیوسته نویسی جدا نویسی کلید واژه های جستجو نتایج متفاوتی را بدست می دهد به عنوان نمونه پایگاه مگ ایران قابلیت جستجوی واژگان دو قسمتی را از طریق استفاده از نقطه فراهم آورده است اما کاربر با جستجوی واژه «خاکبرداری» در سه حالت «خاک. برداری»، «خاکبرداری» و «خاک برداری» به نتایج متفاوتی دست خواهد یافت با جستجوی واژه «خاک.برداری» 5 یافته بازیابی می شود در حالی که با جستجوی «خاک برداری» نتایج جستجو به 16 صفحه نیز می رسد. در میان این 16 صفحه یک نتیجه در برگیرنده واژه «خاک برداری» است در حالی که در میان نتایج 5 یافته اول نبوده و از 5 یافته اول حاصل از جستجوی «خاک.برداری» مورد در نتایج حاصل از جستجوی «خاک برداری» مشاهده نشد. این در حالی است که نتایج حاصل از جستجوی «خاکبرداری» 7 مورد بوده که کاملا متفاوت از دو جستجوی قبلی است. در واقع با جستجوی «خاک برداری» آن دسته از نتایج مرتبط در سایر جستجو ها به عنوان مرتبط ترین شناسایی نشده و در صفحات ،2 4 و 8 پراکنده اند. در پایگاه RCeST نیز جستجوی کلیدواژه «خاک+ برداری» 496 ، «خاکبرداری» 18 و عبارت «خاک برداری» 9 نتیجه را بازیابی می نماید بررسی نتایج جستجوهای «خاکبرداری و «خاک برداری» و مقایسه آن ها با یکدیگر نشان داد که تنها یکی از عنوان ها مشترک بوده است. این بدان معنا است که در هر دو پایگاه

1- البته در نتایج بازیابی شده برای این عبارت ،جستجو هم چنان مسائلی همچون عدم رعایت درست تقطیع به چشم می خورد

2- Letter

جستجو با نگارش های متفاوت به ریزش کاذب منتهی می شود. بدیهی است تعداد زیاد بازیافت ها و کاهش دقت در 10 نتیجه اول، یافتن نتایج مرتبط را برای کاربر زمان بر و دشوار خواهد کرد. باید توجه نمود که واژه هایی مانند خاک برداری کتاب ،شناسی خودکشی و غیره از طریق فرایند ترکیب و اشتقاق ناپایگانی ساخته شده اند (شقاقی 1386 ، 99) و اجزای دوم آن ها یعنی «-برداری»، «-شناسی» به طور مستقل بکار نرفته و استقلال واژگانی و معنایی ندارند بنابراین نظام های بازیابی اطلاعات باید راهکاری را برای جستجو در نظر گیرند تا این اجزا بصورت مجزا جستجو و بازیابی نشده و در ترکیبات مورد نظر به صورت یک واحد جستجو شوند.

مسائل صرفی و بازیابی اطلاعات

صرف (1) یکی از شاخه های زبان شناسی است که به مطالعه واژه و ساختمان درونی آن ها می پردازد. در این نوشتار به دو موضوع صرفی مرتبط به بازیابی اطلاعات یعنی واژه مرکب و وام گیری واژگانی پرداخته می شود. همان طور که پیش تر گفته شد واژه مرکب در فارسی چهار نوع مختلف برون مرکز، درون مرکز، دوسویه و متوازن دارد ویژگی های خط فارسی در اغلب موارد باعث می شود که نتوان از لحاظ نوشتاری تمایزی بین واژه مرکب و گروه نحوی قائل شد. در نتیجه نظام های بازیابی اطلاعات نیز نمی توانند واژه مرکب را از گروه نحوی تشخیص دهند. برای مثال واژه «زیست شیمی» (2) یک واژه مرکب درون مرکز هسته پایانی است که «شیمی» هسته آن را تشکیل می دهد. در حالی که نظام بازیابی اطلاعات این واژه مرکب را هم گروه نحوی «زیستِ شیمی» و هم «زیست+ شیمی» قلمداد می.کند. جستجوی این واژه مرکب در گوگل 11800000 بازیافت در پی دارد که بررسی 10 نتیجه اول نشان می دهد که نظام بازیابی هر سه حالت را مد نظر داشته است. این در حالی است که تنها 50000 بازیافت مرتبط محسوب می شود. لازم به ذکر است که در واژه مرکب «زیست شیمی»، «زیست» پیشوند است و نه واژه مستقل و «شیمی» هسته است در حالی که در واژه «محیط زیست» «زیست» واژه ای مستقل به شمار می رود و «محیط» هسته است. در واقع «زیست» در دو مثال فوق الذکر از دو مقوله و نقش متفاوت برخوردار است که اگر نظام های بازیابی مجهز به امکاناتی برای تشخیص و تعیین نوع آن ها بودند، مسلماً بازیابی بهبود چشم گیری می یافت. از دیگر چالش های صرفی بازیابی اطلاعات وام گیری واژگانی است. همان طور که می دانیم زبان ها از شیوه های مختلفی برای افزایش و غنی سازی واژگان خود بهره می برند که یکی از آن ها وام گیری است. در وام گیری ،زبانی زبان مقصد عناصر زبانی را از زبان مبدأ به صورت های مختلفی مانند وام گیری ،مستقیم غیر مستقیم ترجمه قرضی تعبیر ،قرضی تغییر قرضی، ترجمه و تعبیر ،قرضی آمیزش قرضی و تبادل قرضی وام می گیرد یکی از معضلات بازیابی اطلاعات خصوصاً در حوزه های تخصصی و علمی حاصل از وام گیری مستقیم است. در وام گیری مستقیم واژه ای از زبان

1- Morphology

2- Biochemistry

الف مستقیماً وارد زبان ب می شود این واژه پس از ورود به زبان ب بر طبق قواعد آوایی زبان مقصد تغییراتی پیدا می کند تا تلفظ آن برای گویشوران آسان شود (شقاقی 1386، 127-131). با توجه به این که در حوزه های تخصصی سرعت استفاده و جذب علم و یا فناوری بالاتر از سرعت واژه گزینی در زبان مقصد می باشد، لذا بسیاری از واژگان به همان صورت زبان مبدأ در زبان مقصد آوانگاری می شوند. بدیهی است چنین فرایندی عمیقاً تحت تأثیر نحوه تلفظ متخصص زمینه موضوعی بوده به طوری که امکان دارد یک واژه با املاهای متفاوتی نوشته شود؛ برای مثال سندروم / سندرم کلسیم / کالسیم، ئیدروژن/ هیدروژن این چندگانه نویسی باعث می شود که نه تنها ذخیره بلکه بازیابی اطلاعات نیز با چالش های فراوانی رو به رو شود به عنوان نمونه جستجوی معادل های آوانگاری شده واژه Psychology در پایگاه تخصصی مجلات نور ادعای پیش گفته را تأیید می کند برای واژه «پسیکولوژی» 95 یافته «پسیکولوگی» 3 یافته و «سایکولوژی» 8 یافته به دست آمد. لازم به ذکر است این نگارش های متفاوت به ترتیب بر اساس تلفظ ،فرانسه آلمانی و انگلیسی وارد زبان فارسی شده است.

مسائل معنایی و بازیابی اطلاعات

زبان طبیعی از بُعد معنایی آکنده از ابهام است با وجود این کاربران زبان با استفاده از دانش زبانی و زمینه گفتمان در برخورد با این مسائل به ابهام زدایی می پردازند فلاحتی (1385) به طور کلی انواع ابهام های واژگانی را که می توانند به ابهام معنایی بی انجامند به 5 گروه عمده تقسیم می کند. ابهام مقوله ای ابهام حاصل از هم آوایی و هم نگاشتی ابهام چند معنایی و ابهام انتقالی ابهام مقوله ای در پیوند با معانی متفاوت یک واژه در بافت ها و نقش های مختلف همچون اسم، فعل، صفت و قید ایجاد می شود. در زبان فارسی این گونه از ابهام ها معمولا به واسطه گذاشتن علائم واکه های کوتاه تا حد زیادی برطرف می شوند؛ اما مشکل اساسی این جاست که در متون مخصوص بزرگسالان و افراد باسواد (از جمله متون تخصصی و نوشتارهای علمی) استفاده از این علائم و مشخص کردن آن ها مرسوم نیست بنابراین در بسیاری از مواقع این خواننده است که با مراجعه به متن و بستر واژه، کار ابهام زدایی را انجام می دهد لازم به ذکر است در پاره ای از موارد امکان دارد تفاوتی میان تلفظ دو واژه وجود نداشته باشد اما همچنان به واسطه متفاوت بودن نقش دستوری هر واژه در جمله معنای متفاوتی از واژه مورد نظر دریافت گردد. به عنوان نمونه می توان به واژه «بردار» اشاره کرد.

بَردار: وزنه بردار دیابتی! قوی ترین مرد جهان (ورزش)

بُردار: پیشرفت هایی در کنترل بردار رانش (ریاضیات)

بَردار: داستان بردار کردن حسنک وزیر (ادبیات)

چنان چه این واژه در بستر های مختلف جستجو شود با توجه به این که در متن اصلی استفاده از علائم نشان دهنده واکه های کوتاه مرسوم نیست و حتی در دو مورد اول و سوم نمایش واکه نیز تفاوتی در ابهام معنایی این تک واژه ایجاد نمی کند لذا در صورت جستجو، حداقل در سه گروه

متفاوت مقالات و مطالبی بازیابی می شود این در حالی است که فعل امر «بَردار» از این گروه حذف شده است و حال آن که در جستجو و بازیابی اطلاعات احتمال بازیابی گروهی از مطالب که دارای این واژه هستند نیز وجود دارد تصاویر 1 (الف و ب) شمایی از نتایج جستجو در دو پایگاه اطلاعاتی فارسی را به نمایش می گذارد همان گونه که در این تصویر مشاهده می شود الگوریتم های بازیابی اطلاعات در نظام های جستجو و بازیابی اطلاعات در شرایط معمول با استفاده از میزان حضور واژه در متن و این که واژه در کجای متن باشد به رتبه بندی نتایج جستجو می پردازند. چنین روشی در پیوند با حالت های معمول زبانی مشکلی ایجاد نمی کند اما در مواجهه با مشکلات زبانی می تواند منجر به بروز ریزش کاذب در نتایج شود. لازم به توضیح است دو نمونه ارائه شده در تصویر الف هر دو برگرفته از صفحه اول نمایش نتایج هستند که انتظار می رود مرتبط ترین نتایج را در اختیار کاربر قرار دهند.

ابهام های حاصل از هم آوایی هم نگاشتی و چند معنایی گونه های دیگری از ابهام های زبان فارسی به شمار می آیند در این گروه از ابهام های واژگانی واژه ها یا تلفظ یکسانی دارند و یا به گونه یکسانی نوشته می شوند در نظام های بازیابی مبتنی بر جستجوی متنی ابهام های حاصل از هم نگاشتی مشکلاتی را برای جستجوگران اطلاعات به همراه دارد این مشکلات در پاره ای از اوقات با ابهام مقوله ای هم پوشانی پیدا می کنند یکی از نمونه های بارز این نوع از ابهام حداقل در پیوند با حوزه ای فیزیک واژه «گَرم» و «گِرَم» است (تصویر 2)

عکس

تصویر ۱(الف و ب). نتایج بازیابی شده از دو پایگاه مقالات فارسی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری و بانک اطلاعات نشریات کشور (مگ ایران). در هر دو مورد بازیابی بر مبنای الگوریم‌های معمول محاسبه ربط با توجه به حضور واژه در جمله انجام گرفته است.

تصویر 1 (الف و ب). نتایج بازیابی شده از دو پایگاه مقالات فارسی مرکز منطقه ای اطلاع رسانی علوم و فناوری و بانک اطلاعات نشریات کشور (مگ ایران). در هر دو مورد بازیابی بر مبنای الگوریم های معمول محاسبه ربط با توجه به حضور واژه در جمله انجام گرفته است

ص: 167

عکس

تصویر ۲. نمایی از نتایج جستجو در پایگاه مجلات تخصصی نور

در همین راستا در حوزه معنایی[1] چالش دیگری تحت عنوان هم‌معنایی مطرح می‌شود. هم‌معنایی به وضعیتی گفته می‌شـــود که دو یا چند واژه به یک مفهوم واحد اشاره داشته باشند (صفوی ۱۳۸۷، ۱۶۰). بایســـتی به این نکته توجه داشت که بســـیاری از علوم در ایران بومی نبوده و پژوهشگران و دانشـــجویان غالباً از طریق ترجمه آثار غربی با آنها آشـــنا می‌شـــوند. بنابراین، اصطلاح‌گزینی برای واژگان علمـــی توســـط یک مرجع واحد صورت نگرفته و گاه مترجمان و پژوهشـــگران با ســـلیقه شخصی دست به واژه‌گزینی می‌زنند. به همین دلیل با وجود اینکه یکی از مهم‌ترین عوامل موفقیت ارتباط علمی شـــفافیت واژگانی است، در بسیاری از زمینه‌های موضوعی برای بیان یک مفهوم واحد از چندین واژه متفاوت فارســـی استفاده می شـــود. برای مثال در حوزه زبان‌شناسی برای هر مفهوم به طور متوســـط ســـه واژه علمی فارســـی وجود دارد (احمدی نســـب ۱۳۹۰، ۲۷۶). به همین دلیل، همانگونه که پیشتر نیز بیان شد، بازیابی اطلاعات در بسیاری از نظام‌ها برپایه واژگان موجود در متن صورت می‌پذیرد، و از آنجا که در متون مختلف مفاهیم به صورت‌های مختلف ثبت می‌شوند، لذا با جســـتجوی یک صورت از مفهوم، کاربر امکان دسترسی به صورت‌های دیگر همان مفهوم که منطقاً مرتبط با نیاز اطلاعاتی وی خواهد بود را از دســـت می‌دهد. برای مثال برای مفهوم Ontology در فارسی سه معادل هستی‌شناسی، هستان‌شناسی و آنتولوژی بکار می‌رود که جستجوی هرکدام از این کلیدواژه‌ها تنها مدارکی را بازیابی می‌کند که صرفاً همان صورت واژگانی در آنها بکار رفته است در حالی‌که هر سه واژه به یک مفهوم اشاره دارد (جدول ۱).

1.Synonymy

تصویر 2. نمایی از نتایج جستجو در پایگاه مجلات تخصصی نور

در همین راستا در حوزه معنایی (1) چالش دیگری تحت عنوان هم معنایی مطرح می شود. هم معنایی به وضعیتی گفته می شود که دو یا چند واژه به یک مفهوم واحد اشاره داشـته باشند صفوی (1387، 160) بایستی به این نکته توجه داشت که بسـیاری از علوم در ایران بومی نبوده و پژوهش گران و دانشجویان غالباً از طریق ترجمه آثار غربی با آن ها آشنا می شوند بنابراین اصطلاح گزینی برای واژگان علمی توسط

یک مرجع واحـد صورت نگرفته و گاه مترجمان و پژوهش گران با سـلیقه شخصـی دست به واژه گزینی می زنند به همین دلیل با وجود این که یکی از مهم ترین عوامل موفقیت ارتباط علمی شـفافیت واژگانی است در بسیاری از زمینه های موضوعی برای بیان یک مفهوم واحد از چندین واژه متفاوت فارسـی اسـتفاده می شود. برای مثال در حوزه زبان شناسی برای هر مفهوم به طور متوسط سه واژه علمی فارسی وجود دارد (احمدی نسب 1390، 276). به همین دلیل همان گونه که پیش تر نیز بیان شد بازیابی اطلاعات در بسیاری از نظام ها بر پایه واژگان موجود در متن صورت می پـذیرد، و از آن جا که در متون مختلف مفاهیم به صورت های مختلف ثبت می شونـد، لـذا با جسـتجوی یک صورت از مفهوم کاربر امکان دسترسـی به صورت های دیگر همان مفهوم که منطقاً مرتبط با نیاز اطلاعاتی وی خواهد بود را از دست می دهـد. برای مثـال برای مفهوم Ontology در فارسـی سه معادل هستی شناسـی، هسـتان شناسـی و آنتولوژی بکار می رود که جسـتجوی هرکدام از این کلید واژه ها تنها مدارکی را بازیابی می کند که صـرفاً همان صورت واژگانی در آن ها بکار رفته است در حالی که هر سه واژه به یک مفهوم اشاره دارد (جدول 1)

1- Synonymy

عکس

جدول ۱. تفاوت بازیافت‌ها در پیوند با سه صورت واژگانی متفاوت از یک مفهوم واحد

	پایگاه مقالات فارسی مرکز منطقه ای (Ricest)	بانک اطلاعات نشریات کشور (مگ ایران)	پایگاه مجلات تخصصی نور
هسته شناسی	276	420	3877
هستان شناسی	8	7	69
انتولوژی	8	5	100

ابهـــام انتقالی در بحث ترجمه از یک زبان به زبان دیگر مطرح می‌شـــود. زمانی که واژه از زبان مبدأ به زبان مقصد آمده اما در زبان مقصد چندین برابرنهاده پیدا می‌کند. برابرنهاده‌هایی که می‌توانند معانی متفاوت داشـــته و یا معانی مشـــابه و نگاشت‌های متفاوتی داشـــته باشند. این دست از واژگان ابهام‌برانگیز زمانی ایجاد مشـــکل می‌کنند که ابزار جســـتجو و بازیابـــی اطلاعات در قالب یک نظام بازیابی اطلاعات بین‌زبانی نیز فعالیت کند. به عنوان نمونه ۴ واژه Transpiration (تعرق)، Sweating (تعرق)، Evapotranspiration (تعرق و تبخیر)، و Guttation (تعریق) هرکدام در زبان انگلیســـی به یک معنا به کار برده می‌شوند مثلا در زبان مبدأ Transpiration برای گیاهان و در حوزه کشاورزی، و Sweating در پیوند با انسان به کار می‌رود. این درحالی است که واژه‌های پیش‌گفته در زبان فارسی معادل نســـبتاً یکسانی دارند. هم اکنون گوگل تاحدی قابلیت جستجوی بین زبانی در سطح وب را فراهم آورده است. اما زمانی که معادل واژه Transpiration در صفحات فارسی جستجو شود نتایجی مشابه با تصویر ۳بدست خواهد آمد. این درحالی است که نتایج زمانی که واژه تعرق به فارسی برای بازیابی صفحات انگلیســـی به کار رود به مراتب از پراکندگی بیشتری برخوردار خواهد بود؛ چرا که واژه تعرق در زبان فارسی برای بیان سه مفهوم متفاوت در زبان انگلیسی انتخاب شده است.

تصویر ۳. نمایی از صفحه اول نتایج بازیابی شده بر اساس اولین برابرنهاده پیشنهادی گوگل برای واژه Transpiration

جدول 1 تفاوت بازیافت ها در پیوند با سه صورت واژگانی متفاوت از یک مفهوم واحد

ابهام انتقالی در بحث ترجمه از یک زبان به زبان دیگر مطرح می شود. زمانی که واژه از زبان مبـدأ به زبان مقصـد آمده اما در زبان مقصـد چندین برابر نهاده پیدا می کند برابر نهاده هایی که می توانند معانی متفاوت داشـته و یا معانی مشابه و نگاشته ای متفاوتی داشته باشند. این دست از واژگـان ابهـام برانگیز زمانی ایجاد مشـکل می کننـد که ابزار جسـتجو و بازیابی اطلاعات در قالب یک نظام بازیابی اطلاعات بین

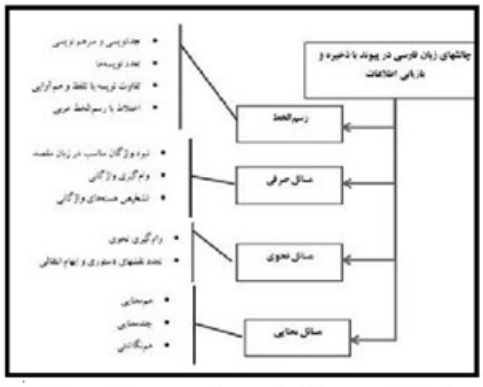
زبانی نیز فعالیت .کنـد به عنوان نمونه 4 واژه Transpiration (تعرق)، Sweating (تعرق)، Evapotranspiration (تعرق و تبخیر) و Guttation (تعریق) هر کـدام در زبان انگلیسـی به یک معنا به کار برده می شونـد مثلا در زبان مبـدأ Transpiration برای گیاهان و در حوزه کشاورزی، و Sweating در پیونـد بـا انسان به کار می رود این در حالی است که واژه های پیش گفته در زبان فارسـی معادل نسـبتاً یکسـانی .دارنـد هم اکنون گوگـل تـا حـدی قـابلیت جسـتجوی بین زبـانی در سـطح وب را فراهم آورده است. امـا زمـانی که معادل واژه Transpiration در صـفحات فارسی جستجو شود نتایجی مشابه با تصویر 3 بـدست خواهد آمد. این در حالی است که نتایج زمانی که واژه تعرق به فارسـی برای بازیابی صفحات انگلیسی به کار رو د به مراتب از پراکندگی بیش تری برخوردار خواهد بود؛ چرا که واژه تعرق در زبان فارسی برای بیان سه مفهوم متفاوت در زبان انگلیسی انتخاب شده است.

تصویر 3 نمایی از صفحه اول نتایج بازیابی شده بر اساس اولین برابر نهاده پیشنهادی گوگل برای واژه Transpiration

مسئله ایجاد یک نظام برای به هم نزدیک سازی ذهنیت پدیدآورنده و استفاده کننده از مسائل مهم در امر طراحی پایگاه های اطلاعاتی است. دسترسی مناسب به معنای ایجاد شرایط لازم از نظر زبان و واژگان اختصاص یافته به مدارک است؛ واژگانی که به کاربر در کاوش منابع یاری می رساند (مک ایوان 1379، 4) حال مشکل از این جا آغاز می شود که در پاره ای از پایگاه های اطلاعاتی، شخص سومی تحت عنوان نمایه ساز وارد شده و پاره ای از کلیدواژه ها را به مدارک با هدف بازیابی بهتر اختصاص می دهد در این فرایند با توجه به این که حلقه سومی در زنجیره ارتباطی میان پدیدآور و مخاطب ایجاد می شود اگر چه هدف بهبود بازیابی است اما گاه احتمال افزایش ضریب خطا اگر چه چه در ورود اطلاعات و چه در اختصاص واژگان کلیدی نیز وجود دارد (شاپوری 1379؛ Gross Taylor 2005) مسلماً دانش حوزه ای از طریق مطالعه حوزه ای حاصل می شود تحقیقات متعددی در تأیید این مطلب صورت گرفته است که وایلدرموث (Wildermuth، 2004) ضمن برشمردن برخی از آن ها تأکید می کند دانش حوزه ای و واژگانی افراد که عموماً برگرفته از متون تخصصی ،است نقش عمده ای در انتخاب واژگان جستجو دارد در چنین شرایطی وظیفه نمایه ساز به عنوان حلقه واسط نزدیک کردن دو زبان پدیدآور و مخاطب است خصوصاً در هنگامی که نظام بازیابی اطلاعات قابلیت برقراری این ارتباط را بصورت مؤثر ندارد با وجود ،این حجم زیاد متون ذخیره شده و کمبود نیروی انسانی توانمند متخصصان حوزه بازیابی اطلاعات را به سمت هرچه هوشمند ساختن بازیابی اطلاعات سوق داده است. تصویر 4 به منظور جمع بندی ساختواره ای درختی از اهم چالش های متصور در زمینه ذخیره و بازیابی اطلاعات را به نمایش می گذارد

عکس

مسئله ایجاد یک نظام برای به هم نزدیک‌سازی ذهنیت پدیدآورنده و استفاده‌کننده از مسائل مهم در امر طراحی پایگاه‌های اطلاعاتی است. دسترسی مناسب، به معنای ایجاد شرایط لازم از نظر زبان و واژگان اختصاص یافته به مدارک است؛ واژگانی که به کاربر در کاوش منابع یاری می‌رساند (مک‌ایوان ۱۳۷۹، ٤۹). حال، مشکل از این‌جا آغاز می‌شود که در پاره‌ای از پایگاه‌های اطلاعاتی، شخص سومی تحت عنوان نمایه‌ساز وارد شده و پاره‌ای از کلیدواژه‌ها را به مدارک با هدف بازیابی بهتر اختصاص می‌دهد. در این فرایند با توجه به اینکه حلقه سومی در زنجیره ارتباطی میان پدیدآور و مخاطب ایجاد می‌شود، اگرچه هدف بهبود بازیابی است، اما گاه احتمال افزایش ضریب خطا چه در ورود اطلاعات و چه در اختصاص واژگان کلیدی نیز وجود دارد (شاپوری ۱۳۷۹؛Gross & Taylor ۲۰۰۵). مسلماً دانش حوزه‌ای از طریق مطالعه حوزه‌ای حاصل می‌شود. تحقیقات متعددی در تأیید این مطلب صورت گرفته است که وایلدرموث (Wildermuth, ۲۰۰۴) ضمن برشمردن برخی از آنها، تأکید می‌کند دانش حوزه‌ای و واژگانی افراد که عموماً برگرفته از متون تخصصی است، نقش عمده‌ای در انتخاب واژگان جستجو دارد. در چنین شرایطی وظیفه نمایه‌ساز به عنوان حلقه واسط، نزدیک کردن دو زبان پدیدآور و مخاطب است. خصوصاً در هنگامی که نظام بازیابی اطلاعات قابلیت برقراری این ارتباط را بصورت مؤثر ندارد. با وجود این، حجم زیاد متون ذخیره شده و کمبود نیروی انسانی توانمند، متخصصان حوزه بازیابی اطلاعات را به سمت هرچه هوشمند ساختن بازیابی اطلاعات سوق داده است. تصویر ۴ به منظور جمع‌بندی ساختواره‌ای درختی از اهم چالش‌های متصور در زمینه ذخیره و بازیابی اطلاعات را به نمایش می‌گذارد.



تصویر ۴. ساختواره درختی چالش‌های عمده زبانی در پیوند با ذخیره و بازیابی اطلاعات[1]

1. لازم به توضیح است این ساختواره کامل‌تر از بحث‌های مطرح در این نوشتار است. دلیل این امر لزوم توجه به برخی از مهمترین موارد به منظور جلوگیری از طولانی شدن مبحث بوده است.

تصویر 4. ساختواره درختی چالش های عمده زبانی در پیوند با ذخیره و بازیابی اطلاعات (1)

ص: 170

1- لازم به توضیح است این ساختواره کامل تر از بحث های مطرح در این نوشتار است. دلیل این امر لزوم توجه به برخی از مهم ترین موارد به منظور جلوگیری از طولانی شدن مبحث بوده است

سخن پایانی: پیشنهادهایی در جهت بهبود وضعیت

راهبردهـای متفاوتی در بحث رفع این مسائل مطرح شـده است که می توان آن ها را به گروه های مختلف دسـته بنـدی نمود از میان آن ها عمـده ترین موارد شامـل تـدوین پیکره زبـان فارسـی یـادگیری ماشـینی ریشه یـاب هـا، نرمال سازی ،نویسه ها اسـتفاده از الگوریتم های احتمالاتی همچون N-gram ها و مهار واژگانی می شود برای رفع چالش های معنایی بازیابی اطلاعات به نظر می رسد که باید بر اساس یـک پیکره زبانی و اسـتخراج قواعـد واژه سازی زبان فارسـی و تعیین اجزای واژگانی همچون ونـدها و ریشه ها عمل کرد این کار از طریق یـادگیری ماشـینی از یکسو و ایجاد پیکره های زبان فارسـی به صورت الکترونیک امکان پـذیر است حرکت هایی در این زمینه در پیونـد با زبان فارسی صورت گرفته است اما همچنان در مراحل ابتدایی است در این راستا یکی از کارهای ارزشمند انجام شده پارس مورف متعلق به مواجی، اسلامی و وزیر نژاد (1390) است. این نرم افزار قابلیت تحلیل صـرفی زبان فارسی را داراست. در همین رابطه می توان به ریشه یاب تهیه شده توسط احسان و فیلی (1390) نیز اشاره کرد که میزان دقت در نظام های بازیابی را 5 درصد افزایش می دهد. از دیگر تلاش ها می توان به ریشه یاب برنجیان (1390) برای افعال ماضی و مضارع ناگذر اشاره کرد. لازم به یادآوری است که حرکت های مشابه برای زبـان عربی مـدت هاست آغاز شـده و به مرحله ارزیابی و بهسازی رسـیده است. به عنوان نمونه می توان به پژوهش های محمود ماجـد و خلدون (Mahmoud, Majed Kaldoun 2011) و یا انیس و دیگران (Anis etal. In Press) اشاره کرد.

از راهبرد های پیشنهادی در پیوند با برطرف سازی مشکلات مرتبط با تعدد نویسه هایی همچون الف و یا همزه در زبان عربی نرمال سازی است. در این راهبرد با استفاده از یک الگوریتم ساده، کلیه صورت های مختلف یک نویسه به طور خودکار به یک شکل یکسان تبدیل می شود به عنوان مثال کلیه نگارش های متفاوت «آ» و «ا» به صورت «1» در نظر گرفته می شود (Mahmoud, Majed Kaldoun 2011). اما با بررسی های صورت گرفته و همان گونه که پیش تر در این مقاله نیز مطرح گردید زبان فارس به واسطه ویژگی های معنایی خاص خود بر تابنده چنین تغییری، حداقل در پیوند با این مثال نمی باشد لذا نرمال سازی در پیوند با نویسه های زبان فارسـی بایستی با دقت و مطالعه بیش تری صورت گرفته و نمی توان به یافته های مطالعات مرتبط با زبان عربی در این زمینه اسـتناد کرد راهبرد دیگر در این راستا استفاده از ریشه یاب ها است اسـتفاده از ریشه یاب امکان بازیابی صورت های مختلف واژگانی که به یک حوزه مفهومی واحـد اشاره دارند را فراهم می کند. این در حالی است که این راهبرد می تواند به صورت بالقوه پاسخی به مشکلات نرم افزارهای بازیابی اطلاعات در پیوند با صورت های مفرد و جمع واژگان خصوصاً جمع های مکسر و یا واژه های هم خانواده باشد (Mahmoud, Majed Kaldoun 2011) با توجه به این که زبان فارسـی از نظر رسم الخط مشابه با زبان عربی است پیشـنهاد می شود در این زمینه از مطالعات صورت گرفته در پیونـد با خط عربی و بازیابی اطلاعات اسـتفاده شود در حالی که در زمینه ویژگی های صـرفی و معنایی زبان فارسـی به زبان انگلیسـی نیز نزدیک بوده و مطالعات صورت گرفته در این زمینه تا حدود زیادی

راهگشاست.

اگر چه در بسیاری از متون بحث رعایت فاصله گذاری مناسب به پدیدآور نسبت داده می شود اما هم چنان نیازمند نظام هایی با قابلیت تشخیص مناسب فاصله ها در واژگان مرکب (1) هستیم. در زبان فارسی یکی از شیوه های پر کاربرد واژه سازی استفاده از فرایند ترکیب است لذا تعداد واژگان ترکیبی در زبان فارسی بسیار زیاد است این در حالی است که در زبان فارسی به واسطه ماهیت زبانی و دستوری و نیز سهولت ،خواندن واژگان مرکب معمولا به صورت ناپیوسته همراه با یک فاصله یک حرفی ثبت می شوند. لازم به یادآوری نیست که در زبان انگلیسی این مشکل کمتر دیده می شود. به عنوان نمونه کلیه واژگانی که با «شناسی» همراه هستند. معمولا به صورت دو جزء جداگانه ثبت می شوند در حالی که در زبان انگلیسی "logy" بخشی از واژه به شمار می آید. بدیهی است رفع این مشکل تنها از طریق استفاده از عملگر And در یک فیلد جستجو امکان پذیر نمی باشد، چرا که منجر به افزایش تعداد بازیافت های غیر مرتبط می شود این در حالی است که چنان چه کاربری کلمات وارد شده در یک فیلد به شکل عبارتی تغییر یابد امکان بروز این خطا کمتر خواهد شد. اما با نگاهی به متون موجود در پیوند با رفع این دسته از چالش ها، حداقل در پیوند با زبان عربی استفاده از الگوریتم های N-gram پیشنهاد می.شود دلیل این امر قابلیت ارائه بافت واژه مورد جستجو همراه با واژه در قالب پیشنهادهایی برای بهسازی جستجوست (Mahmoud, Majed Kaldoun 2011).

مهار واژگانی یکی دیگر از راهبرد های مفید در زمینه بهسازی جستجو از طریق کم اثر سازی تعدد صورت های واژگانی و یا معنایی است بدین معنا که با ارائه صورت های مرجح، نامرجح و واژگان شامل و زیر شمول کلیه صورت های متفاوت یک مفهوم را ذیل یک مفهوم واژه پذیرفته شده گرد می آورد. واژگان مهار شده معمولا نمایان گر ساختواره ای از روابط معنایی و سلسله مراتبی بوده و استفاده از آن خصوصاً در برطرف ساختن چالش های هم معنایی، چندمعنایی، املاهای متعدد و وام گیری واژگانی کاراست (Svenonius 2003). این در حالی است که استفاده همزمان از اصطلاحنامه ها در کنار کلید واژه های زبان طبیعی یکی از کاراترین روش ها در این راستا به شمار می آید که از جمله پایگاه های موفق در این زمینه می توان به INSPEC ، Eric و LISA اشاره کرد.

در پایان خاطر نشان می سازد که تاکنون پیکره های فارسی مختلفی همچون همشهری محک بی جن خان و دادگان زبان فارسی تدوین شده است از این میان درودی و دیگران (1387) مجموعه محک وب را تهیه کرده اند که از آن می توان برای انجام مطالعات در حوزه بازیابی اطلاعات فارسی در وب استفاده نمود. این مجموعه dotIR نام دارد و استفاده از آن برای امور غیر تجاری رایگان است. مجموعه پیش گفته از یک پیکره استاندارد 50 پرس وجوی استاندارد 18 هزار داوری تعیین ربط پرس وجوها به اسناد پیکره و 50 هزار بردار ویژگی استخراج شده از اسناد، تشکیل شده و بیش تر مناسب پژوهش در رابطه با بازیابی اطلاعات است همشهری (2) مجموعه دیگری است که

1- در این جا برای سهولت بحث واژه مرکب به واژه غیر بسیط اطلاق شده و معنای دقیق زبان شناختی آن مدنظر نیست 2.

http://ece.ut.ac.ir/dbrg/hamshahri/faindex.html

2- http://ece.ut.ac.ir/dbrg/hamshahri/faindex.html

توسط رهگذر و دیگران بر اساس اخبار روزنامه همشهری تهیه شده و دارای دو نسخه 1 و 2 است که نسخه دوم از امکانات بیش تری مانند پیوند به تصاویر و به اصل صفحات وب برخوردار است. از این مجموعه نیز می توان برای انجام پژوهش در حوزه بازیابی اطلاعات ،فارسی، پردازش زبان طبیعی و تدوین الگوریتم های ریشه یابی استفاده نمود. بر اساس این مجموعه پژوهش های متعددی صورت گرفته که از جمله می توان به آل احمد و دیگران (2007) اشاره نمود که با استفاده از این مجموعه مدل فضای برداری بر اساس n-gram و کلمه را ارزیابی کرده و نشان داده اند که بازیابی متون فارسی بر اساس مدل فضای برداری gram-4 و طرح وزن دهی Inu ltu نتایج قابل قبول و تحلیل محلی متن (1) بهترین نتایج را در برخواهد داشت. از پیکره های زبانی دیگر می توان به پیکره بی جن خان (2) اشاره کرد که از متون خبری و متون عمومی شامل 4300 موضوع تشکیل شده و بر اساس 40 مقوله دستوری فارسی بر چسب دهی شده است و مناسب پژوهش های مرتبط به پردازش زبان طبیعی است. نمونه دیگر پایگاه دادگان فارسی (3) است که توسط عاصی و دیگران در پژوهش گاه علوم انسانی و مطالعات فرهنگی تهیه شده است این پایگاه گونه های مختلف زبانی نوشتاری و گفتاری را در بر می گیرد و در آن امکان جستجوی واژه ها ترکیب ها و بررسی بسامد آن ها به همراه گزارش های آماری از متون وجود دارد. در این جا ذکر این نکته خالی از لطف نیست که متون موجود و پژوهش های انجام شده نشان دهنده این امر است که علی رغم وجود این پیکره های زبانی ،فارسی، پژوهش های چندانی بر اساس این پیکره ها صورت نگرفته است شاید بتوان این امر را به عدم اطلاع اکثر پژوهش گران از اهمیت و نقش پیکره های زبانی در تحقیقات در حوزه پردازش زبان ،طبیعی بازیابی اطلاعات و تهیه ریشه یاب ها و همچنین عدم آموزش کافی جامعه علمی در این زمینه نسبت داد. در این نوشتار نیز با توجه به چارچوب زمانی در نظر گرفته شده برای بررسی ،حاضر متأسفانه محققان قادر به بررسی دقیق این پیکره ها و استفاده از آن ها در این بررسی نشدند. بنابراین لازم است تا پژوهشی جامع، میزان موفقیت راهبردهای پیشنهادی این بررسی بر پایه پیکره های پیش گفته را بسنجد. هم چنین توصیه می شود که در طراحی و به سازی پایگاه های اطلاعاتی زبان فارسی کارگروه هایی متشکل از متخصصان علم اطلاعات و دانش شناسی زبان شناسی و علوم رایانه نقش اصلی را ایفا نمایند.

منابع

آشوری، داریوش 1375 چند پیشنهاد دیگر برای اصلاح خط فارسی سرهم نویسی و جدا نویسی. نگاه نو. 28:101-117.

احمدی نسب، فاطمه . 1390 تهیه و تدوین اصطلاحنامه تک زبانه فارسی زبان شناسی رساله دکترا دانشگاه علامه طباطبایی.

اسلامی محرم 1381 دشواری های پردازش رایانه ای خط فارسی نشر دانش دوره 19(3) 28-32

1- Local context analysis

2- http://ece.ut.ac.ir/DBRG/Bijankhan

3- http://ece.ut.ac.ir/DBRG/Bijankhan

اسلامی، محرم 1386 خط فارسی و رسانه های گروهی دو فصل نامه پردازش علائم و داده ها. 8(2):93-98.

برنجیان شاپور رضا 13900. ریشه یاب ماضی و مضارع از مصدر افعال ناگذر در زبان فارسی. شیراز نوید شیراز مرکز منطقه ای اطلاع رسانی علوم و فناوری

دبیر مقدم، محمد. 1383. زبان شناسی نظری پیدایش و تکوین دستور زایشی (ویراست دوم) تهران انتشارات سمت.

درودی احسان و دیگران 1387 پیکره محک استاندارد برای تحقیقات بازیابی اطلاعات وب فارسی گزارش ،فنی گروه تحقیقاتی پایگاه داده ها دانشگاه تهران شماره: DBRG-TR-138702. http://ccc.ut.ac.ir/DBRG/webir/files/Papers/WebIR.pdf (دسترسی در 1391/10/19).

شاپوری، سودابه (1379) بررسی مشکلات جستجوی موضوعی استفاده کنندگان رایانه ای کتابخانه مرکزی دانشگاه فردوسی مشهد پایان نامه کارشناسی ارشد مشهد دانشکده علوم تربیتی و روانشناسی

شقاقی ویدا . 1386. مبانی صرف تهران: انتشارات سمت.

صفار مقدم، احمد. 1386 . فاصله گذاری در خط فارسی نامه فرهنگستان 9/4: 123-137.

صفوی ،کورش 1387. درآمدی بر معنی شناسی. تهران: انتشارات سوره مهر (حوزه هنری سازمان تبلیغات اسلامی)

فرهنگستان زبان و ادب فارسی 1389. دستور خط فارسی چاپ نهم تهران: فرهنگستان ادب و زبان فارسی (نشر آثار).

فلاحتی فومنی، محمدرضا 1385 ابهام در ماشین ترجمه کتابداری و اطلاع رسانی. 9 (3): 38 - 21

کامری ،برنارد جان ماونتفورد ویوین ،لا و د. آ. کروز 1990. زبان های دنیا چهار مقاله در زبان شناسی ترجمه کورش صفوی 1384 تهران: انتشارات سعاد.

گل تاجی، مرضیه و سعیده برزگر. 1389 بررسی مشکلات ریخت شناسی زبان فارسی در سه پایگاه اطلاعاتی مرکز منطقه ای اطلاع رسانی علوم و فناوری پژوهشگاه اطلاعات و مدارک علمی ایران و جهاد دانشگاهی کتابداری و اطلاع رسانی 13 (2): 191-214

مک ایوان اندرو .(1379) استفاده از سر عنوان های موضوعی کتابخانه کنگره: هزینه همکاری برای رسیدن به دسترس پذیری ترجمه مجتبی اسدی گزیده مقالات ایفلا 98 (آمستردام: 21 - 16 اوت 1998) (ص. 59 - 47) تهران کتابخانه ملی جمهوری اسلامی ایران.

محقق زاده محمد صادق و کاظم زارعیان 1383 ارائه راه حل برای برخی مسائل اتوماسیون و نگارش فارسی. فصلنامه اطلاع- رسانی 19 (3 و 4): 1-10.

معصومی همدانی حسین 1381 خط فارسی و رایانه نشر دانش. 19 (2) 2-6

مواجی وحید محرم اسلامی و بهرام وزیر نژاد 1390. پارس مورف: تحلیل گر صرفی زبان

فارسی. دو فصلنامه پردازش علائم و داده ها 15 (1)

Aleahmad, Abolfazl. et al. 2007. N-Gram and Local Context Analysis for Persian Text Retrieval, International Symposium on Signal Processing and Its Applications, Sharjah U.A.E. Retrieved 2013-01-8 From http://ece.ut.ac.ir/dbrg/hamshahri/files/ Papers/isspa.pdf

Anis, Z. et al. In Press. Contribution to Semantic Analysis of Arabic language. Advances in Artificial /Intelligence. Retrieved 2012-11-10 From http://www.hindawi.com/ journals/aai/aip/620461

.Crystal, David. 2008. A dictionary of linguistics and phonetics. Six Edition. Blackwell Publishing

Gross, Tina, Arlene G. Taylor. 2000. What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results. College Research Libraries. 66(3): 212 -230. Retrieved August 2, 2006, From www.ala.org/ala/acrl/acrlpubs/crljournal/ backissues2005a/crlmay05/Gross.pdf

Mahmoud, R., S. Majed, and Z. Kaldoun. 2011. Improving Arabic information retrieval system using N-.gram method. WSEAS Transactions on Computers 10 (4): 125-133

Svenonius, E. 2003. Design of Controlled Vocabularies. Encyclopedia of Library and Information Science) .(822-838). New York: Marcel Dekker (reprinted from the first edition, 1989

Wildermuth, Barbara M. 2004. The Effects of Domain Knowledge on Search Tactic Formulation. Journal of .the American Society for Information Science and Technology. 55(3): 246-258

چکیده

هدف: هدف اصلی از انجام این پژوهش ارزیابی کاربرد پذیری وب سایت نهاد کتابخانه های عمومی کشور است.

روش شناسی: پژوهش حاضر که از نوع کاربردی است، با استفاده از شیوه ارزیابانه به بررسی و ارزیابی معیارها و مولفه های مطرح در کاربرد پذیری وب سایت نهاد کتابخانه های عمومی کشور می پردازد. ابزار گردآوری اطلاعات در این پژوهش سیاهه ارزیابی محقق ساخته مشتمل بر 11 معیار و 160 مؤلفه است که وب سایت مورد نظر با آن سنجیده شد. در سیاهه مزبور از روش روایی صوری و محتوایی استفاده شده منظور تجزیه و تحلیل یافته های پژوهش از آمار توصیفی فراوانی درصد و میانگین برای به توصیف کشیدن وضیعت موجود وب سایت مورد بررسی استفاده شد.

یافته ها: در مجموع نتایج به دست آمده نشان داد در وب سایت نهاد کتابخانه های عمومی کشور نیمی از استانداردها 346 امتیاز (52/2 درصد) از 663 امتیاز وب سایت شاخص رعایت گردیده است.

کلید واژه ها: ارزیابی، کاربرد پذیری وب سایت، نهاد کتابخانه های عمومی کشور

ارزیابی کاربرد پذیری وبگاه نهاد کتابخانه های عمومی کشور

اشاره

صدیقه محمد اسماعیل (1) | ماهرخ ناصحی اسکویی (2)

مقدمه

امروزه از وب به عنوان یکی از مهم ترین و اصلی ترین ابزار ها برای دسترسی به اطلاعات استفاده می کنند لذا ضروری است، کاربران وب میزان کاربرد پذیری منابع موجود بر روی آن (به ویژه وب سایت های آن) را مورد بررسی و ارزیابی مستمر قرار دهند در این راستا پژوهش حاضر بر آن است تا وب سایت نهاد کتابخانه های عمومی کشور را مورد ارزیابی قرار دهد تا از این طریق مشخص گردد وب سایت نهاد به لحاظ کاربردپذیر بودن در چه شرایطی قرار دارد و تا چه حد در طراحی آن اصول و ضوابط طراحی وب سایت به لحاظ رعایت مؤلفه های مربوط به این امر (اعم از: اعتبار ا، صحت، روزآمد بودن، سطح پوشش و مخاطبان خاص، وجود نماهای تعاملی و تبادلی عینیت ،اطلاعات قابلیت ناوبری، نماهای غیر متنی، دسترس پذیری ، کارآمدی ، ویژگی های ظاهری رعایت گردیده است. بدیهی است، پرداختن به وب سایت نهاد کتابخانه های عمومی کشور با عنایت به جایگاه محوری آن در بحث اطلاعات و اطلاع رسانی امری حائز ارزش و توجه .است

1- استادیار گروه کتابداری و اطلاع رسانی دانشگاه آزاد اسلامی واحد علوم تحقیقات تهران M.esmaeili2@gmail.com

2- دانشجوی کارشناسی ارشد کتابداری و اطلاع رسانی دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران mahrokh.nasehi@gmail.com

2-روش پژوهش و توجیه روایی آن

روش پژوهش حاضر، روش کتابخانه ای (سندی) و پیمایشی از نوع ارزیابانه است. بدین معنا که، با استفاده از سیاهه کنترل یا به ارزیابی کاربردپذیری وب سایت نهاد کتابخانه های عمومی کشور (بر مبنای معیارهای متعدد موجود در چک لیست) پرداخته شده است. در توجیه روایی استفاده از چنین روشی جهت انجام این پژوهش ذکر این نکته ضروریست که از آن جایی که شاخص های بیرونی متعددی در قالب سیاهه وارسی بوده که چگونگی کاربردپذیری صفحات وب نهاد کتابخانه های عمومی کشور با آن ها سنجیده شده است لذا مناسب ترین راه برای آگاهی از وضعیت این وب سایت ها (بر اساس معیارها و مؤلفه های ذکر شده) روش ارزیابانه فوق بوده است.

3-شیوه گردآوری اطلاعات و تجزیه و تحلیل آن ها

در این پژوهش گردآوری اطلاعات با استفاده از روش مشاهده مستقیم وب سایت و بر اساس سیاهه ای با 11 معیار و 160 مؤلفه صورت گرفت و بر این مبنا وب سایت نهاد کتابخانه های عمومی کشور از نقطه نظر کاربرد پذیری و معیارهای مطرح در آن ارزیابی گردید. برای این منظور، تلاش شد تا ابتدا نسبت به تهیه سیاهه مربوطه تحت عنوان سیاهه ارزیابی کاربردپذیری وب سایت (1) اقدام گردد. از این رو، متون و منابع و پیشینه های موجود در زمینه کاربردپذیری وب سایت ها در داخل و خارج کشور مورد مطالعه و تفحص قرار گرفت و به منظور اطمینان از روایی محتوایی هر چه بیش تر در اختیار بررسی 5 تن از مدیران وب سایت های کتابخانه ای قرار گرفت سپس نظرات آن ها ،بررسی و نسبت به تهیه نسخه نهایی سیاهه (مشتمل بر 11 معیار کلی و 160 مؤلفه) اقدام شد و با استفاده از آن وضعیت وب سایت نهاد کتابخانه های کشور در طول مدت زمان انجام این پژوهش سه ماهه اول (1391)، مورد بررسی و ارزیابی قرار گرفت. به منظور تجزیه و تحلیل یافته های پژوهش از آمار توصیفی (فراوانی درصد و میانگین) برای به توصیف کشیدن وضعیت موجود استفاده شد. در سیاهه یاد شده از دو مقیاس وجود و نبود بلی(ü) و خیر (-) استفاده شده که امتیازات در نظر گرفته شده برای آن ها به ترتیب عبارتند از بلی = 1 (یک) و خیر = 0 (صفر). (نکته: اعداد مندرج در قسمت تواتر در جدول به شماره منابع ارجاع خورده است).

4-تجزیه و تحلیل داده ها و ارائه یافته ها

همان گونه که قبلاـ نیز گفته شد در این پژوهش از یک وب سایت مفروض استفاده شده با این فرض که در طراحی آن هر 160 مؤلفه یا ویژگی به گونه ای رعایت شده است که می توان آن را به مثابه شاخصی برای ارزیابی و سنجش دیگر وب سایت جامعه مورد مطالعه به کار برد. یافته های پژوهش بیان می کند که در وب سایت شاخص امتیاز کل رعایت همه معیارهای مطرح در کاربرد پذیری سایت برابر با 663 امتیاز است که از این امتیاز 107 امتیاز مربوط به رعایت مؤلفه های اعتبار اطلاعات (16 درصد)

1- Website Usability Checklist

32 امتیـاز مربوط به صـحت اطلاعات (5 درصـد) ، 26 امتیـاز مربوط به روزآمـد بودن (4 درصد)، 21 امتیاز مربوط به سـطح پوشـش و مخاطبان خاص (3 درصد31) امتیـاز مربوط به نماهـای تعاملی و تبادلی (5 درصد)، 34 امتیـاز مربوط به عینیت اطلاعات (5 درصد)، 166 امتیاز مربوط به رعایت معیار کلی ناوبری در سایت (جمعا: 25 درصد)، به تفکیک شامل: 9 امتیاز مرتبط با ویژگیهای عنوان مرورگر (1 درصد) 22 امتیـاز مرتبط با ویژگیهای عنوان صـفحه (3 درصد) 92 امتیـاز مرتبط با پیونـدهای متنی و فرامتنی (14 درصد)، 6 امتیاز مرتبـط با نشانه اینترنتی (1 درصد)، 25 امتیـاز مرتبـط با نقشه سایت یا نمایه (4 درصد) 12 امتیاز مرتبط با موتور جسـتجوی داخلی (2 درصد) ، 64 امتیاز مربوط به نماهای غیرمتنی (10 درصد)، 77 امتیاز مربوط به قابلیت دسترس پذیری (12 درصد)، 99 امتیاز مربوط به کارآمد پذیری (15درصد) و 6 امتیاز مربوط به ویژگی های ظاهری (1 درصد) است

همان گونه که از جـدول 1 مسـتفاد می شود یـافته هـای تحقیق نشـان می دهـد که وب سـایت نهاد کتابخانه های عمومی کشور از لحاظ رعـایت معیارهای «اعتبار اطلاعات» 45 امتیاز (42 درصـد)، «صـحت اطلاعات» امتیاز (53 درصـد)، «روزآمد بودن» 7 امتیاز (26 درصد)، «سـطح پوشش و مخاطبین خاص» 10 امتیاز (47 درصـد) «نماهای تعاملی و تبادلی» دارای 12 امتیاز (38 درصد)، «عینیت اطلاعات» 15 امتیاز (44 درصد) از لحاظ معیارهای کلی «ناوبری» 103 امتیاز (62 درصد) «نماهای غیر متنی» 18 امتیاز (28 درصد) «دسترس پذیری» 48 امتیاز (62 درصـد)، «کارآمدی» 51 امتیاز (51 درصـد) و در نهـایت در رابطه بـا آخرین معیار مورد بررسـی که ویژگی های «ظاهری» 6 امتیاز (1000 درصد) را به طور کامل به خود اختصاص داده است. یافته های این پژوهش با عصاره که به بررسی وب سایت های 58 کتابخانه ملی جهان با هـدف شـناسایی ویژگی های شاخص در طراحی وب سایت یک کتابخانه ملی به روش تحلیل محتوا و ارزیابی بر مبنای سـیاهه وارسی (وب سایت کتابخانه ملی آمریکا 80 درصد، وب سایت کتابخانه بیمارستانی انگلستان 28 درصد و وب سـایت کتابخانه ملی کشورهای مالزی و دانمارک 72 درصد با سـیاهه وارسی و وب سایت کتابخانه ملی ایران 51 درصد با سیاهه همخوانی داشـتند) و نیز با یافته های تحقیق حاجی زین العابدینی که به بررسـی وضعیت وب سایت های کتابخانه های ملی جهان از نظر میزان مورد اسـتفاده قرار گرفتن از طریق پیوندهای آن ها می پردازد همسو است. و در بعد دسترس پذیری نیز با یافته های محمد اسماعیل و کاظمی در تحقیقی با عنوان دسترس پذیری وب سایت کتابخانه های ملی ،خاورمیانه با استفاده از سیاهه وارسی ای مشتمل بر 13 مؤلفه که در آن به بررسـی تطبیقی وب سـایت سازمـان اسـناد و کتابخـانه ملی جمهوری اسـلامی ایران بـا وب سـایت هـای کتابخانه های ملی کشورهای اسلامی در منطقه خاورمیانه می پردازد کاملا مطابقت و همسویی دارد.

ص: 179

عکس

جدول ۱. سیاهه ارزیابی کاربردپذیری وب سایت نهاد کتابخانه‌های عمومی کشور از نظر میزان رعایت مؤلفه‌ها (به تفکیک)

	اعتبار اطلاعات	نهاد کتابخانه‌های عمومی کشور	
ردیف	تواتر	مؤلفه‌ها	
۱	۱،۴۵،۱۰،۱۵،۱۷،۱۸،۱۹،۲۳،۲۵	بیان نام مولف	-
۲	۱،۱۰،۳۵،۱۵،۱۶،۱۷،۱۸،۱۹،۲۳	بیان ویژگی‌های مولف (صلاحیت،شهرت،اعتبارو...)	-
۳	۶۸،۱۰،۴۵،۱۹،۴۲ ،۱	چگونگی تماس با سازمان یا شخص مسئول محتویات سایت	✓
۴	۱،۶،۱۵،۱۸،۲۱،۲۳،۲۶	دارا بودن تائیدیه رسمی از سوی سازمان یا فرد مسئول محتوی سایت	-
۵	۱و۴۲،۶،۲۵،۸،۱۵،۱۷	ذکر مشخصات سازمان یا شخص مسئول محتویات سایت	✓
۶	۱،۸،۱۵،۲۵،۳۷،۴۲	ارائه فهرستی از کارگزاران اصلی و مشخصات آن‌ها	-
۷	۲۴،۲۵،۲۹،۴۰،۲۷،۱،۸،۱۰،۱۵،۱۹	ذکر راهی برای تماس با مولف صفحه (نشانی پستی ،شماره تلفن و پست الکترونیک و ...)	✓
۸	۱،۱۵،۱۶،۱۸،۱۹،۲۰	بیان راهی برای اثبات ویژگی‌های مولف (تجارب وی در زمینه موضوعی خاص،عضویت وی در سازمان‌های حرفه‌ای و...)	-
۹	۱،۸،۲۷،۱۵،۱۷،۱۹	ذکر نام سازمان مسئول محتویات سایت	✓
۱۰	۶،۱۸،۱۹ ،۱	بیان ماهیت حامی	-
۱۱	۷،۴۲ ،۱،۸	ذکر مدت زمان تاسیس سازمان	-
۱۲	۱،۸،۱۵،۳۵،۲۲	بیان حق مالکیت معنوی منابع اطلاعاتی عرضه شده در سایت (ذکر نام)	✓
۱۳	۱،۸،۱۵،۴۲	بیان نام سازمان پشت پرده سایت	-
۱۴	۱،۸،۱۰،۱۵	ارائه سیاهه‌ای از نام سازمان‌ها یا سایت‌هایی که این سایت را توصیه می‌کنند	-
۱۵	۱،۱۵،۲۹،۱۶،۲۵	مشخص بودن هدف از طراحی و نشر صفحه	-
۱۶	۱،۸،۱۰،۱۵	ارائه فهرستی از اسامی و مشخصات افراد مسئول نظارت بر سازمان	-
۱۷	۱،۱۵	ارائه فهرستی از منابع چاپی منتشر شده توسط سازمان	-
۱۸	۲۷،۲۲،۴۰	قرار دادن نام و آرم سازمان سایت در هر صفحه	✓
۱۹	۳۰	نشان دادن آرم سازمان بصورت واضح و برجسته	✓
۲۰	۲۲	قرار دادن نام سازمان سایت در بالای صفحه	✓
۲۱	۱،۱۵	ذکر نام دارنده حق مالکیت معنوی	✓
۲۲	۹	قراردادن پرچم کشور	✓
	جمع امتیازها	۱۰۷	۴۵

	صحت اطلاعات	نهاد کتابخانه‌های عمومی کشور	
ردیف	تواتر	مؤلفه‌ها	
۲۳	۸،۲۵،۳۶،۱۶،۱۹ ،۲۷،۲۲،۴۰،۱،۶،۱۰	نبود خطاهای املائی ، گرامری و تایپی	✓
۲۴	۱،۱۰،۲۲،۱۵،۱۷،۲۴،۱۸،۲۳	ذکر نام و مشخصات کتابشناختی منبع اصلی	-
۲۵	۱،۱۰،۱۵،۱۹،۲۸،۱۷،۱۹	وجود شاخصی دال بر بررسی صحت و سقم اطلاعات توسط ویراستار یا مصحح در طول فرآیند بازنگری مجدد منابع	-
۲۶	۸،۳۵ ،۲۷،۳۹،۱،۱۰	عنوان بندی روشن و واضح گراف، نمودار یا جداول موجود در صفحه	✓
	جمع امتیازها	۳۲	۱۷

جدول 1 سیاهه ارزیابی کاربردپذیری وب سایت نهاد کتابخانه های عمومی کشور از نظر میزان رعایت مؤلفه ها (به تفکیک)

ص: 180

عکس

روزآمد بودن اطلاعات		نهاد کتابخانه‌های عمومی کشور	
ردیف	تواتر	مؤلفه‌ها	
۲۷	۱، ۷،۱۰،۱۵،۱۷،۱۸،۱۹،۲۳،۲۵	ذکر تاریخ آخرین تجدید نظر در محتویات صفحه	-
۲۸	۸،۱۶،۱۷،۱۸،۱۹،۲۳ ،۱،۱۰،۳۵	ذکر نخستین تاریخ قرار گرفتن منبع اطلاعاتی (با هر فرمتی) بر روی صفحه وب	-
۲۹	۶،۸،۱۰،۱۵،۱۹،۴۲ ،۱	ذکر فواصل زمانی به روز کردن اطلاعات دارای حساسیت زمانی	-
۳۰	۱،۶،۱۵،۱۸،۲۱،۲۳،۲۶	وجود اطلاعات آماری در صفحه	✓
۳۱	۱،۶،۲۵،۸،۱۵،۱۷،۴۲	بیان تاریخ گردآوری آمار	-
۳۲	۱،۸،۱۵،۲۵،۳۷،۴۲	ارائه تاریخ ها در یک فرمت بین المللی	-
	جمع امتیازها	۲۶	۷
سطح پوشش مخاطبان خاص		**نهاد کتابخانه‌های عمومی کشور**	
ردیف	تواتر	مؤلفه‌ها	
۳۳	۴۴،۴۰،۱،۶،۱۰،۸،۱۵،۱۶،۱۸،۲۸	مشخص بودن نام و نوع منابع موجود در صفحه	✓
۳۴	۱،۶،۱۰،۸،۱۵،۱۸،۲۸	تعیین مخاطبان خاص صفحه	-
۳۵	۱،۶،۱۰،۱۵	درج زمان تخمینی تکمیل صفحه در دست ساخت	-
	جمع امتیازها	۲۱	۱۰
نماهای تعاملی و تبادلی		**نهاد کتابخانه‌های عمومی کشور**	
ردیف	تواتر	مؤلفه‌ها	
۳۶	۲۷،۳۰،۱،۸،۴،۱۵،۱۱،۲۶،۳۸	وجود نظام مشخصی جهت بازخورد کاربران (feedback)	✓
۳۷	۱،۸،۴،۱۵،۲۶،۳۸	وجود نظام مشخصی برای کاربران برای درخواست اطلاعات بیشتر از سازمان	-
۳۸	۱،۲۷،۱۵،۲۶،۳۸	ذکر مشخصه زمانی لازم برای دریافت پاسخ از سوی سازمان	-
۳۹	۱،۲۶،۳۸	وجود امکان عضویت در سایت	-
۴۰	۱،۲۶،۳۸	وجود نظام مشخصی جهت عضویت کاربران در سایت	✓
۴۱	۱،۱۵	آگاهی کاربر از وجود مکانیزم کوکیز در سایت	-
۴۲	۲۷	پاسخگویی سریع و دقیق سیستم FAQ به پرسشهای کاربر	-
۴۳	۲۷	قابلیت پاسخگوئی به سئوالات کاربران	-
۴۴	۲۷	آگاهی کاربران از زمان دریافت پاسخ	-
	جمع امتیازها	۳۱	۱۲
عینیت اطلاعات		**نهاد کتابخانه‌های عمومی کشور**	
ردیف	تواتر	مؤلفه‌ها	
۴۵	۱،۱۰،۱۵،۱۶،۲۳،۴	مشخص بودن ارتباط میان شخص مولف با سازمان و فرد مسئول محتویات سایت	-
۴۶	۱،۱۵،۱۶،۴،۲۳	مشخص بودن دیدگاه مولف	-
۴۷	۱،۸،۱۵،۴،۳۸	مشخص بودن دیدگاه شخص یا سازمان مسئول تهیه اطلاعات	✓
۴۸	۱،۱۰،۱۵،۱۸	نبود در تبلیغات در صفحه	✓
۴۹	۱،۶،۱۷،۱۵،۲۷،۳۰	ارائه مطالبی در خصوص اهداف شخص یا سازمان گردآورنده اطلاعات	-
۵۰	۱،۱۵	توجیه اطلاعات غیر مرتبط با خدمات سازمان در صفحه	-

عکس

۵۱	۱۵،۳۹،۲۷	وجود وجه تمایزی میان محتوی اطلاعاتی و محتوی تفریحاتی	✓
۵۲	۲۷	شبیه نبودن اطلاعات به تبلیغات از لحاظ نوع طراحی	✓
۵۳	۴۰،۳۹	اجتناب از بکاربردن مفاهیم و اطلاعات نا مرتبط	✓
	جمع امتیازها	۳۴	۱۵
	ناوبری	**نهاد کتابخانه‌های عمومی کشور**	
ردیف	**تواتر**	**الف : ویژگیهای عنوان مرورگر**	
۵۴	۱،۱،۲۱	مبین نام سازمان یا فرد مسئول محتویات سایت	✓
۵۵	۱،۱۵،۲۶	مبین صفحه اصلی بودن صفحه	✓
۵۶	۱،۱۵	کوتاه بودن عنوان مرورگر	✓
۵۷	۱۵	منحصر بودن عنوان مرورگر برای سایت	✓
	جمع امتیازها	۹	۹
ردیف	**تواتر**	**ب: ویژگی‌های عنوان صفحه**	
۵۸	۱،۱۰،۳۴،۱۵،۱۶،۲۱	مبین متعلق بودن صفحه اصلی بودن صفحه	✓
۵۹	۱،۱۰،۲۷،۱۵،۲۶	مبین متعلق بودن صفحه به یک سایت مشخص (حداقل توسط یک آرم)	-
۶۰	۱و۷	کوتاه بودن عنوان صفحه	✓
۶۱	۳۴،۴۰،۴۴،۱۵	منحصر بودن عنوان صفحه برای سایت	✓
۶۲	۲۲،۲۷	بکاربردن کلمات کلیدی و مهم برای عناوین	✓
۶۳	۲۷،۲۹،۲۲	بکارگیری عنوان مناسب جهت بیان دقیق محتوی متن	✓
	جمع امتیازها	۲۲	۱۷
ردیف	**تواتر**	**ج: پیوندهای متنی و فرا متنی**	
۶۴	۱،۶،۱۰،۴،۱۵،۳۸،۲۵،۲۶	یکنواختی ظاهر پیوندهای تکراری	✓
۶۵	۲۷،۳۰،۲۴،۱،۶،۱۰،۴،۱۵،۲۱،۲۵	امکان ناوبری آسان میان صفحات	✓
۶۶	۱،۶،۳۰،۱۰،۴،۱۵،۲۱،۲۶	قرارگیری هماهنگ پیوندهای مستقیم داخلی در صفحه	✓
۶۷	۱،۱۵،۱۷،۲۱،۲۵،۲۶،۳۵	وجود تناسب میان عنوان پیوند با آنچه پیوندبدان ختم می شود	✓
۶۸	۱،۱۵،۱۶،۱۷،۱۸،۲۶	ادامه بکار پیوندها متناسب با اهداف از پیش تعریف شده	-
۶۹	۲۷،۴۰،۳۲،۱،۶،۴،۱۵،۲۱،۲۶	قابل درک بودن ساختار سایت / صفحه برای کاربران	✓
۷۰	۲۷،۱،۱۰،۱۵،۲۱،۲۵	تناسب چیدمان صفحات مختلف سایت با یکدیگر	✓
۷۱	۱،۱۵،۲۱،۲۶،۳۸	وجود میانبرهایی در صفحه اصلی جهت دسترسی هرچه آسانتر و موثر تر کاربران به صفحات پرمخاطب سایت	-
۷۲	۱،۱۰،۳۹،۱۵،۲۱،۲۶	وجود امکان دسترسی از صفحه اصلی سایت به بخش های اصلی آن	✓
۷۳	۱،۱۰،۱۵،۲۱	اجتناب از بکاربردن متن های زیر خط دار در کنار پیوندها	✓
۷۴	۱۵،۲۱،۲۵	عنوان دهی مناسب بوک مارک ها	-
۷۵	۱،۱۵،۱۹	امکان انتخاب اقلام اطلاعاتی مورد نظر از روی فهرست به جای تایپ آنها	-
۷۶	۱۵،۲۶	قرار دادن پیوند آرم سایت در صفحه اصلی	✓
۷۷	۳۰	شناساندن پیوندها از طریق زیر خط دار کردن آنها یا کاربرد نوعی رنگ خاص	-
۷۸	۲۷	وجود تناسب بین عنوان صفحه و پیوند	✓
۷۹	۴۰	نمایان بودن پیوندهای بازدید شده و بازدید نشده با استفاده از تغییر رنگ پیوند	-

عکس

۸۰	۲۷،۴۴	قرار دادن پیوند صفحه اصلی جهت مشخص بودن صفحه اصلی	✓
۸۱	۴۴	منطقی بودن تعداد پیوندها	-
۸۲	۳۰	پیوند دادن تصویر به صفحه مربوطه	-
۸۳	۲۴،۲۷	امکان شناسایی آسان پیوندها	✓
۸۴	۲۴	استفاده از عبارات مناسب برای پیوندها(عدم استفاده از عبارت clik here , more)	✓
۸۵	۲۲،۳۰	قرار دادن پیوند متنی در ابتدای پاراگراف	-
۸۶	۲۴	وجود تناسب بین عنوان صفحه و پیوند	✓
۸۷	۲۲	استفاده از علائم دیداری مانند رنگ، اندازه برای نشان دادن ارتباط میان پیوندها	-
۸۸	۴۰	بکارگیری متن کافی برای توضیح پیوند	-
	جمع امتیازها	۹۲	۶۸
ردیف	**تواتر**	**د: نشانه اینترنتی صفحه**	
۸۹	۱،۱۵،۲۱	درج نشانه اینترنتی صفحه در بدنه اصلی آن	-
۹۰	۲۷	مختصر بودن نشانه اینترنتی صفحه	✓
۹۱	۲۷	عدم تغییر در نشانه اینترنتی سایت	✓
۹۲	۳۰	کاربرپسند بودن نشانه اینترنتی صفحه	✓
	جمع امتیازها	۶	۳
ردیف	**تواتر**	**و: نقشه سایت یا نمایه**	
۹۳	۲۷،۳۹،۱،۶،۱۵،۲۱،۲۶،۳۷	وجود نقشه سایت یا نمایه ها بر روی صفحه اصلی و صفحات پیوندی	-
۹۴	۱،۶،۱۵،۲۱،۲۶،۳۷	دارا بودن موضوعات اصلی سایت	-
۹۵	۱،۶،۱۵،۲۱،۳۷	امکان خواندن آسان نمایه ها یا نقشه سایت	-
۹۶	۱،۲۷،۶،۱۵،۲۱،۳۷	سازماندهی منطقی نمایه ها یا نقشه سایت	-
	جمع امتیازها	۲۵	۰
ردیف	**تواتر**	**ی: موتور جستجوی داخلی**	
۹۷	۱،۴،۱۵،۲۱،۲۵،۲۸	وجود یک موتور جستجوی داخلی	✓
۹۸	۱،۴،۱۵،۲۱،۲۵،۲۸	مرتبط بودن اطلاعات بازیابی شده توسط موتور جستجوی داخلی	-
	جمع امتیازها	۱۲	۶
نماهای غیرمتنی		*نهاد کتابخانه های عمومی*	
ردیف	**تواتر**	مؤلفه ها	
۹۹	۲۷،۳۹،۱،۶،۱۵،۲۱،۲۶،۳۵،۴۲،۳۷	عدم استفاده از تصاویر متحرک (انیمیشن) بی مورد	✓
۱۰۰	۲۲،۴۰،۱،۱۰،۱۵،۲۱،۲۶،۳۷،۳۵	بکارگیری تصاویر گرافیکی، فایل های صوتی و تصویری به منظور افزایش کارآیی سایت	-
۱۰۱	۱،۱۰،۱۵،۲۱،۲۶،۳۷،۳۵	بیان نام نرم افزار خاص مورد نیاز و چگونگی دسترسی به آن	-
۱۰۲	۱،۱۰،۱۵،۲۱،۲۶،۳۷،۳۵	وجود جایگزینی برای فایلهای نیازمند نرم افزار خاص جهت دسترس پذیری اطلاعات برای کلیه کاربران	-
۱۰۳	۱،۱۰،۱۵،۲۶،۳۵،۳۷	بیان نام نرم افزار مرورگر مورد نیاز ، یا ویرایش خاصی از آن، جهت دسترسی به	-

عکس

		صفحات وب(در صورت لزوم)	
۱۰۴	۱،۱۰،۱۵،۲۷،۲۱،۲۶،۳۷،۳۵	وجود جایگزین متنی برای تصاویر و گرافیک های موجود در صفحه برای کلیه کاربران	-
۱۰۵	۱،۱۰،۳۰،۲۱،۳۵،۳۷،۴۲	عدم استفاده از تکنیک فلش (علائم چشمک زن)	✓
۱۰۶	۱،۱۵،۲۱،۲۶،۳۵،۳۷	آگاهی کاربر از بار شدن فایلی حجیم در صورت پیروی از یک پیوند	-
۱۰۷	۲۷	مرتبط بودن تصاویر گرافیکی با محتوی متن	✓
۱۰۸	۲۷	اضافه کردن متون به تصاویر برای درک بیشتر	-
۱۰۹	۳۰	استفاده از تگ ALT مناسب برای تصاویر	-
۱۱۰	۳۲	هدفمند بودن تصاویر گرافیکی	-
	جمع امتیازها	۶۴	۱۸
دسترس پذیری		***نهاد کتابخانه‌های عمومی کشور***	
ردیف	**تواتر**	**مؤلفه‌ها**	
۱۱۱	۱،۶،۱۰،۱۵،۱۶،۲۵،۴۱،۳۷،۴۳،۳۵	دسترس پذیر بودن با استفاده از مرورگر اینترنت اکسپلورر(Internet explorer 6.0)	✓
۱۱۲	۱،۶،۱۰،۱۵،۱۶،۲۶،۴۱،۳۷،۴۳،۳۵	دسترس پذیر بودن با استفاده از مرورگر نت اسکیپ نویگیتور (Netscape Navigator 6.5)	✓
۱۱۳	۳۷،۴۳،۳۵ ،۱،۶،۱۰،۱۵،۱۶،۲۶،۴۱	دسترس پذیر بودن با استفاده از مرورگر موزیلا برد (Mozilla Bird 0.7)	-
۱۱۴	۳۷،۴۳،۳۵ ،۱،۶،۱۰،۱۵،۱۶،۲۶،۴۱	دسترس پذیر بودن با استفاده از مرورگر اپرا (Opera 7.2)	-
۱۱۵	۱،۱۵،۱۶،۱۷،۴۳،۲۵،۲۶	دسترس پذیری سایت از طریق موتورهای جستجوی عمومی	✓
۱۱۶	۱،۴،۳۹،۲۶،۳۵،۳۸،۴۱	اندازه صفحه کمتر از ۵۰ کیلو بایت	-
۱۱۷	۱۰،۶،۱۵،۲۶	امکان پشتیبانی از سکوی عملیاتی کاربر (ویندوز me،۹۸،۲۰۰۰ ، لینوکس REDHAT)	✓
۱۱۸	۱۲ ،۴،۱۵،۳۵،۳۸،۲۷،۳۰،۲۴،۴۴،۳۲،۳۴	استفاده از قلم (فونت) های استاندارد	✓
۱۱۹	۱۵،۳۸	قابل رویت بودن تمامی اجزای سایت	✓
۱۲۰	۲۷	دسترسی کاربر به منابع مورد نیاز در کمتر از ۳ بار کلیک کردن	-
۱۲۱	۲۷	شبیه بودن اطلاعات به هرم واژگون (دستیابی سریع به اطلاعات مهمتر)	-
۱۲۲	۲۴و۲۷	سهولت استفاده ازسایت برای همه کاربران (مبتدی و متخصص)	✓
۱۲۳	۳۰و۳۹	عدم استفاده از تصاویر متحرک (انیمیشن) بی مورد	✓
	جمع امتیازها	۷۷	۴۸
کارآمدی		*نهاد کتابخانه های عمومی*	
ردیف	**تواتر**	**مؤلفه ها**	
۱۲۴	۲۷و۲۴و۳۹و۳۲و۳۴و۱و۱۰و۲۱و۱۶و۴۲،۳۷،۳۵	استفاده از رنگ های استاندارد در طراحی صفحات	✓
۱۲۵	۲۵،۲۶ ،۱،۱۵،۲۱،۱۷	عنوان بندی مناسب قابلیت های سایت/صفحه	✓
۱۲۶	۱۵،۲۸،۲۵،۱۹،۳۵	امکان تشخیص سایت بر اساس عنوان دامنه صفحه (Domin)	✓
۱۲۷	۱،۱۵،۳۹،۲۶،۲۸،۳۸	تناسب زبان سایت با فرهنگ و روحیات کاربر	✓
۱۲۸	۱،۱۵،۲۶،۳۸	وجود تعاریف حاشیه ای برای تشریح اقلام اطلاعاتی موجود بر روی صفحه	-

عکس

۱۲۹	۳۰،۲۲،۱،۱۵،۲۶،۳۸	اطلاع کاربر از عملیات در حال انجام	-
۱۳۰	۱،۱۵،۲۶،۳۸	تناسب اطلاعات موجود در صفحات با ماموریت سایت	-
۱۳۱	۱۵،۲۶،۳۷	امکان انجام تمامی قابلیت های سایت در درون سایت	✓
۱۳۲	۱۰،۱۵،۲۵	وجود قابلیت لغو عملیات صورت گرفته در سایت	-
۱۳۳	۱،۱۵،۲۶	وضوح موقعیت کاربر در سایت	-
۱۳۴	۱،۳۷،۳۹،۲۶	در نظر گرفتن پیشینه ذهنی کاربر	✓
۱۳۵	۱۷،۲۶	جذابیت سایت	-
۱۳۶	۱،۱۵،۲۶،۳۸	در نظر گرفتن ابزارهایی جهت کمک به کاربر	-
۱۳۷	۱۵،۳۴،۲۵،۳۰	ارائه عنوان کامل تمام سرواژه های مهم بکاررفته در صفحات	-
۱۳۸	۶،۴۲	طومارنوردی آسان صفحه خانگی	✓
۱۳۹	۲۵،۱۵	امکان چاپ اطلاعات بدون اعمال تغییرات در تنظیم سیستم رایانه	✓
۱۴۰	۱۵	بیان تعداد مراجعان سایت در یک بازه زمانی	✓
۱۴۱	۲۷،۴۴،۲۶	استفاده از عبارات کوتاه برای تشریح موارد موجود در صفحه	-
۱۴۲	۲۷،۳۴	ایجاد جلب توجه کاربران از طریق نوع طراحی (نوع رنگ ، فونت، و ....)	-
۱۴۳	۲۷،۳۹،۴۰	بهم فشرده نبودن مطالب در یک صفحه و ایجاد فضای خالی بین آنها	✓
۱۴۴	۳۹	دسترسی آسان به بخش Help	-
۱۴۵	۳	ارائه قیمت پیشنهادی به کاربر در صورت انتفاعی بودن سایت	-
۱۴۶	۳۰،۲۷	بالا بودن سرعت بارگذاری	-
۱۴۷	۲۷،۳۹،۳۴،۲۲	قابلیت استفاده از عملکرد جستجو	✓
۱۴۸	۲۹	امکان ناوبری توسط تورق	-
۱۴۹	۲۹،۲۷	امکان ناوبری توسط جستجو	-
۱۵۰	۲۷،۴۴	دارا بودن خطی مشی اختصاصی	-
۱۵۱	۲۴	قرار دادن مطالب در کمتر از ۳ صفحه	-
۱۵۲	۲۹	رعایت سلسله مراتب مطالب در صفحات (بر حسب تاریخ ، عنوان و...)	✓
۱۵۳	۴۴	بکار بردن عدد برای نشان دادن ارقام (۲ به جای دو)	✓
۱۵۴	۴۴	عدم استفاده از فایل های pdf بدلیل کاهش جذابیت (به جز برای اسناد)	✓
۱۵۵	۳۹	قابل رویت بودن پیام های اخطار	-
۱۵۶	۳۴،۴۴	پایین بودن زمان دانلود	-
	جمع امتیازها	۹۹	۵۱

ویژگی‌های ظاهری		نهاد کتابخانه‌های عمومی کشور	
ردیف	تواتر	مؤلفه ها	
۱۵۷	۲۲،۳۴،۴۰	صفحه آرایی مناسب	✓
۱۵۸	۲۷	وجود هماهنگی بین طرح ها و صفحات	✓
۱۵۹	۳	وجود هماهنگی بین رنگ ها و سبک نگارش	✓
۱۶۰	۲۴	بکارگیری حاشیه در اطراف متن	✓
	جمع امتیازها	۶	۶

5-نتیجه گیری

در مجموع نتایج به دست آمده از ارزیابی کاربرد پذیری وب سایت نهاد کتابخانه های عمومی کشور به تفکیک معیارهای 11 گانه ی موجود در این پژوهش در مقایسه با وب سایت شاخص فرضی با 663 امتیاز، در سطحی معادل 52 درصد (346 امتیاز)، قرار دارد و این امر گر چه بیان گر آن است که تنها نیمی از استانداردها در وب سایت نهاد کتابخانه های عمومی کشور رعایت گردیده است و این امر با وب سایت شاخص فاصله زیادی دارد لذا لازم است در طراحی این وب سایت دقت و توجه بیش تری به عمل آید افزون بر این با عنایت به کاربردی بودن این پژوهش پژوهش گران در صدند تا با انعکاس نتایج پژوهش حاضر، کتابداران و طراحان وب سایت ها را از نقاط ضعف و قوت وب سایت مطلع ساخته تا با به کارگیری عناصر لازم و ویژگی های مناسب در طراحی وب سایت فضایی را ایجاد نمایند که کاربر به محض ورود به وب سایت امکان بهره گیری از تمامی اطلاعات و خدمات مرتبط با این ،محمل بدون وجود هر گونه محدودیتی را داشته باشد.

6-پیشنهادهای پژوهش

با توجه به یافته های پژوهش در باب رعایت معیارهای مطرح در کاربرد پذیری هر چه بیشتر و بهتر وب سایت نهاد کتابخانه های عمومی کشور توجه به رعایت معیارها و مولفه های : بیان نام مولف و ویژگی های او (صلاحیت شهرت و اعتبار و....) ، ذکر فواصل زمانی به روز کردن اطلاعات دارای حساسیت زمانی درج زمان تخمینی تکمیل صفحه در دست ساخت نماهای تعاملی و تبادلی، وجود امکان عضویت در سایت، مشخص بودن ارتباط میان شخص مولف با سازمان و فرد مسئول محتویات سایت منحصر بودن عنوان صفحه برای سایت نمایان بودن پیوندهای بازدید شده و بازدید نشده با استفاده از تغییر رنگ پیوند درج نشانه اینترنتی صفحه در بدنه اصلی آن بیان نام نرم افزار خاص مورد نیاز و چگونگی دسترسی به آن بیان تعداد مراجعان سایت در یک بازه زمانی توصیه می گردد

منابع

1. اصغری ،پوده احمدرضا 1380. بررسی عناصر و ویژگی های مطرح در طراحی وب سایت کتابخانه های دانشگاهی. پایان نامه کارشناسی ارشد کتابداری و اطلاع رسانی دانشکده علوم تربیتی و روانشناسی، دانشگاه فردوسی مشهد

2. الکساندر ژانت 1383. شناخت وب: چگونه اطلاعات موجود بر روی وب را ارزیابی نموده و صفحاتی اینگونه را پدید آوریم، ترجمه صدیقه محمد اسماعیل، تهران: دبیزش.

3. جانسون، استیو 1382. صفحات سفارش و دریافت مواد منابع کتابخانه ای بر روی وب. ترجمه صدیقه محمد اسماعیل اطلاع شناسی، 1 (پاییز): 187-199.

4. حاجی زین العابدینی ،محسن مکتبی ،فرد ،لیلا عصاره، فریده 1384 تحلیل پیوندهای وب سایت های کتابخانه های ملی جهان مجله مطالعات تربیتی و روانشناسی دوره 7، شماره 1، ص 193-173

5. خوانساری، جیران ساختن سایت های وب موفق برای کتابخانه های کوچک دانشگاهی. صنعت برق 54 (آبان) ص. 28-24.

6. دایره المعارف کتابداری و اطلاع رسانی ج. 1و 2. 1381 تهران: نهاد کتابخانه های عمومی کشور، .1381

7. رضایی شریف آبادی ،سعید، فرودی، نوشین 1381. ارزیابی صفحات وب کتابخانه های دانشگاهی ایران و ارائه الگوی پیشنهادی فصلنامه کتاب دوره سیزدهم 4 (زمستان) ص. 19-12.

8. عصاره، فریده، مرادمند، علی 1384. شناسایی ویژگی های عمده در طراحی وب سایت های کتابخانه های ملی جهان به منظور ارائه الگویی مناسب جهت ارتقاء کیفی وب سایت نهاد کتابخانه های عمومی کشور. فصلنامه اطلاع شناسی پاییز و زمستان شماره 1و 2.ص. 170-190

9. عصاره فریده 1381 معیارهای ارزیابی منابع اینترنتی فصلنامه کتاب، دوره سیزدهم 2 (تابستان). ص. 73-61

10. محمد اسماعیل صدیقه 1383. بررسی کاربرد پذیری وب سایتهای دانشگاه های صنعتی کشور. پایان نامه دکتری کتابداری و اطلاع رسانی تهران دانشگاه آزاد اسلامی واحد علوم و تحقیقات

11. محمد اسماعیل صدیقه 1384 بررسی کاربرد پذیری وب سایت های دانشگاه های صنعتی کشور فصل نامه کتاب دوره 16 شماره 1.

12. نویدی، فاطمه .1386 ارزیابی دسترس پذیری وب سایت وزارتخانه های دولت جمهوری اسلامی .ایران پایان نامه کارشناسی ارشد علوم کتابداری و اطلاع رسانی دانشگاه تربیت مدرس

13. Alexander, J. E. ,Tate, M. A.1999. Web Wisdom: who to evaluate and create information quality on the web. London, Mahwah, NewJersey.LEA: (Lawrence Erlbaum Associates).

14. Back, S. E.1997. Evaluation criteria: the good, The Bad and The Ugly: Or, Why It's a Good Idea to Evaluate Web Sources". [on-line] Available: http:lib.nmsu.edu/instruction/ evalcrit.html

15. Barker. J. 2004 Finding Information on the Internet: A Tutorial University of California. [on-line].Available:

16. www.lib.berkeley.edu/TeachingLib/Guides/Internet/Evaluate.html

17. Barker. J. 2003. Evaluating Web Pages: Techniques to Apply Questions to Ask." VC Berkeley - Teaching Library Internet Workshops, 12sep.2003.[on-line]. Available: http:// www.Lib.berkeley.edu/Teaching Lib/Guides/Internet/Evaluated.html

18. Berger. P. 1999 Web Evaluation Guide: Tramline, Incorporated. [on-line]. Available:

http://www.infosearcher.com/cybertours/tours/touro4/-tourlaunch1.html

Bertot J.C. et al. 2006 Functionality, usability and accessibility: Interactive user-centered evaluation . 19 strategies for digital libraries", Performance Management and Metrics, Vol. 7

ص: 187

.No. 1, pp. 17-28

Braynik. G. 2003. Atomatic Web Usability Evaluation: What Need to be done?. (29jan.2003): 1-16. .20
[on-line]. Available: http://Usable.binghamton.edu.Atomatic Thesis. html

Clyde. L. A. 1996. The Library as information providers, The Homepage The Electronic Library. Vol. .21
.14 no.6 pp:549-558

Engle. M. 1996. Evaluating Websites: Criteria and Tools. New York Library Association Conference, .22
Saratoga Springs, Ny. (October 1996): 1-3.[on-line]. Available: http://www.
.Library.Cornel.edu/Okuref/research/Webeval.html

.The Essential Web Site Usability Checklist. 2007. [on-line]. Available: www.dailybits. com .23

Hupp.J.2008. Test Your Web Site: A 57- Point Checklist for Maximum Usability .[on- online]. . 24
.Available: www.virtualhosting.com

.Jafari. A. Optimizing Campus Web Sites. EDUCASE QUARTERLY, No.2 2000:56-58 .25

.Leggett. D. Quick Usability Checklist.2009, [on-line]. Available: www.uxbooth.com .26

Meyers. P. J. 25-Point Web Site Usability Checklist, 2008. [on-line]. Available: www. usereffect.com .27

Nadler.D.M. and Furman.V.M. (2001). Access board issues final standards for disabled access under . 28
.Section 508 of Rehabilitation Act, Government Contract Litigation Reporter, Vol. 14 No. 19, p. 14

National Science foundation (NSF).Universal Design of College Algebra, 2008. [on- line]. Available: .29
www.usablealgebra.landmark.edu

Nielsen.J.Coyne, K, Tahir, Marie, (2001). Make It Usable-Web Site Usability Magazine, (6 feb.2001): .30
.1-5.[on-line]. Available: http://www.Pcmag.com/article2/0,4149,33821,00. asp, 23jan.2003

Osareh.F. (2003). "A notice on content of library information science (LIS) schools websites, Libri (3): .31
262-265

.The SEO File: Usability Checklist 2009. [on-line]. Available: www.theseofiles.co.uk .32

Sloan, John. Rampo College Web Design Standards. Rampo College of New yersey Statements and . 33
policies.2001.[on-line].Available: http://www.guide.rampo.edu 1/ content/Webstandards.html

34. Stueart, R.D. Library and Information Center management fifth edition. Endewood ,

ص: 188

.Colorado: Libraries Unlimited ;INC, 1998

Sullivan, T. Matson R. Barriers to use: Usability and Content Accessibility on the Web's Most Popular .35
.Sites. Interdisciplinary PhD. program in information Science, University of North Texas, Denton, 2000

Tungar.M. N. Heuristic Evaluation.2002.[on-line]. Available: http://www.cc.gatech. . 36
.edu/~manas/cs8803a/Heuristic.pdf

.Usability Guidelines Information Science Technology, 2004. [on-line]. Available: www. web.mit.edu .37

.Web Usability Checklist. The College of New Jersey. 2008. [on-line]. Available: www. teng.edu" .38

Weibel .S.L. The World Wide Web and Emerging Internet Resource Discovery Standards for Scholarly .39
.Litrature. Library Trends.vol.43, No.4 1995:627-664

.Wilson, S. 1995 World Wide Web Design Guide. Indianapolis. IN: Hayden Books, 1995 .40

Zaphiris. P. Darin, Ellis, R. Website Usability and Content Accessibility of The Top USA Universities .41
2001. Dertoit, MI: Institute of Geronotology and Dept of Industrial and Manufacturing Engineering Wayne
.State University, 2001

Web Usability Tips to Attract and Retain Web Visitors 2009,:[on-line].Available: 50 . 42
.www.doshdosh.com

چکیده

هدف: هدف کلی پژوهش دستیابی به فرایند پردازش و سازماندهی وبگاه ها با استفاده از قواعد و استانداردهای مورد استفاده در سازمان اسناد و کتابخانه ملی ایران است.

جامعه آماری: جامعه آماری این پژوهش را 50 وبگاه تشکیل می دهند که در 20 رده موضوعی از وبگاه پارس ایندکس و به صورت تصادفی انتخاب شده اند.

روش شناسی: این پژوهش به دو روش کتابخانه ای و پیمایش توصیفی انجام شده است- جهت گردآوری اطلاعات از سیاهه وارسی و برای تجزیه و تحلیل داده ها از آمار توصیفی استفاده شده است.

نتایج: نتایج نشان می دهند که استاندارد سازی موضوع سازماندهی و پردازش وب گاه ها به عنوان نوعی از منابع الکترونیکی و بومی سازی آن در سازمان اسناد و کتابخانه ملی ایران امکان پذیر است.

کلیدواژه ها: وب گاه ها سازماندهی امکان سنجی سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

امکان سنجی پردازش وبگاه ها در سازمان اسناد و کتابخانه ملی ایران

اشاره

دکتر رضا خانی پور (1) | محبوبه قربانی (2) | سهیلا فعال (3)

مقدمه

امروزه شبکه جهانی اینترنت منابع اطلاعاتی متنوعی را در دسترس قرار داده است. وب هر روزه پهنه بیش تری از وسعت دنیا را در بر گرفته و در دور افتاده ترین نقاط جهان نیز رسوخ نموده است. شبکه جهانی وب با دسترسی پذیری خود امکانات زیادی را برای مخاطبان فراهم نموده است. دسترسی به آخرین اخبار دنیا و اطلاعات کشفیات و نوآوری ها آشنایی با فرهنگ و تمدن ملل مختلف، دریافت اطلاعات مورد نیاز روزانه مانند آب و هوا و اخبار وقایع مهم از امتیازات دسترسی به وب است. این گستردگی و دسترس پذیری در کنار محاسنی که دارد کاربران را با مشکلاتی روبه رو ساخته است. اطلاعات در لابلای صفحات وب نهفته و نیاز به جستجوهای دقیقی برای کشف آن ها وجود دارد.

عمده ترین مشکل کاربران ،وب دست یابی به بهترین و با کیفیت ترین اطلاعات مورد نیاز و حصول جامعیت و مانعیت در بازیابی اطلاعات در زمینه های تخصصی است (علی محمدی 1381)

ص: 191

1- دکترای کتابداری و اطلاع رسانی عضو هیات علمی و مدیر کل پردازش و سازماندهی سازمان اسناد و کتابخانه ملی ایران -R Khanipour@nlai.ir

2- دانشجوی دکترای کتابداری و اطلاع رسانی معاون اداره کل پردازش و سازماندهی سازمان اسناد و کتابخانه ملی ایران -m ghorbani@nlai.ir

3- کارشناس ارشد کتابداری و اطلاع رسانی رئیس گروه سازماندهی منابع غیر کتابی سازمان اسناد و کتابخانه ملی ایران -s faal@nlai.ir

کتابخـانه هـا و مراکز اطلاـع رسـانی به عنوان متولیـان گردآوری سازمانـدهی و اشـاعه اطلاعات، خود را موظف به گردآوری سازمانـدهی اطلاعات موجود بر روی وب می دانند یکی از چالش های فراروی کتابداران نحوه سازماندهی منابع وب بوده است تلاش های گسترده ای در خصـوص سازمانـدهی اطلاعـات بر روی وب صورت گرفته است تهیه انواع اسـتانداردها و دست نـامه هـا و ابـداع فراداده هـای متنوع سازماندهی منابع وب از جمله تلاش های صورت گرفته در جهت سازماندهی منابع وب است. (نشاط 1382)

وبگاه ها: شناخته شده ترین منابع وب

منـابع وب آثـاری آثاری هسـتند که توسط افراد مختلف در قالب های متنوع تولیـد و انتشار یافته انـد. این منابع وبگاه ها ،وبلاگ ها کتاب های الکترونیکی مقالات تمام نشـریات الکترونیکی و.... را در بر می گیرد. یکی از مهم ترین منابع تحت وب، وبگاه ها (1) هستند وبگاه مجموعه ای از صـفحات وب است که دارای یک دامنه اینترنتی و به صورت مجموعه ای از صـفحات مرتبط که داده هایی نظیر ،متن صدا، تصویر و فیلم روی آن ها ارائه می شود، روی شـبکه اینترنت قرار می گیرد صفحه وب به صورت سندی است که در قالب اچ تی ام ال (2) نوشـته می شود و همواره با اسـتفاده از پروتکل اچ تی تی پی (3) می توان به آن دسترسـی پیدا کرد. مهم ترین قسـمت یک وبگاه در واقع صفحه اصـلی یا صـفحه خانگی آن است وبگاه یک مؤسـسه مرکز تجاری یا سازمان چهره او به سوی جهان و نقطه شـروع بیش تر کاربران تلقی می شود نیلسن (2002) (4)

ارزیابی وبگاه ها

عواملی که مجموع قابلیت به رهوری از یک وبگاه را تعیین می کننـد عبارتنـد از: محتوا، زبان، ساختار، ،طراحی جهت یابی عبور و قابلیت دسترسـی روش هـای گوناگونی جهت ارزیابی قابلیت های بهره برداری از یک وبگاه وجود دارد که عبارتنـد از ارزیابی با مشارکت کاربر و ارزیابی بـدون مشارکت کاربر روش ارزیابی در صورتی مفید خواهد بود که هم زبان و هم ساختار وبگاه مورد ارزیابی به سـهولت توسـط کاربران درک شود. (پل 2007) (5)

استانداردها و طرحهای ابر داده ای سازماندهی منابع وب

استانداردها و طرح های ابر دادهای مختلفی برای توصیف منابع اینترنتی تهیه شده اند که از انواع آن ها می توان به موارد ذیل اشاره نمود:

1. طرح ابر داده ای دابلین کور (6)

1- WebSites.

2- HTML

3- HTTP

4- Nielsen, 4J

5- Roswitha Poll

6- Dublin Core (DC)

2. مارک 21 (1)

3. آر دی اف (2)

4. طرح کدگذاری توصیف آرشیوی (3)

5. قالب ابر دادهای خدمات مکان یاب اطلاعات دولتی (4)

6. قالب ابر داده ای طرح کدگذاری متن (5)

پردازش منابع الکترونیکی در سازمان اسناد و کتابخانه ملی ایران

در سازمان اسناد و کتابخانه ملی ایران برای پردازش و سازماندهی انواع منابع کتابی و غیر کتابی از ویرایش دوم قواعد فهرست نویسی انگلو امریکن (6)، استاندارد بین المللی توصیف کتاب شناختی (7) و استاندارد یونی مارک (8) استفاده می شود. کاربرد مارک (9) در سازماندهی منابع به طور جدی از سال 1385 آغاز شد. در این راستا از مارک ایران که شکل بومی سازی شده استاندارد یونی مارک است استفاده شده است. بر اساس این استاندارد منابع الکترونیکی در گروهی جداگانه و با کد 1 مشخص می شوند. در کتابخانه ملی ایران انواع کار برگه ها با توجه به این تقسیم بندی طراحی شده اند و هر کار برگه از 10 بلوک و فیلدهای اصلی و فرعی مربوط به توصیف هر منبع تشکیل شده است.

بر اساس طرح مقدماتی برای فهرست نویسی منابع الکترونیکی(2003) (10) ، منابع الکترونیکی از لحاظ نوع دسترسی به دو به دو دسته منابع قابل «دسترسی مستقیم» (11) و منابع قابل «دسترسی از راه دور» (12) تقسیم می شوند (عبداللهی 1383). بنابراین وب گاه ها دسته ای از منابع الکترونیکی و از نوع «دسترسی از راه دور» به حساب می آیند منابع الکترونیکی در واحد سازماندهی منابع غیر کتابی اداره کل پردازش، سازماندهی می شوند. تاکنون بیش از 3600 منبع الکترونیکی از نوع «دسترسی مستقیم» در این واحد سازماندهی شده اند.

با توجه به مطالب پیش گفته به نظر می رسد که استانداردها و قواعد توصیف منابع الکترونیکی برای توصیف وبگاه ها نیز کاربردی باشد.

بیان مساله

وبگاه ها به عنوان نوعی از منابع اطلاعاتی وب در چرخه اطلاعات و دانش محسوب می شوند. با توجه به این که پردازش و سازماندهی منابع اطلاعاتی هسته مرکزی و فنی این چرخه به حساب می آید، لازم

1- Machin Readabel cataloging (MARC)

2- Encoded Archival description(EAD)

3- Resource Description Framework(RDF)

Government information locator sources (GILS) –4

Text Encoding Initiative(TEI) –5

Anglo American Cataloging Rules, Second Edition (AACR2) –6

International Standard Bibliographic Description (ISBD) –7

Universal MARC –8

Machine–Readable Cataloging –9

Draft internet guidelines for cataloging electronic resources –10

Direct access –11

Remote access –12

می نماید برای استفاده هر چه کامل تر از محتوای اطلاعاتی وبگاه ها سازمان اسناد و کتابخانه ملی ایران که مسئولیت استاندارد سازی و نظارت بر سازماندهی را بر عهده دارد از پردازش و سازماندهی وبگاه ها نیز غافل نشود با عنایت به اینکه کتابخانه ملی ایران در عرصه پردازش انواع منابع کتابخانه ای اعم از چاپی و الکترونیکی وارد شده است لازم است برای روزآمد سازی کنترل مدیریت و سازماندهی وبگاه ها نیز طرح و برنامه ای داشته باشد. این پژوهش بر آن است تا امکان پردازش وبگاه ها را به عنوان نوعی از منابع الکترونیکی مهم در عرصه های ملی و جهانی توسط سازمان اسناد و کتابخانه ملی ایران مورد پژوهش قرار دهد.

پیشینه پژوهش

پیشینه پژوهش در ایران:

حاجی زین العابدینی (1381) در پژوهشی به بررسی مشکلات اینترنت در زمینه سازماندهی و بازیابی اطلاعات پرداخته و دست نامه فهرست نویسی منابع اینترنتی را ارائه کرده است در تهیه دست نامه از قواعد فهرست نویسی انگلو امریکن قواعد به کار گرفته شده در طرح های فهرست نویسی منابع اینترنتی و الکترونیکی، اصول و قواعد فهرست نویسی منابع فارسی و نمونه های منابع اینترنتی فهرست نویسی شده، استفاده شده است. با استفاده از این دستنامه امکان فهرست نویسی منابع اینترنتی و سازماندهی اطلاعات دل خواه در اینترنت فراهم می شود هم چنین کاربرگ های ایجاد شده است که بر اساس فیلدهای اطلاعاتی موجود در منابع ،اینترنتی دستورالعمل های دست نامه نمونه های موجود در طرح هایی چون اینترکت، رهنمود های ایجاد مارک و فیلد های اطلاعاتی در راهنمای نرم افزار مارکایت طراحی شده است.

فتاحی حسن زاده (1385) به منظور مطالعه و ارزیابی شیوه های سازماندهی اطلاعات در وبگاه های کتابخانه های دانشگاهی پژوهشی انجام داده اند نتایج کلی پژوهش بیان گر آن است که در سازماندهی ،اطلاعات صفحه اول وبگاه ها، شیوه دسته بندی بر اساس نوع خدمات بیش از سایر انواع رایج در حالی که برای سازماندهی سایر صفحات از شیوه های نسبتا متنوعی استفاده می شود همچنین شیوه الفبایی عنوان و موضوع بیش ترین کاربرد و شیوه دایرکتوری (موضوعی سلسله مراتبی) کم ترین کاربرد را برای سازماندهی انواع منابع اطلاعاتی در وب سایت کتابخانه های دانشگاهی دارند.

پیشینه پژوهش در خارج از ایران:

کوچ (1) و همکارانش (1997) نقش طرح های رده بندی را در توصیف و بازیابی منابع اینترنتی مورد بررسی قرار داده اند. آن ها استفاده از این طرح ها را در سازماندهی محتویات وب گاه ها توصیه کرده اند، ولی به کاربرد شیوه های مختلف سازماندهی و نظریات کتابداران و کاربران در مورد چگونگی آن ها نپرداخته اند.

ویلیامسون (1997) (2) با تأکید بر ساختار دانش موجود در منابع اینترنتی در راستای سازماندهی دانش و

1- Koch

2- Williamson

بازیابی اطلاعات، لزوم سازماندهی آن ها را بیان نموده است. در تحقیق وی، بررسی امکان کاربرد شیوه های متعارف سازماندهی برای سازماندهی محتوای اطلاعاتی وبگاه ها توصیه شده است.

وارد (2001) (1) در مقاله ای به تشریح اهمیت فهرست نویسی منابع اینترنتی پرداخته است. در ادامه فهرستی از فعالیت های انجام شده در کتابخانه های ایالات متحده در خصوص سازماندهی منابع اینترنتی را ارائه کرده است.

وایلر (2) و همکارانش (2008) در پژوهشی به ارزیابی هزینه های پردازش و سازماندهی منابع وب در کتابخانه و دانشگاه ملی کرواسی پرداخته اند نتایج این پژوهش نشان داده است که زمان پردازش منابع وبی با پردازش منابع چاپی یکسان است.

یانگه ی (2011) (3) در پژوهشی به بررسی فاصله بین ایجاد کنندگان منابع وب و ایجاد کنندگان فراداده های آن ها به منظور بهبود عناصر فراداده پرداخته است. یافته های پژوهش نشان گر این بوده است که رضایت کاربر تا حد بالایی به مفید بودن سهولت استفاده و دریافت اطلاعات و اثر بخشی عناصر اطلاعاتی بستگی دارد.

اهمیت پژوهش

سازماندهی و پردازش وبگاه ها در سازمان اسناد و کتابخانه ملی ایران می تواند از جنبه های زیر حائز اهمیت باشد:

1. ایجاد بانک اطلاعاتی وبگاه ها و غنی سازی کتاب شناسی ملی ایران؛

2. امکان جستجوی وبگاه ها با استفاده از قابلیت های جستجوی نرم افزار کتابخانه ای (رسا)؛

3. ایجاد نقاط دسترسی توصیفی و تحلیلی برای برآوردن نیازهای اطلاعاتی کاربران نهایی؛

4. ایجاد جامعیت و مانعیت در بازیابی پیشینه های کتاب شناختی وبگاه ها؛

5. استاندارد سازی و یکپارچه سازی پردازش توصیفی و تحلیلی وبگاه ها.

اهداف پژوهش

هدف کلی پژوهش دست یابی به فرایند پردازش و سازماندهی وبگاه ها با استفاده از قواعد و استاندارد های مورد استفاده در سازمان اسناد و کتابخانه ملی ایران است به این منظور اهداف فرعی زیر دنبال می شوند:

1. بررسی چگونگی دستیابی و دریافت اطلاعات وبگاه ها؛

2. توصیف کتاب شناختی وبگاه ها بر اساس استانداردها و قواعد مورد استفاده در سازمان اسناد و کتابخانه ملی ایران؛

3. دسترسی جامع و مانع به وبگاه ها با ایجاد نقاط بازیابی و تحلیل موضوعی مورد استفاده در سازمان اسناد و کتابخانه ملی ایران؛

Ward –1

Willer –2

Younghee –3

4. کاربرد نظام های موضوعی مورد استفاده در سازمان اسناد و کتابخانه ملی ایران برای نمایه سازی وبگاه ها

پرسش های اساسی

1. دستیابی و دریافت اطلاعات وبگاه ها از چه راه هایی امکان پذیر است؟

2. توصیف کتاب شناختی وبگاه ها بر اساس استانداردها و قواعد مورد استفاده در سازمان اسناد و کتابخانه ملی ایران چگونه است؟

3. دسترسی جامع و مانع به وبگاه ها با ایجاد نقاط بازیابی و تحلیل موضوعی مورد استفاده در سازمان اسناد و کتابخانه ملی ایران چگونه است؟

4. کاربرد نظام های موضوعی مورد استفاده در سازمان اسناد و کتابخانه ملی ایران برای نمایه سازی وبگاه ها چگونه است؟

تعاریف عملیاتی

امکان سنجی: منظور از امکان سنجی در این پژوهش بررسی طرح اولیه جهت ارائه و پیاده سازی فرایند پردازش و سازماندهی وبگاه ها در اداره کل پردازش و سازماندهی سازمان اسناد و کتابخانه ملی ایران است.

پردازش: منظور از پردازش در این پژوهش فرایند سازماندهی ،منبع شامل دریافت، توصیف و تحلیل است.

جامعه آماری

جامعه آماری این پژوهش را 50 وبگاه تشکیل می دهند این جامعه مورد مطالعه در 20 رده موضوعی از وبگاه پارس ایندکس (1) و به صورت تصادفی انتخاب شده اند.

روش پژوهش و ابزار گردآوری داده ها

در این پژوهش دو روش کتابخانه ای و پیمایش توصیفی به کار گرفته شده است. جهت گردآوری اطلاعات از سیاهه وارسی و برای تجزیه و تحلیل داده ها از آمار توصیفی استفاده شده است. روایی سیاهه وارسی بر اساس تجربه پدید آوران مقاله در مدت بیش از یک دهه فعالیت تخصصی در حوزه سازماندهی منابع الکترونیکی و همچنین از طریق مشاوره با استادان به دست آمده است

یافته های پژوهش و پاسخ به پرسش های اساسی

پرسش اول: دست یابی و دریافت اطلاعات وبگاه ها از چه راه هایی امکان پذیر است؟

1- http://www.parsindex.com

در این پژوهش جهت دستیابی و دریافت اطلاعات وبگاه ها دو روش مورد بررسی قرار می گیرد:

1. استفاده از فهرست وب گاه ها

برخی وب گاه ها و پایگاه های اطلاعاتی فهرست وب گاه های مختلف را بر اساس تقسیم بندی های موضوعی ارائه می دهند به این ترتیب یکی از راه های دست یابی به اطلاعات وب گاه ها، استفاده از فهرست وب گاه ها .است نمونه فهرست های وب گاه ها را می توان در وب گاه ایران (1) پارس ایندکس و الکسا (2) و غیره جستجو کرد.

در این پژوهش برای دسترسی به حجم نمونه از جامعه مورد مطالعه وبگاه پارس ایندکس انتخاب شد و امکان دست رسی به بسیاری از وب گاه ها امکان پذیر گردید.

2. ورود اطلاعات وبگاه ها در وبگاه سازمان اسناد و کتابخانه ملی ایران

منابع اینترنتی گروهی از منابع اطلاعاتی به حساب می آیند و کتابخانه ملی به لحاظ رسالت خود، لازم است مانند کتاب ها و سایر منابع اطلاعاتی اقداماتی در خصوص فراهم آوری منابع اینترنتی نیز به انجام رساند. در این راستا وب گاه سازمان اسناد و کتابخانه ملی ایران می تواند درگاهی جهت دست یابی به اطلاعات وب گاه ها .باشد. بنابراین با طراحی «فرم الکترونیکی ورود اطلاعات وب گاه ها» و قرار دادن آن در وبگاه سازمان می توان به نشانی وب گاه ها و اطلاعاتی که برای پردازش آن ها مورد نیاز است، دست پیدا کرد.

جدول 1 فیلدهای پیشنهادی برای طراحی فرم الکترونیکی ورود وب گاه ها را نشان می دهد.

عکس

در این پژوهش جهت دستیابی و دریافت اطلاعات وبگاه‌ها دو روش مورد بررسی قرار می‌گیرد:

۱. استفاده از فهرست وبگاه‌ها

برخی وبگاه ها و پایگاه‌های اطلاعاتی، فهرست وبگاه‌های مختلف را بر اساس تقسیم بندی‌های موضوعی ارائه می‌دهند، به این ترتیب یکی از راه‌های دستیابی به اطلاعات وبگاه‌ها، استفاده از فهرست وبگاه‌ها است. نمونه فهرست‌های وبگاه ها را می توان در وبگاه ایران[1]، پارس ایندکس و الکسا[2] و غیره جستجو کرد.

در این پژوهش برای دسترسی به حجم نمونه از جامعه مورد مطالعه، وبگاه پارس ایندکس انتخاب شد و امکان دسترسی به بسیاری از وبگاه ها امکان‌پذیر گردید.

۲. ورود اطلاعات وبگاه‌ها در وبگاه سازمان اسناد و کتابخانه ملی ایران

منابع اینترنتی گروهی از منابع اطلاعاتی به حساب می‌آیند و کتابخانه ملی به لحاظ رسالت خود، لازم است مانند کتاب‌ها و سایر منابع اطلاعاتی اقداماتی در خصوص فراهم‌آوری منابع اینترنتی نیز به انجام رساند. در این راستا وبگاه سازمان اسناد و کتابخانه ملی ایران می‌تواند درگاهی جهت دستیابی به اطلاعات وبگاه‌ها باشد. بنابراین با طراحی «فرم الکترونیکی ورود اطلاعات وبگاه‌ها» و قرار دادن آن در وبگاه سازمان میتوان به نشانی وبگاه‌ها و اطلاعاتی که برای پردازش آنها مورد نیاز است، دست پیدا کرد.

جدول ۱ فیلدهای پیشنهادی برای طراحی فرم الکترونیکی ورود وبگاه‌ها را نشان می‌دهد.

جدول ۱. فیلدهای پیشنهادی برای طراحی فرم الکترونیکی ورود اطلاعات وبگاه‌ها

نام فیلد	عنوان فارسی	عنوان به زبان دیگر	صاحب امتیاز	طراح رایانه ای	سال تولید	زمان آخرین ویرایش	زبان	حوزه موضوعی	کلید واژه‌ها	نشانی اینترنتی	توضیحات
نوع فیلد	ضروری	عادی	ضروری	عادی	ضروری	ضروری	ضروری	ضروری	ضروری	ضروری	عادی

پرسش دوم: توصیف کتاب‌شناختی وبگاه‌ها بر اساس استانداردها و قواعد مورد استفاده در سازمان اسناد و کتابخانه ملی ایران چگونه است؟

جدول ۲ عناصر پیشنهادی توصیف وبگاه‌ها را با استفاده از قواعد فهرستنویسی انگلوامریکن در بخش منابع الکترونیکی و در بستر فرادادهای یونی مارک نشان می‌دهد.

1. http://www.iran.ir/directory
2. http://www.alexa.com

جدول 1 فیلدهای پیشنهادی برای طراحی فرم الکترونیکی ورود اطلاعات وب گاه ها

پرسش دوم: توصیف کتاب شناختی وب گاه ها بر اساس استانداردها و قواعد مورد استفاده در سازمان اسناد و کتابخانه ملی ایران چگونه است؟

جدول 2 عناصر پیشنهادی توصیف وب گاه ها را با استفاده از قواعد فهرست نویسی انگلوامریکن در بخش منابع الکترونیکی و در بستر فراداد های یونی مارک نشان می دهد.

1- http://www.iran.ir/directory

2- http://www.alexa.com

عکس

جدول ۲. فراوانی کاربرد قواعد انگلوامریکن و استاندارد یونی مارک در توصیف کتابشناختی ویگاه‌ها

ناحیه	عناصر توصیف	انگلوامریکن (قواعد)	یونی مارک (فیلد)	یونی مارک (بلوک)	درصد فراوانی
عنوان و پدیدآور	عنوان کامل	B۹/۱	$a۲۰۰	اطلاعات توصیفی (۲)	٪۷۶
	وجه تسمیه عام: [منابع الکترونیکی]	C۹/۱	$b۲۰۰		
	عنوان به زبان دیگر	D۹/۱	$d۲۰۰		
	دیگر اطلاعات عنوان	E۹/۱	$e۲۰۰		
	شرح پدیدآور	F۹/۱	$f۲۰۰		
وضعیت ویراست		B۹/۲	$a۲۰۵		٪۶۸
بر اساس آخرین ویرایش قواعد انگلوامریکن ناحیه ۳ (نوع انتشار) برای منابع الکترونیکی ذکر نمی‌شود					
نشر و تولید	محل تولید	C۹/۴	$a۲۱۰	اطلاعات توصیفی (۲)	٪۷۰/۶
	ناشر و صاحب امتیاز	D۹/۴	$c۲۱۰		
	تاریخ تولید	F۹/۴	$d۲۱۰		
بر اساس قواعد انگلوامریکن، ناحیه مشخصات ظاهری برای منابع الکترونیکی دسترسی از راه دور ذکر نمی‌شود					
یادداشت	ماهیت و دامنه	۱B۹/۷	۳۰۰	یادداشت ها (۳)	٪۵۱/۴
	زبان	۲B۹/۷	۳۰۰		
	منبع عنوان کامل	۳B۹/۷	۳۰۴		
	شرح های پدیدآور	۶B۹/۷	۳۰۴		
	ویراست و تاریخچه کتابشناختی اثر	۷B۹/۷	۳۰۵		
	نشر، پخش و غیره	۹B۹/۷	۳۰۶		
	مشخصات ظاهری	۱۰B۹/۷	۳۰۷		
	عنوان های مرتبط	۴B۹/۷	۳۱۲		
	مسوولیت معنوی اثر	۶B۹/۷	۳۱۴		
	مندرجات	۱۸B۹/۷	۳۲۷		
	چکیده	۱۷B۹/۷	۳۳۰		
	مخاطبان	۱۴B۹/۷	۳۳۳		
	نوع منابع الکترونیکی	۸B۹/۷	۳۳۶		
	روش دسترسی	۸B۹/۷	۳۳۷		
درصد فراوانی کل					٪۶۰

جدول 2 فراوانی کاربرد قواعد انگلو امریکن و استاندارد یونی مارک در توصیف کتاب شناختی وب گاه ها

ص: 198

همان گونه که در جدول 2 مشاهده می شود بیش ترین کاربرد قواعد و اسـتانداردهای مورد مطالعه برای توصیف وب گاه ها در ناحیه عنوان و پدیدآور با 76 درصد و کم ترین آن متعلق به ناحیه یادداشت با 51/4 درصد است. بنابراین در پاسخ به پرسش دوم می توان گفت امکان توصـیف کتاب شـناختی وب گاه ها اساس اسـتانداردها و قواعد مورد اسـتفاده در سازمان اسناد و کتابخانه ملی ،ایران به میزان 60 درصد است.

پرسـش سوم: دسترسـی جامع و مانع به وب گاه ها با ایجاد نقاط بازیابی و تحلیل موضوعی مورد استفاده در سازمان اسناد و کتابخانه ملی ایران چگونه است؟

بخش دوم قواعد فهرست نویسی انگلوامریکن قواعد مربوط به نقاط بازیابی و تحلیل موضوعی را ارائه می دهد جدول 3 فیلد های بکار رفته در یونی مارک را بدین منظور نشان می دهد.

عکس

همان‌گونه که در جدول ۲ مشاهده می‌شود بیشترین کاربرد قواعد و استانداردهای مورد مطالعه برای توصیف وبگاه‌ها در ناحیه عنوان و پدیدآور با ۷۶ درصد و کمترین آن متعلق به ناحیه یادداشت با ۵۱/٤ درصد اســـت. بنابراین در پاسخ به پرسش دوم می توان گفت امکان توصیف کتاب‌شناختی وبگاه‌ها بر اساس استانداردها و قواعد مورد استفاده در سازمان اسناد و کتابخانه ملی ایران، به میزان ٦۰ درصد است.

پرسش سوم: دسترسی جامع و مانع به وبگاه‌ها با ایجاد نقاط بازیابی و تحلیل موضوعی مورد استفاده در سازمان اسناد و کتابخانه ملی ایران چگونه است؟

بخش دوم قواعد فهرستنویسی انگلوامریکن قواعد مربوط به نقاط بازیابی و تحلیل موضوعی را ارائه می‌دهد. جدول ۳ فیلدهای بکاررفته در یونی مارک را بدین منظور نشان می‌دهد.

جدول ۳. فراوانی کاربرد استاندارد یونی مارک در نقاط بازیابی و تحلیل موضوعی وبگاه‌ها

نام فیلد	شماره فیلد	شماره بلوک	درصد فراوانی
عنوان اصلی به زبان دیگر	۵۱۰	عنوان‌های مرتبط (۵)	٪۶۱/۵
عنوان های گونه گون دیگر	۵۱۷		
عنوان گسترده	۵۳۲		
عنوان ترجمهشده	۵۴۰		
نام شخص به منزله موضوع	۶۰۰	تجزیه و تحلیل موضوعی (۶)	٪۸۲
نام تنالگان به منزله موضوع	۶۰۱		
عنوان به منزله موضوع	۶۰۵		
موضوع (اسم عام یا عبارت اسمی عام)	۶۰۶		
نام جغرافیایی به منزله موضوع	۶۰۷		
نام شخص به منزله شناسه افزوده	۷۰۲	مسئولیت معنوی اثر (۷)	٪۵۳/۳
نام تنالگان به منزله سرشناسه	۷۱۰		
نام تنالگان به منزله شناسه افزوده	۷۱۲		
نشانی اینترنتی	۸۵۶	بین المللی کاربرد (۸)	٪۱۰۰
درصد فراوانی کل			٪۷۰/۴

با توجه به جدول ۳، بیشترین کاربرد استاندارد یونی مارک مربوط به بلوک ۸ با ۱۰۰ درصد است، بعـــد از آن بلـــوک ٦ با ۸۲ درصد و بلوک ۵ با ٦۱/۵ درصد قرار دارد. کمترین کاربرد نیز در بلوک ۷ با ۵۳/۳ درصد مشاهده می‌شود. بنابراین در پاسخ به پرسش سوم میتوان گفت امکان دسترسی جامع و مانع

جدول 3 فراوانی کاربرد استاندارد یونی مارک در نقاط بازیابی و تحلیل موضوعی وب گاه ها

با توجه به جدول، بیش ترین کاربرد اسـتاندارد یونی مارک مربوط به بلوک 8 با 100 درصد است بعد از آن بلوک 6 با 82 درصد و بلوک 5 با 61/5 درصـد قرار دارد کم ترین کاربرد نیز در بلوک 7 با 53/3 درصـد مشاهده می شود بنابراین در پاسخ به پرسـش سوم می توان گفت امکان دسترسی جامع و مانع

ص: 199

به وب گاه ها با ایجاد نقاط بازیابی و تحلیل موضوعی مورد استفاده در سازمان اسناد و کتابخانه ملی ایران حدود 70 درصد است.

پرسش چهارم: کاربرد نظام های موضوعی مورد استفاده در سازمان اسناد و کتابخانه ملی ایران برای نمایه سازی وب گاه ها چگونه است؟

در کتابخانه ملی، ایران منابع غیر کتابی با اسـتفاده از زبان کنترل شده نمایه سازی می شوند. جدول 4 نظام های موضوعی مورد استفاده را نشان می دهد لازم به ذکر است علاوه بر نظام های موضوعی زیر، از سر عنوان های موضوعی کتابخانه کنگره (1)، سایر اصطلاحنامه ها و واژه نامه های موضوعی، دایره المعارف ها و بانک های اطلاعاتی عمومی و موضوعی جهت مستندسازی توصیف گر ها و پیشنهاد آن ها به نظام های موضوعی پیش گفته استفاده می شود

عکس

به وبگاه‌ها با ایجاد نقاط بازیابی و تحلیل موضوعی مورد استفاده در سازمان اسناد و کتابخانه ملی ایران حدود ۷۰ درصد است.

پرسش چهارم: کاربرد نظام‌های موضوعی مورد استفاده در سازمان اسناد و کتابخانه ملی ایران برای نمایه‌سازی وبگاه‌ها چگونه است؟

در کتابخانه ملی ایران، منابع غیرکتابی با استفاده از زبان کنترل‌شده نمایه‌سازی می‌شوند. جدول ٤ نظام‌های موضوعی مورد استفاده را نشان می‌دهد. لازم به ذکر است علاوه بر نظام‌های موضوعی زیر، از سرعنوان‌های موضوعی کتابخانه کنگره[1]، سایر اصطلاحنامه‌ها و واژه‌نامه‌های موضوعی، دایره المعارف‌ها و بانک‌های اطلاعاتی عمومی و موضوعی جهت مستندسازی توصیفگرها و پیشنهاد آنها به نظام‌های موضوعی پیشگفته استفاده می‌شود.

جدول ٤. فراوانی کاربرد نظام‌های موضوعی در نمایه‌سازی وبگاه‌ها

نظام موضوعی	حوزه موضوعی	درصد فراوانی
اصطلاحنامه فرهنگی فارسی(اصفا)	علوم انسانی	٪۶۶
اصطلاحنامه پزشکی فارسی	پزشکی	٪۱۰۰
اصطلاحنامه های علوم	فنی- مهندسی	٪۱۰۰
سرعنوان های موضوعی فارسی	همه علوم	٪۷۲
درصد فراوانی کل		٪۸۴/۵

بر اساس جدول ٤، هر کدام از نظام‌ها در حوزه موضوعی خود در جامعه مورد مطالعه، بررسی شده‌اند. بنابراین در پاسخ به پرسش چهارم می‌توان نتیجه گرفت که کاربرد نظام‌های موضوعی مورد استفاده در سازمان اسناد و کتابخانه ملی ایران برای نمایه‌سازی وبگاه‌ها بالای ٨٤ درصد امکان‌پذیر است.

بحث و نتیجه‌گیری

یافته‌های پژوهش نشان می‌دهند پردازش وبگاه ها در سازمان اسناد و کتابخانه ملی ایران، بیش از ٦٥ درصد امکان‌پذیر است. بر اساس نتایج این پژوهش برای طراحی فرایند پردازش وبگاه‌ها می‌توان فرم الکترونیکی ثبت وبگاه‌ها را طراحی کرد و آن را در وبگاه سازمان دسترس‌پذیر نمود. در برخی موارد نیز با استفاده از فهرست وبگاه‌ها می‌توان نوعی فراهم‌آوری را انجام داد. در مرحله بعدی با استفاده از قواعد فهرستنویسی انگلوامریکن در بستر فرادادهای یونی مارک و بومی سازی آن و با استفاده از کاربرگه

1. (LCSH) Library of Congress Subject Headings

جدول4 فراوانی کاربرد نظام های موضوعی در نمایه سازی وب گاه ها

بر اساس جدول 4، هر کدام از نظام ها در حوزه موضوعی خود در جامعه مورد مطالعه بررسی شده اند. بنابراین در پاسخ به پرسش چهارم می توان نتیجه گرفت که کاربرد نظام های موضوعی مورد استفاده در سازمان اسناد و کتابخانه ملی ایران برای نمایه سازی وب گاه ها بالای 84 درصد امکان پذیر است.

بحث و نتیجه گیری

یـافته هـای پژوهش نشـان می دهنـد پردازش وب گاه ها در سازمان اسـناد و کتابخانه ملی ایران، بیش از 65 درصـد امکان پذیر است. بر اساس نتایـج این پژوهش برای طراحی فرایند پردازش وب گاه ها می توان فرم الکترونیکی ثبت وب گاه ها را طراحی کرد و آن را در وب گاه سازمان دسترس پـذیر نمود. در برخی موارد نیز با اسـتفاده از فهرست وب گاه ها می توان نوعی فراهم آوری را انجام داد در مرحله بعدی با استفاده از قواعد فهرست نویسی انگلوامریکن در بستر فرادادهای یونی مارک و بومی سازی آن و با استفاده از کاربرگه

1- (LCSH) Library of Congress Subject Headings

منابع الکترونیکی اقدام به توصیف و تحلیل وبگاه ها نمود سپس آن ها را بر اساس نقاط بازیابی مختلف دسترس پذیر ساخت. با توجه به نتایج پژوهش، نمایه سازی و تحلیل موضوعی وب گاه ها نیز با استفاده از نظام های موضوعی پیش گفته امکان پذیر خواهد بود

در مقایسه نتایج این پژوهش با پژوهش های پیشین می توان خاطر نشان کرد نمونه سازماندهی وب گاه ها در کتابخانه ها در پژوهش «وایلر» و همکارانش و همچنین «وارد» مورد بررسی قرار گرفته است پردازش وب گاه ها از لحاظ هزینه و رضایت کاربران همان گونه که «وارد»، «وایلر» و همکارانش و هم چنین «یانگهی» در پژوهش خود اشاره کرده اند از اهمیت و ارزش والایی برخوردار است. فتاحی» و «حسن زاده» و همچنین «کوچ» و همکارانش نیز بر اعمال رده بندی و فهرست موضوعی، جهت دسترسی به وب گاه ها تاکید کرده اند. همچنین استفاده از قواعد و استانداردهای مرسوم سازماندهی از جمله قواعد فهرست نویسی انگلوامریکن و مارک برای سازماندهی وب گاه ها در پژوهش «حاجی زین العابدینی»، «ویلیامسون» و «کوچ» و همکارانش مورد بررسی و تایید قرار گرفته است

به طور کلی از این پژوهش می توان نتیجه گرفت که استاندارد سازی موضوع سازماندهی و پردازش وب گاه ها به عنوان نوعی از منابع الکترونیکی و بومی سازی آن در سازمان اسناد و کتابخانه ملی ایران امکان پذیر است

پیشنهاد های برخاسته از پژوهش

پس از انجام این پژوهش و با عنایت به یافته های آن پیشنهاد می شود سازمان اسناد و کتابخانه ملی ایران نسبت به انجام موارد زیر که جنبه عملیاتی دارند اقدام نماید .

1. طراحی و ایجاد سامانه ورود اطلاعات وب گاه ها در وب گاه سازمان اسناد و کتابخانه ملی ایران جهت فراهم آوری و پردازش آن ها؛

2. ارائه آموزش های تخصصی در حوزه پردازش و سازماندهی وب گاه ها به نیروی انسانی و کارشناسان ذیربط؛

3. برنامه ریزی در جهت اشاعه اطلاعات وب گاه ها به صورت جامع و مانع.

منابع

پارس ایندکس http://parsindex.com/default.aspx (دسترسی در 1391/7/20)

پژوهشگاه علوم و فناوری اطلاعات ایران 1390 اصطلاح نامه های علوم http://thesauri.irandoc.ac.ir (دسترسی در 1391/7/20)

پل روزیتا. 1390. ارزیابی وب گاه کتابخانه ای پژوهش های آماری و معیارهای کیفیت ترجمه رضا خانی پور گزیده مقالات ایفلا 2007 دوربان آفریقای جنوبی، 23-19 اوت 2007 تهران نشر کتابدار

ص: 201

سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران 117-103

حاجی زین العابدینی، محسن . 1381. بررسی مسائل فهرست نویسی منابع اینترنتی و ارائه دست نامه پیشنهادی برای کتابخانه های ایران پایان نامه کارشناسی ارشد دانشگاه علوم پزشکی و خدمات بهداشتی درمانی ایران دانشکده مدیریت و اطلاع رسانی پزشکی.

سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران بانک مستندات http://opac.nlai.ir (دسترسی در 1391/7/20).

عبداللهی، مریم 1383 ارزیابی وضعیت فهرست نویسی منابع الکترونیکی بر اساس قواعد انگلوامریکن در کتابخانه های دانشگاهی و آرشیوهای موجود در شهر تهران پایان نامه کارشناسی ارشد. دانشگاه آزاد ،اسلامی واحد تهران شمال

علی محمدی ،داریوش 1381 . ابر پایگاه های اطلاعاتی در شبکه جهانی .وب فصلنامه کتاب. 49(1): 93- 100

فتاحی، رحمت الله. حسن زاده محمد. 1385 نظر سنجی از کتابداران متخصص پیرامون شیوه های سازماندهی اطلاعات در وب سایت کتابخانه های دانشگاهی گزارشی از مرحله دوم یک طرح پژوهشی فصلنامه کتابداری و اطلاع رسانی 36 (4): 6-30.

کمیته ملی مارک ایران 1381 مارک ایران تهران: کتابخانه ملی جمهوری اسلامی ایران

نشاط، نرگس. 1382 چالش های سازماندهی موضوعی منابع وب اطلاع شناسی. 1: 37-54.

IFLA. 2008. UNIMARC concise bibliographic format. Available at: http://archive.ifla.org/ VI/8/unimarc-concise-bibliographic-format-2008.pdf [September. 2012].

Koch, Traugott et al. 1997. The role of classification schemesin Internet resource description and discovery. . [Available at:www.ub.lu.se/desire/radar/reports/D3,2,3/class-v10.html [September. 2012

Nielsen, J.2002. Top 10 guidelines for homepage usability. Available at: http://www.useit.com/alertbox/20020512.htm1 [September. 2012].

Steering Committee of American Library Association, et al.. 2002. Anglo American Cataloging Rules. Joint .Chicago: American Library Association

Ward, Diane. 2001. Internet resource cataloging: the SUNY Buffalo Libraries' response. OCLC Systems .Services, 17 (1):19 - 26

Willer, Mirna. Buzina, Tanja. Holub, Karolina. Zajec, Jasenka. Milinovic, Miroslav. Topolscak, Nebojsa .2008. Selective Archiving of Web Resources: A Study of Processing Costs. Program: Electronic Library .and Information Systems. v42. n4: p341- 364

Williamson N. J.1997. Knowledge structures and the Internet. Knowledge organization for information retrieval: proceedings of thesixth international study conference on classification research. University .College London 16–18 June: pp23–27

Younghee Noh. 2011. A study on metadata elements for web–based reference resources system developed .through usability testing. Library Hi Tech, 29(2):242–265

چکیده

محاسبه هزینه ها دانسته های کنونی ما را در خصوص ایجاد آرشیو تحت وب و اینکه این نوع آرشیو در مقایسه با شیوه فراهم آوری منابع چاپی، فعالیتی پرهزینه بوده را تأیید می کند مقاله حاضر اطلاعات سودمندی را در خصوص ایجاد یک آرشیو گزینشی منابع تحت وب و چگونگی ارتقای برنامه های نرم افزاری ارزیابی مجموعه های فراهم آوری شده را در اختیار خواننده قرار می دهد. و در مقاله راهکارهایی ارائه می شود که منجر به افزایش دانش در زمینه بهره گیری از فناوری های مربوط به ایجاد آرشیو تحت ،وب برای افزایش کارآمدی تمامی پروژه های مربوط به ایجاد آرشیو تحت وب می.شود یکی از این پروژه ها که در مقاله بررسی شده است ، آرشیو کتابخانه ملی استرالیاست که اخیراً با استفاده از رویکرد گزینشی، موفق شده که هزینه های سرانه خود را در یک بازده زمانی کاهش دهد.

مقدمه

طی دهه ،گذشته تعدادی اندکی – اما – اما رو به رشد – از کتابخانه های ملی اقدام به تدوین برنامه هایی در زمینه ایجاد آرشیو تحت وب کرده اند. برنامه هایی که یک یا بیش از چهار رویکرد اصلی زیر را در برداشته اند:

1. ایجاد آرشیو گزینشی، برای مثال، آرشیوهای کتابخانه ملی ،کانادا، ژاپن و استرالیا؛

2. تهیه و تدارک دامنه وب در کل کشور به طور دوره ای، برای مثال نمونه های انجام گرفته توسط آرشیو های کشور های اسکاندیناوی، از جمله سوئد؛

3. گردآوری موضوعی، برگرفته از کتابخانه کنگره مجموعه مینروا از منتخب های سال 2000 ، 2002، و 11 سپتامبر 2001؛ و

مجموعه های واسپاری نظیر STORS کتابخانه ایالتی تاسمانی و واسپاری الکترونیکی (1) کتابخانه ملی هلند

کتابخانه ملی ،استرالیا به دلیل مزایای موجود در رویکرد گزینشی این روش را انتخاب کرد این دلایل عبارت اند از:

• هر یک از اقلام موجود در آرشیو کتابخانه از لحاظ کیفیت مورد ارزیابی قرار گرفته و تا سر حد امکان با بهره گیری از قابلیت های فنی موجود در این زمینه کاربرد پذیر می شوند.

• هر یک از اقلام موجود در آرشیو را می توان به طور کامل فهرست نویسی، و به بخشی از کتاب شناسی ملی تبدیل کرد و داده های کتاب شناختی را به اشتراک گذاشت هم چنین در فهرست کتابخانه، اطلاعات کتاب شناختی منابع تحت وب با منابع دیگر ادغام شده و کاربران می توانند به صورت یکجا به تمامی منابع دسترسی پیدا کنند.

• هر کدام از اقلام موجود در وب می تواند به سرعت از طریق محیط وب برای مخاطبان خود دسترس پذیر شود؛ زیرا پیش تر تلاش شده اجازه انجام این کار از ناشران اخذ شود.

• ویژگی های هر یک از منابع موجود در آرشیو و شیوه طبقه بندی آن ها برای مدیران مجموعه شناخته شده است. این امر در نخستین گام توانایی ما را در توسعه روش ها و ابزار های گردآوری منابع ذخیره سازی و دسترس پذیری به آن ها افزایش می دهد افزون بر این شناخت و آگاهی، مدیران مجموعه را بر آن می دارد که سیاست ها و راهبرد های حفاظت و نگهداری منابع را به گونه ای اتخاذ کنند که امکان دسترس پذیر سازی منابع برای مدت زمان طولانی فراهم شود.

• سایت هایی را که برای روبات های جست وجوگر موجود دسترس پذیر نیستند نیز می توان شناسایی و گردآوری نمود و با استفاده از روش های دیگر- برای مثال بر حسب نام ناشر - مرتب کرد. این امر، نام های تجاری - که نیازمند رمز عبور ناشر هستند- و پایگاه داده ها را نیز در برمی گیرد.

با وجود مزایای فوق هر یک از رویکرد های موجود در زمینه ایجاد آرشیو تحت وب، معایبی نیز دارد. رویکرد گزینشی نیز از این قاعده مستثنا نیست این ،امر به کارمندان کتابخانه بستگی دارد که در محیطی مملو از اطلاعات کاملاً جدید به ایفای وظیفه می پردازند، عملکرد آن ها در این زمینه، باید به گونه ای باشد که در نهایت به قضاوت در خصوص آن چه که در آینده باید مورد تحقیق و تفحص قرار گیرد، بی انجامد و عرصه تحقیقات آینده را روشن سازد. رویکرد گزینشی همچنین به استخراج منابع، جدا از محتوا و پیوند میان آن ها با دیگر منابع خارجی می پردازد. از نظر توسعه و مدیریت آرشیو، بزرگ ترین نقطه ضعف رویکرد آرشیوی این است که روند کار فشرده و پر زحمت است و هزینه انجام کار به ازای هر منبع که قرار است آرشیو شود، بالاست.

ایجاد آرشیو گزینشی منابع تحت وب

اشاره

ایجاد آرشیو گزینشی منابع تحت وب (2)

بررسی هزینه های مربوط به فراهم آوری منابع تحت و در کتابخانه ملی استرالیا

نوشته: مارگارت فیلیپس (3) | ترجمه: صدیقه محمد اسماعیل (4)

تاریخچه

کتابخانه ملی استرالیا در 1996، به صورت گزینشی و آزمایشی اقدام به ایجاد آرشیوهای تحت وب کرد.

ص: 206

1- e-Depot

2- Selective Archiving of Web Resources: A Study of Acquisition Costs at the National Library of Australia

3- Margaret Philips, Director, Digital Archiving, National Library of Australia

4- استادیار کتابداری و اطلاع رسانی و دانشیار گروه کتابداری و اطلاع رسانی دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران

اقدامی که در تاریخ خود حرکتی بسیار نو محسوب می شد و کمتر نوشته ای بدان پرداخته بود. حتی کسی نبود که بتوان از او آموخت. ما در مورد ایجاد آرشیوگان تحت وب اطلاعات کافی نداشتیم تا بتوانیم در این خصوص اقدام به تدوین طرح یا راهبردهای مناسبی جهت برآورد هزینه های مربوط به این کار کنیم. تنها در سایه برداشتن گام های کوچک عملی و با عنوان خودآموزی توانستیم به راهمان ادامه داده و رشد کنیم هیچ گونه کمک مالی خاصی برای انجام این کار جدید دریافت نمی کردیم، فقط مجبور بودیم آن دسته از کارکنان مجموعه سازی را که جدا از وظایف کاملاً سنتی ،خود، به کار جدید ایجاد آرشیوگان تحت وب علاقه مند بوده و از خود استعدادهایی در این زمینه بروز می دادند، برای انجام این کار جدید تربیت کنیم برای مثال شخصی در بخش فناوری اطلاعات باید بخشی از وقت خود را صرف این می کرد که چگونه می تواند منابع را از وب گاه های ناشران دریافت و اطلاعات آن ها را بر روی سرور کتابخانه نگهداری کند ما برای در اختیار گرفتن و مدیریت فایل های دریافت شده، تنها مجبور به استفاده از نرم افزارهایی بودیم که به طور رایگان موجود بود با اعمال چنین تدابیری نه تنها هزینه های اولیه ایجاد آرشیو تحت وب کاملاً اندک، بلکه عمدتاً در بودجه مربوط به کارکنان پنهان می شد.

با گذشت 9 سال عملیات مربوط به مراتب پیچیده تر شدند کتابخانه استرالیا نرم افزار پاندورا (1) یعنی آرشیو تحت وب استرالیا (2) را راه اندازی کرد که به مثابه آرشیوی کارآمد، وارد طرح شراکت با 9 کتابخانه دیگر استرالیا و نیز دیگر سازمان های مسئول گردآوری منابع فرهنگی، یکی پس از ،دیگری شد. در مجموع می توان گفت که این امر روندی رو به رشد و توسعه داشت، اما نهادها را با افزایش هزینه ها مواجه ساخت حجم افزونتر فعالیت های مربوط به ایجاد آرشیو و نیاز آن به پشتیبانی از مشارکت کنندگانی که از مکان های دور در این طرح شرکت می کردند، ضرورت ایجاد زیر ساخت های پیچیده فنی متشکل از سیستم ارسال و تحویل مدارک، مدیریت آرشیو، و ذخیره سازی را ایجاب می کرد این ،امر به مفهوم وجوب هزینه های ضروری در امر توسعه بود، هر چند که این هزینه ها همچنان از طریق بودجه های موجود پرسنلی برای توسعه مجموعه سازی و بخش فناوری اطلاعات فراهم می شد.

کتابخانه مذکور نمی توانست نظام مدیریت آرشیو مناسبی برای خریداری پیدا کند، بنابراین، نظام ایجاد آرشیو دیجیتالی پاندورا (پانداس) (3) به همراه نظام تحویل مدارک پاندورا، پا به عرصه وجود

گذاشتند. همچنین کتابخانه اقدام به خرید سیستم ذخیره سازی شیء دیجیتال (دی او اس اس) کرد که زمینه اشتراک منابع را میان پاندورا و دیگر مجموعه های دیجیتالی کتابخانه را فراهم می کرد.

یکی از چیزهایی که در 9 سال گذشته تغییر نکرده این است که: کتابخانه هنوز هیچ بودجه اضافی برای انجام این فعالیت ها در اختیار ندارد به طرزی عجیب شواهد موجود بر سودآور بودن چنین اقدامی دلالت دارد؛ لذا چنین مسئله ای ما را بر آن می دارد که بودجه ای را برای تحقق این امر از محل بودجه دولت استرالیا که هر ساله رقم آن رو به تزاید است تعریف نماییم با این حال هیچ گونه طرح کوتاه

1- PANDORA

2- Australia's Web Archive

3- PANDORA Digital Archiving System (PANDAS)

مـدت خاصـی که بتوانـد بودجه ریزی این فعالیت را به گونه ای توجیه کنـد تا هنگام [طرح] اتمام به ثبات این فعالیت بی انجامد و آن را در حالت بی ثباتی و ناپایداری رها نکند، وجود ندارد.

آرشیو پاندورا در حال حاضر

آرشیو پاندورا مجموعه قابل توجهی از نشریات برخط و وب گاه های استرالیاست که توسط کتابخانه ملی این کشور و شرکای آن ایجاد شده و گسترش یافته .است آرشیوی که در کتابخانه ملی کانبرا (1)، به شیوه ای متمرکز ذخیره، اداره، و حفظ و نگهداری می شود. از 30 آوریل 2005، که آرشیو مشتمل بر 8235 عنوان ،بود، تا کنون نرخ متوسط رشـد مجموعه سالیانه در حدود 2400 عنوان می باشد. این عناوین شامـل فایـل مجزایی از اسـناد متنی نظیر فایل های پی دی اف (2)، یـا اقلام تحت ،وب ،نظیر وب گاه های حجیم و پیچیـده، متشـکل از هزاران فایـل در اشـکال و فرمت هـای ،مختلف شامـل ،متن ،صـدا ،تصویر و یـا ویـدئوست که بر پایه نظمی مشـخص و به منظور ایجاد «شاخصـی» جدید از عناوین آرشـیوی ای که در حال حاضر رقم آن ها در آرشیو موجود به معادل 16/736 مورد بالغ می گردد گردآوری شده است. هدف این بود که عناوین منتخب به شـیوه ای ایمن و صحیح در آرشیو ذخیره سازی و به طور مستمر برای همگان دسترس پذیر شود.

آرشـیو مزبور شامل هر دو نوع انتشارات برخط ایستا و پویا (3) بوده و وب گاه ها را نیز در بر می گیرد، و نشان دهنـده طیف گسترده ای از انواع انتشارات و فرمت های اسـتفاده شده توسط ناشران و پدید آورندگان بر روی وب است. افزون بر این نشریات برخط و وب گاه هایی را که هم اکنون در محیط زنده وب ناپدید گردیده و در هیچ کجای دیگر نیز در در دسترس نمی باشند نیز در برمی گیرد.

بسـیاری از عناوین موجود در این آرشـیو می توانـد با اتصال به اینترنت به راحتی و آزادانه در دسترس تمامی افراد در هر نقطه از جهان قرار گیرد تقریبا پنج سال است – اندکی کمتر – که به دلایل تجاری دسترسی به بخش بسیار کوچکی از آرشیو محدود شده است. با وجود این، اطلاعات مربوط به این عناوین دارای این محدودیت را می توان بر روی کامپیوتری واحد در اتاق مطالعه اصلی کتابخانه مشاهده کرد.

دسترسـی به محتویـات آرشـیو را می توان یا از طریق ایجاد پیونـدی مناسب به پیشـینه فهرست نویسـی منبعی خاص یا از طریق موضوع و فهرست عنـاوین موجود در وب گاه پانـدورا امکان پـذیر کرد. موتور های جست و جوی متعارف ماننـد گوگل و یاهو، به نمایه سازی منابع آرشیوی بر حسب آن چه که در عنوان آن ها آمده است می پردازند.

محدوده وظایف

وظایفی که کارکنان کتابخانه ملی و دیگر شـرکای آن ها به عنوان بخشـی از برنامه ایجاد آرشـیو تحت وب پانـدورا انجام می دهنـد متأثر از مجموعه عوامل متعـددی نظیر تصـمیمات مربوط به سـیاست گذاری ها و شـرایط محیطی است؛ مجموعه سـیاست ها و عواملی که نه به شکلی ،حداقل بلکه به شیوه ای کاملاً مؤثر بر

1- Canberra

2- PDF

هزینه های انجام طرح اثر گذار هستند.

افراد شرکت کننده در طرح پاندورا به بالاترین شکل بر حفظ اصل به طرفه العینی به نتایج مناسب (رسیدن خصوصیات ظاهری و کارآمد پذیری منبع) در باب منبع منتشره و یا وب گاه، و نیز توجه همزمان بر محتویات آن تأکید می ورزند به محض آن که فرد گردآوری کننده نسخه ای از یک منبع را بر روی سرور کتابخانه ملی قرار دهد افراد همکار در سازمان های دیگر نیز نسخه مزبور را جهت کسب اطمینان از جامعیت و کارآمدی آن و قبل از ارسال آن منبع به آرشیو جهت استفاده عموم، مورد بازبینی قرار می دهند. این تضمین کیفیت فرآیندی بسیار وقت گیر و در نتیجه گران قیمت و پرهزینه است که در واقع، گران ترین جنبه فرآیند فراهم آوری منابع نیز محسوب می شود.

هر عنوان در آرشیو بر مبنای رکوردهای موجود در کتابخانه ملی فهرستگان پیوسته کتابخانه های ،دیگر و برمبنای استفاده از اطلاعات پایگاه کتابشناختی ملی (فهرستگان مشترک اطلاعات کتاب شناختی منابع مربوط به بیش از 850 کتابخانه استرالیا به همراه فراهم آوری امکان دست یابی به منابع توسط خدمات مرورگر کین تکا (1)، فهرست نویسی گردیده است) تصمیم در اتخاذ چنین سیاستی از آن جهت صورت گرفته که امکان شناخت و استخراج منابع برخط یکپارچه سازی شده با دیگر منابع کتابخانه ای، امری بسیار مهم تشخیص داده شده است با وجود ،این محرز است که انجام چنین اقداماتی بر هزینه های مربوط به ایجاد آرشیو منابع تحت وب می افزاید.

عامل مهم و یاری کننده دیگری که به صرفه جویی زمانی و استفاده بهینه از اوقات کاری کارکنان منجر می شود شرایط مطرح در قوانین وا سپاری کشور استرالیاست. در بسیاری از ایالت های استرالیا، حتی آن هایی که مشترک المنافع هستند نیز قانون و اسپاری منابع و از آن جمله منابع برخط همچنان مهجور مانده و به تصویب نرسیده است تنها در بخش های شمالی این کشور اخیرا قانونی به تصویب رسیده است که به وضوح امکان واسپاری منابع برخط را میسر می سازد بدان معنا که تمامی کتابخانه های همکار در این طرح از جمله کتابخانه ملی باید قبل از تهیه نسخه ای از یک منبع نسبت به کسب اجازه از ناشر آن ،اقدام سپس آن را برای استفاده همگان در آرشیو دسترس پذیر کنند.

نیروی انسانی و افراد شاغل در پاندورا

ایجاد آرشیو پاندورا اجرا و گسترش برنامه های کاربردی وابسته به آرشیو حفظ و نگهداری سیستم های مربوط، تخصص گرایی و برنامه ریزی به منظور محافظت بلند مدت از آن مستلزم به کارگیری نیروهای کاری شش واحد در دو بخش کتابخانه ملی است

کارکنان شاغل در آرشیو دیجیتال بخش مجموعه سازی مسئول انتخاب منابع موجود در آرشیو و محتوای آن می باشند واحد برنامه های کاربردی بخش فناوری اطلاعات نیز مسئول توسعه زیر ساخت های فنی آرشیو هستند بخش مربوط به خدمات وبگاه مسئولیت توسعه رابط کاربر را برای هر دو سیستم پاندورا و پانداس بر عهده دارد و نظام های حمایتی سیستم های تجاری خود عاملی جهت حفظ تولیدات

1- Kinetica

آزمایش آموزش و ارزیابی محسوب می شوند و موجب می گردند اهداف مذکور به خوبی اجرا و محقق .شود افزون بر ،این کارکنان واحد خدمات حفاظتی بخش مجموعه سازی مسئولیت دسترس پذیر سازی طولانی مدت محتوای آرشیو را بر عهده دارند.

در سال 2004 ، در زمان تخمین و برآورد هزینه های زیر سهم کمک های کتابخانه ملی به آرشیو پاندورا در قالب استفاده از نیروهای تمام وقت به شرح سطوح زیر بود:

شعبه آرشیو دیجیتال:

- 1 نفر نیروی تمام وقت در سطح مدیر - کتابدار
- 2 نفر نیروی تمام وقت در سطح یک نفر سرپرست یک نفر کارشناس پروژه های خاص –
- کتابداران
- 4 نفر نیروی تمام وقت در سطح کارکنان ستادی – کتابداران؛ و
- 17 نفر نیروی تمام وقت در سطح کارشناس رفع مشکلات فنی – دارای 5 سال سابقه کار در حوزه فناوری اطلاعات
- 2 نفر مدیر تمام وقت که 10 درصد از وقت خود را هر کدام صرف توسعه و نگهداری سیستم کنند؛

و

- 2 نفر مدیر تمام وقت که 25 درصد از وقت خود را هر کدام صرف توسعه و نگهداری سیستم نمایند. همان طور که در بالا ذکر شد نیروی انسانی شاغل در بخش حفاظت از منابع آرشیوی، به صورت تمام وقت در این طرح حضور داشتند، با وجود این هزینه های مربوط در این برآورد منظور نگردیده است.

هزینه دستیابی به وب گاه ها و نشریات برخط

گر چه برای کتابخانه کاملاً مشخص بود که هزینه سرانه ایجاد واحد آرشیو منابع برخط در مقایسه با منابع چاپی نظیر کتاب ها و پیایند ها بالا- می باشد؛ با این حال، تا همین اواخر هیچ گونه اطلاعات جامع و دقیقی در خصوص هزینه های ایجاد آرشیو تحت وب در اختیار نداشتیم. در سال 2004، کتابخانه تصمیم گرفت تا هزینه های مربوط به فعالیت های واسپاری منابع پیایندی و تک نگاشت ها را برآورد کند. افزون بر این، از آن جا که به نظر می رسید گردآوری منابع برخط استرالیا نیز امتداد بسط همان مسئولیت های مذکور باشد، تصمیم گرفته شد هزینه اجرای این کار نیز به همان روال برآورد شود.

محدوده برآورد هزینه

مطالعات مربوط به برآورد هزینه به بررسی و مطالعه هزینه های مربوط به فراهم آوری منابع کتابخانه ملی به عنوان یک «نمونه» (1) پرداخته سپس آن را به هزینه های پاندورا می افزاید افزون بر این مرز میان هزینه هایی که باید لحاظ یا حذف شوند نیز مشخص می گردد.

Instance -1

هزینه های مستقیم عبارت بودند از:

- هزینه های کارکنان بخش آرشیو دیجیتال کتابخانه ملی؛
- هزینه های اداری بخش آرشیو دیجیتال از قبیل ،مسافرت ،آموزش شرکت در کنفرانس و
- تجهیزات اداری (شامل هزینه های فرد تأمین کننده)؛ و
- هزینه های مربوط به توسعه زیر ساخت ها و نگهداری از آن ها نظیر هزینه کارکنان بخش فناوری اطلاعات و خرید تجهیزات سخت افزاری و نرم افزاری

هزینه هایی که باید حذف شوند عبارت بودند از:

- هزینه های غیر مستقیم مانند تأمین ایستگاه های کاری کارکنان
- نورپردازی و محافظت از ساختمان و
- هزینه های مربوط به محافظت از محتوای منابع موجود در آرشیو

در برآورد هزینه ،حاضر فقط هزینه های انجام شده توسط کتابخانه ملی در نظر گرفته شد بدین معنی که هزینه های مربوط به استخدام کارکنان آرشیو که سازمان های شریک در نظر داشتند، حذف گردید.

روش شناسی انجام کار

جهت برآورد هزینه کارکنان اقدام به تهیه نموداری از تمامی فعالیت ها و فرآیندهای (وظایف صورت گرفته توسط بخش آرشیو دیجیتال) مورد نیاز جهت کسب شاخصی در زمینه ایجاد آرشیو کردند.در نمودار مزبور، هزینه های (فعالیت های) اساسی مشخص شده بود ،سپس کارکنان اقدام به تخمین مدت زمانی کردند که به طور متوسط هر یک از کارکنان در هر روز صرف می کردند یک روز کاری شامل 441 دقیقه است

- شناسایی و انتخاب مواد و منابع - 30 دقیقه
- تماس با ناشر، مذاکره برای کسب اجازه جهت افزودن منبعی به آرشیو و بایگانی مکاتبات – 30 دقیقه
- گردآوری تضمین کیفیت و آرشیو منابع – 210 دقیقه
- فهرست نویسی – 81 دقیقه
- فعالیت های دیگر (شامل انجام مکاتبات با افراد و نهاد های مسئول در امر نمایه سازی و چکیده نویسی، سوالات مرجع آرشیو مدار و همکاری کارکنان آرشیو دیجیتال برای بسط و توسعه پانداس) – 60 دقیقه

• برقراری ارتباط با شرکا و نیز پشتیبانی – 30 دقیقه (این فعالیت در برآورد هزینه منظور نشده است) البته تمامی کارکنان کلیه وظایف فوق را به یک نسبت انجام نمی دهند برخی بیش تر و برخی کم تر کار می کنند برای مثال سرپرست بخش عهده دار انجام وظایف اداری و نظارت بر کارکنان نیز می باشد این بدان معنی است که او کم تر از دیگران به کارهایی چون گردآوری ،آوری کنترل کیفیت، آرشیو سازی،

و فهرست نویسی می پردازد.

از میان مجموعه وظایفی که در بالا اشاره شد، مدیر بخش ایجاد آرشیو دیجیتال، تنها متعهد به حمایت از شرکا و برقراری تماس با ناشران است اوقات وی بیش تر صرف ،مدیریت اعمال سیاست گذاری های ،توسعه تبلیغات و ارتباط با دیگر سازمان های فعال در امر ایجاد آرشیو تحت وب است. با وجود این حقوق و دستمزد وی جزء ضروری هزینه های کلی ایجاد آرشیو دیجیتال کتابخانه محسوب می شود و باید در برآورد کلی هزینه ها منظور شود.

در این میان کارمندانی نیز عهده دار انجام اموری هستند که کاملاً از حیطه وظایف آن ها خارج است. برای مثال، دو نفر از کتابداران هر هفته باید وقت خود را صرف میز مرجع موجود در اتاق مطالعه کند، و باید این مدت زمان را از مجموع مدت زمان حضور آن ها در آرشیو کسر کرد.

کتابخانه ملی به عنوان شریک اصلی پاندورا و تأمین کننده زیر ساخت های فنی دارای نقشی حمایتی نسبت به دیگر شرکاست و این ،امر به نوبه خود به صرف هزینه های بیش تری می انجامد. تمامی کارکنان بخش آرشیو دیجیتال مقادیر قابل توجهی از وقت خود را صرف برقراری ارتباط با دیگر همکاران و ارائه خدمات فنی پشتیبانی کننده به آن ها می کند کارمندان ،ستادی هر روز به طور متوسط 30 دقیقه از وقت خود را صرف انجام این کار ،یعنی برقراری ارتباط با دیگر همکاران و حمایت از آن ها می کند، در حالی که همزمان مدیر آن ها تنها 10 دقیقه از وقت خود را در هر روز صرف انجام این مهم می کند. لازم به ذکر است که هزینه های مربوط به موارد فوق در صورت هزینه های برآورد شده منظور نشده است.

نحوه محاسبه هزینه ها

پس از برآورد هزینه ها در مرحله بعد با استفاده از برنامه اکسل روشی برای محاسبه هزینه های مربوط به هر یک از فعالیت های انجام گرفته در آرشیو، طراحی و تدوین شد.

سپس در هر روز مقادیر مربوط به حقوق و دستمزد کارکنان آرشیو دیجیتال و نیز مقادیر زمانی صرف شده توسط آن ها در خصوص انجام هر کار وارد برنامه صفحه گستر اکسل شد و در نهایت، کل تعداد دقایق صرف شده برای انجام هر کار و نیز هزینه های مربوط به هر کارمند، محاسبه گردید.

سپس هزینه های تأمین کنندگان مختلف به شرحی که پیش تر گفته شد نیز اضافه گردید، و هزینه های مربوط به توسعه زیر ساخت های نگهداری نیز در این مرحله لحاظ شدند.

در مجموع می توان گفت که از کل 937 موردی که توسط کتابخانه ملی در ماه های ژوئیه تا اکتبر سال 2004 وارد آرشیو شد به طور متوسط در هر روز 13 عنوان به ثبت رسید.

هزینه های فراهم آوری منابع آرشیوی

با ورود تمامی این اطلاعات در برنامه صفحه گستر اکسل مشخص شد که هزینه های تمام شده برای هر منبع آرشیو شده، صرف نظر از هزینه های مربوط به انجام فعالیت هایی چون برقراری ارتباط با همکاران و حمایت از آن ها رقمی معادل 178/68 دلار استرالیا است.

اطلاعات مربوط به هر یک از اجزای تشکیل دهنده

ص: 212

هزینه های مربوطه در زیر آمده است:

- هزینه های مربوط به استفاده از نیروی انسانی جهت آرشیو هر منبع در آرشیو دیجیتال – به ازای هر مورد 168/36 دلار استرالیا؛
- هزینه های فراهم کنندگان اقلام آرشیوی به ازای هر مورد 3/41 دلار استرالیا؛ و
- هزینه های مربوط توسعه زیرساخت ها و تعمیر و نگهداری اقلام آرشیوی به ازای هر مورد 6/791 دلار استرالیا

همان گونه که ملاحظه می کنید، حقیقت تلخ پر هزینه بودن این کار سخت و پرمشقت، در رویکرد گزینشی ایجاد آرشیو وب به خوبی نمایان و قابل مشاهده است. هزینه های مربوط به نیروی انسانی، 94 درصد کل هزینه ها را تشکیل می دهد

مقایسه با نوع چاپی

تفاوت هزینه های بالای دستیابی به انتشارات ،وبی در مقایسه با نشریات چاپی را همچنین می توان از قیاس میان هزینه های مربوط به تهیه این گونه مواد (وبی) و موارد مشابه چاپی (شامل پیاپند ها و تک نگاشت ها) که به طور همزمان از طریق قانون و اسپاری فراهم گردیده اند، احراز نمود.

- هزینه فراهم آوری تک نگاشت ها با استفاده از قانون واسپاری منابع چاپی به ازای هر مورد 12/29 دلار استرالیا؛ و
- هزینه فراهم آوری پیایندهای چاپی با استفاده از قانون واسپاری منابع چاپی به ازای هر مورد 11/29 دلار استرالیا.

تلاش برای مقایسه هزینه های مزبور اندکی شبیه تلاش برای مقایسه هزینه های مربوط به حمل و نقل هندوانه موز و انگور در بازار در حالت عمده در قیاس با قیمت تمام شده برای هر یک از این نوع اقلام است. در حالت عادی قیمت تمام شده هر یک از این اقلام با قیمت عمده آن ها برابر نیست. در قیاس با این مسئله، قیمت و هزینه تمام شده برای تهیه منابع تحت وب، با قیمت تمام شده تهیه هریک از این مواد در حالت چاپی نیز برابری .ندارد گر چه اختلاف فاحشی میان ماهیت مواد مزبور (منابع تحت وب) و فرآیند تهیه آن ها با نوع چاپی وجود دارد کاملاً آشکار است که در کل، هزینه تهیه و فراهم آوری منابع تحت وب در مقایسه با منابع چاپی بالا تر می باشد

باید این جا به این نکته توجه داشت که در میان اعداد و ارقام موجود به هیچ عنوان هزینه های مربوط به خرید منابع منظور نشده .است ،کتابخانه منابع چاپی را بر اساس طرح واسپاری و بدون پرداخت هزینه ای دریافت می کند منابع تحت وب نیز بدون پرداخت هزینه از طریق وبگاه ناشران قابل دستیابی است. این هزینه ها صرفا مربوط به فرآیند خرید منابع است هزینه های مربوط به قفسه آرایی و آماده سازی اقلام بر روی قفسه ها ،منظور ،اما هزینه های بعد آن نظیر و جین منابع و مراقبت از مجموعه در محاسبات مربوط به برآورد هزینه ها لحاظ نمی شود.

هزینه فعالیت های خاص

افزون بر این کتابخـانه ملی علاـقه منـد به تحلیل هزینه های مربوط به ارائه خـدمات خاص نیز بود از این رو جزئیات هزینه های مربوط به استفاده از افراد به ازای انجام هر یک از موارد کاری زیر عبارت است از:

- شناسایی و انتخاب مواد و منابع – 10/16 دلار استرالیا
- تماس با ناشر، مذاکره برای کسب اجازه جهت افزودن منبعی به آرشیو و بایگانی مکاتبات– 10/34 دلار استرالیا
- گردآوری تضمین کیفیت و آرشیو منابع – 7/09 دلار استرالیا،
- فهرست نویسی – 27/42 دلار استرالیا، و
- دیگر فعالیت ها (نظیر بسـیاری از فعالیت های مدیریتی انجام مکاتبات با افراد و نهاد های مسـئول در امر نمایه سازی و چکیده نویسـی پرسش های مرجع آرشیو مدار و همکاری کارکنان آرشیو دیجیتال برای بسط و توسعه پانداس – 59,67 دلار استرالیا.

توجه داشـته باشید که این برآورد هزینه چندان نیز مفید ،نیست زیرا این برآورد هزینه تنها به شکل موردی و در سطح فهرست نویسی ساده و ابتدایی (توصـیفی) صورت می پذیرد نه در سـطحی تحلیلی و محتوایی در حالی که هر عنوان موجود در آرشـیو دارای حداقل دو سـطح فهرست نویسی است. به این ترتیب قیمت واقعی تر رقمی معادل 54/84 دلار استرالیا می باشد.

امکان کاهش هزینه ها

بـا اطلاعات حاصل از این روش (تخمین هزینه فعالیت های صورت گرفته) توانستیم دریابیم که چگونه می توان هزینه ها را کاهش .داد ما توانستیم هزینه های خود را با ایجاد تغییر در رویکرد و اقدام جهت ایجاد آرشـیو تحت وب (تغییر سـیاست های کاری خود) با یافتن شیوه هایی جهت انجام مؤثرتر وظایف خود در این زمینه و با اعمال همزمان هر دو روش کاهش دهیم برای دستیابی به آینده ای قابل پیش بینی، تلاش خواهیم کرد تا تـدابیر و سـیاست های موجود در این عرصه برای مثال تدوین فهرسـتگان اطلاعات کتاب شـناختی منابع یا تضـمین کیفیت را حفظ کنیم. البته موقعیت های دیگری نیز برای صرفه جویی در هزینه ها وجود دارد.

شناسایی و انتخاب عناوین، اقدامی مهم و اساسـی برای کاربرد رویکرد گزینشـی در ایجاد آرشیو دیجیتال است. در نگاه اول و در نخستین گام مشکل به نظر می رسد که در یابیم چگونه می توان مدت زمان صرف شده برای انجام این قبیل فعالیت ها را کاهش داد به هر حال ما با ناشـران دولتی از آن جهت همکاری می کنیم که بتوان از طریق آن ها به ابر داده های مربوط به اطلاعات کتاب شـناختی منابع برخط که به صورت گروهی توسط آن ها در پانـداس بارگـذاری شده و امکان دستیابی خودکار به منابع برخط را فراهم می سازد – دست یابیم در واقع ناشـران آن چه را که پانـدورا آرشـیو خواهـد شد – به واسـطه تأمین ابر داده های مربوط به ،آن ها مشـخص می نمایند. پانداس همچنان در صـدد است تـا از این طریق امکـان پردازش و دستیابی به منابع اطلاعاتی را به شـکلی انبوه برای همگان محقق سازد با وجود این، این امر زمانی محقق

خواهـد شد که قیمت متوسط فراهم آوری منابع اطلاعاتی توسط سازمان های همکار کاهش یابد. هم اکنون فراداده ها توسط تعداد اندکی از ناشـران به طور خودکار به رکوردهای مارک ارسال می شوند تا ضـمن افزوده شدن به پایگاه داده های کتابشناختی ،ملی به طور همزمان در فهرستگان برخط کتابخانه نیز بارگذاری شوند این امر به کاهش هزینه ها خواهد انجامید.

محاسـبات بیشتر در خصوص جزئیـات مربوط به هر فعـالیت (جمع آوری، تضـمین کیفیت، و آرشـیو منـابع) بیان گر آن بود که تضـمین کیفیت 86 درصد کل هزینه ها را تشـکیل می دهد. چنان چه می توانستیم از نرم افزارهای قابل اعتماد در زمینه ارزیابی کیفیت در پانداس استفاده کنیم صرفه جویی در هزینه ها بسیار با ارزش تر و معنی دارتر می شد

افزون بر این کتابخـانه مزبور تلاـش می کنـد با انجام واسـطه گری و لابی های مناسب، دولت استرالیا را متقاعـد به بسط و گسترش امکان اسـتفاده از قانون واسـپاری منابع برخط ،کنـد تا از این طریق بتواند نسـبت به رفع موانع موجود در زمینه ضـرورت اخذ مجوز از ناشـران به منظور تهیه نسخه منابع آرشیوی اقدام نماید. بدیهی است تحقق این امر موجب کاهش زمان کاری کارکنان آرشیو خواهد شد.

بهره گیری از فناوری به پیشـرفت این امر کمک خواهد کرد، و ما را قادر خواهد ساخت که هزینه های مربوط به انجام این فعالیت را کاهش دهیم کنسرسـیوم بین المللی محافظت از منابع اینترنتی که کتابخانه ملی نیز از اعضای فعال آن می باشد، در صدد تهیه مجموعه ای از ابزار های لازم جهت ایجاد آرشیو وب نظیر میانجی کاربر به طور اخص طراحی شده برای کتابخانه ملی .است انتظار میرود تحقق این امر نیز به کارآمد شدن این فعالیت بیانجامد.

نتیجه گیری

محاسبه هزینه ها، دانسته های کنونی ما را در خصوص ایجاد آرشیو تحت وب و این که این نوع آرشیو در مقایسه با شیوه فراهم آوری منابع چاپی فعالیتی پر هزینه بوده (حتی اخیراً هزینه های آن به دلیل افزایش هزینه های نیروی کار نسـبت به گذشته فزونی یافته است) تأیید کرد. علاوه بر این، مطالعه حاضـر اطلاعات سودمندتری را درباره چگونگی ارتقای برنامه ها در اختیار ما قرار داد آیا ما باید از سـیاست واحدی در زمینه فهرست نویسـی عناوین آرشـیوی اسـتفاده کنیم؟ بله شاید از این طریق بتوانیم راه های کم زحمت تری را جهت ثبت پیشینه های کتـاب شـناختی منـابع به دست آوریم تا چه میزان بایـد وقت و انرژی خود را صـرف فرآینـدگران قیمت و پرهزینه تضـمین کیفیت نماییم؟ مسلماً ،خیلی اما بهره گیری از یک نرم افزار ارزیابی مناسب می تواند بسیار سودمند باشد.

بـدیهی است افزایش پختگی در زمینه بهره گیری از فناوری های مربوط به ایجاد آرشـیو تحت وب، منجر به افزایش کارآمـدی تمامی پروژه های مربوط به این زمینه (ایجاد آرشـیو تحت وب)، نظیر، آرشـیوهایی که اخیراً اقدام به استفاده از رویکرد پر زحمت گزینشی کرده و موفق شده اند از هزینه های سرانه پایین تری در یک بازه زمانی برخوردار گردند خواهد شد.

قدردانی

مـایلم در اینجـا اعلام کنم که بیش ترین زحمات در برآورد هزینه ها بر عهـده ناظم یوسف، مـدیر مالی کتابخانه استرالیا بوده است. روش شناسی این کار شامل تخمین هزینه ها و تفسیر نتایج مربوط به آن ها نیز بر عهده وی بوده است.

منابع

Library and Archives Canada. Electronic Collection: a Virtual Collection of Monographs and Periodicals. .1 http://www.collectionscanada.ca/ electroniccollection/ (accessed 22 March 2005); National Diet Library of Japan. Web Archiving Project (WARP). http://warp.ndl.go.jp/ (accessed 22 March 2005); National Library of Australia. www.nla.gov.au 9|10 Creative Commons Attribution-NonCommercial-ShareAlike 2.1 Australia 19 March 2009 Staff paper Selective Archiving of Web Resources 10 | 10 www.nla.gov.au 19 March 2009 Creative Commons Attribution-NonCommercial-ShareAlike 2.1 AustraliaPANDORA, .(Australia's Web Archive. http://pandora.nla.gov.au/index. html (accessed 22 March 2005

.(National Library of Sweden. Kulturawa3. http://www.kb.se/kw3/ (accessed 22 March 2005 .2

.(Library of Congress. MINERVA. http://www.loc.gov/minerva/ (accessed 22 March 2005 .3

State Library of Tasmania. STORS: Long Term Storage of Tasmanian Electronic Documents. . 4 http://www.stors.tas.gov.au/ (accessed 22 March 2005); Oltmans, E. and H. van Wijngaarden. 2004. Digital .preservation in practice: the e-Depot at the Koninklijke Bibliotheek. Vine 34 (1): 21-26

Information about PANDORA, Australia's Web Archive, and access to its contents are available at . 5 .http://pandora.nla.gov.au/index.html

.Information about PANDAS is available at http://pandora.nla.gov.au/pandas. html .6

.(7National Library of Australia. Kinetica. http://www.nla.gov.au/kinetica/ (accessed 22 March 2005 .7

Levels and pay scales are explained in Attachment A – Salary Table of the .8

.National Library of Australia Certified Agreement 2004–2007 available here

An "instance" is a single gathering of a title. It includes the gathering of a monograph that has been archived .9 once only, the first gathering of a serial or integrating title (for example, a website that changes over time), .and all subsequent gatherings

International Internet Preservation Consortium http://netpreserve.org/about/ index.php (accessed 22 . 10 .(March 2005

ص: 217

چکیده

بایگانی اینترنت اقدامی پیشگامانه و تلاشی ابتکاری برای بایگانی منابع وب انجام داده است. معمولا جست وجوی این نوع بایگانی ها با استفاده از فراداده های توصیفی یا خدمات نمایه سازی متن ،کامل امکان پذیر نیست. مسائل بسیاری وجود دارد که بایگانی وب را مشکل می کند و اگر برنامه ریزی سازمانی، برای بایگانی وب از محدوده کوچک تری آغاز شده باشد و بیش تر محدودیت های آن محدویت های بشری یا محدودیت های منابع سخت افزاری باشد، به برخی مسائل باید با رویکردی متفاوت نگاه کرد. برای نشان دادن برخی مسائل اصلی، بایگانی رقومی برای مطالعه زبان چینی (SHCAD) به عنوان مطالعه موردی در مقاله حاضر مورد بررسی قرار گرفته است.

بایگانی وب علمی در مقیاس کوچک

اشاره

(DACHS (1

هانو الشر (2) ا :ترجمه حمزه علی نور محمدی (3)

1. چرایی بایگانی علمی در مقیاس کوچک

با توجه به پیچیدگی های بایگانی وب و نیاز های سخت افزاری و نرم افزاری و همچنین تخصص و پرسنل آیا چنین پروژه هایی فقط برای مؤسسه هایی در مقیاس بزرگ مانند کتابخانه های ملی امکان پذیر است؟ یا اینکه مؤسسه های کوچک تر مانند موزه ها دانشگاه ها و مانند آن ها نیز قادر به انجام وظایف مورد نیاز برای بایگانی وب- با چشم انداز بلند مدت قادر به این کار خواهند بود؟

حتی اگر پاسخ مثبت باشد این سؤال باقی می ماند که آیا در واقع این کار ضروری است یا خیر؟ به هر حال می توان فکر کرد که در حال حاضر بایگانی اینترنت همراه با تلاش برای افزایش تعداد کتابخانه های ملی بسیاری از منابع وب را تحت پوشش دارد و بایگانی می کند.

اجازه دهید با سؤال دیگری شروع کنم بایگانی اینترنت اقدامی پیشگامانه را به عنوان نخستین تلاش ابتکاری ،جامع جهت بایگانی منابع وب انجام داده است. موفقیت در انجام این کار، انقلابی بوده و بنیاد و اساس پروژه های بسیاری را پی ریزی کرده است با این حال، بررسی آن چه که بایگانی اینترنت و پروژه های جامع دیگر می توانند به آن دست یابند جهت کشف برخی محدودیت ها، آسان است. از آن جا که اساس

1- Small Scale Academic Web Archiving: DACHS: in Masanes, Julien (ed.), Web Archiving. Berlin. Heidelberg New York: Springer.pp.213-224.

2- Hanno E. Lecher

3- استادیار دانشگاه شاهد

کار مجموعه، بسیار گسترده است، باید برای بخش بزرگی از فعالیت ها برای جمع آوری اطلاعات وب ها به صورت خودکار، تا آن جا که امکان پذیر است به روبات ها تکیه کرد این نوع جمع آوری اطلاعات غالباً بسیار سطحی است بخش های زیادی از دست رفته بسیاری از صفحات به طور ناقص بارگذاری می شوند و بعضی از انواع فایل و نیز وب های مخفی به طور کلی نادیده گرفته خواهند شد.

علاوه بر این، از آن جا که دریافت به طور خودکار انجام می شود در فواصل زمانی نامنظم نمی توان انتخابی آگاهانه از منابع داشته باشیم امکان در نظر گرفتن یا تشخیص مطالب مهم موجود – که ممکن است عمر کوتاهی داشته و یا به سختی شناسایی شوند – وجود ندارد ،البته بایگانی اینترنت و سایر پروژه های بایگانی وب در مقیاس بزرگ برخی امور خاص را انجام داده اند که در آن تلاش زیادی صرف توسعه مجموعه ها در محدوده خاصی از موضوع هایی شده است که انتخاب کرده اند. در هر صورت، تعداد این موضوع ها بسیار محدود است و طرح های پژوهشی زیادی باید آرشیوهای خود را توسعه دهند.

بیش ترین مشکل در پروژه های جامع بایگانی ،وب دسترسی محدود به محتواست. مهم ترین اصل شناسایی نشانی دقیق اینترنتی یک سند یا وبگاه است تا بازیابی اطلاعات مقدور شود. معمولاً جست وجوی این نوع بایگانی ها با استفاده از فراداده های توصیفی یا خدمات نمایه سازی متن کامل، امکان پذیر نیست. حتی اگر گزینه جست و جوی متن کامل در دسترس باشد عدم انتخاب آگاهانه منابع، تنها نتایج نامنظمی شبیه جست وجوی وب امروزی ارائه می دهد

با نگاه به این محدودیت ها بدیهی است که بایگانی اینترنت و برخی منابع دیگر، نه بایگانی اینترنت را به طور کامل انجام می دهند و نه راه های دسترسی مناسبی برای بسیاری از اهداف علمی و یا پژوهشی دیگر فراهم می کنند. در نتیجه بایگانی علمی در مقیاس کوچک به یک نیاز مهم تبدیل شده است. اما آیا بازگشت به پرسش مطرح شده در ابتدا امکان پذیر است؟

به برخی مسائل باید با رویکردی متفاوت نگاه کرد اگر برنامه ریزی سازمانی، برای بایگانی وب از محدوده کوچک تری آغاز شده باشد و بیش تر محدودیت های آن محدویت های بشری/ محدودیت های منابع سخت افزاری باشد برای نشان دادن برخی مسائل اصلی بایگانی دیجیتال برای مطالعه زبان چینی (DACHS) به عنوان مطالعه موردی مورد بررسی قرار خواهد گرفت. هر چند باید در ذهن داشته باشید که پروژه های متفاوت غالباً به روش ها یا راه حل های متفاوتی نیاز دارند.

2.بایگانی دیجیتال برای مطالعات زبان چینی

اشاره

اهداف اصلی بایگانی دیجیتال مطالعات زبان چینی (DACHS)، اطمینان از دسترسی درازمدت برای شناسایی و بایگانی منابع اینترنت مربوط به مطالعات زبان چینی است انتخاب نقش مهمی در این فرآیند دارد و به طور ویژه در گفتمان سیاسی و اجتماعی بر آن تأکید و در اینترنت چینی منعکس شده است.

در حال حاضر پروژه DACHS توسط کتابخانه های دو مؤسسه چین شناسی با نام های «مؤسسه چین» در شهر هاید لبوگ آلمان و «مؤسسه چین شناسی» در دانشگاه لیدن هلند به کار گرفته شده است. بدین ترتیب، زیر ساختی کاملا متفاوت با کتابخانه های بزرگ ملی دارد (جدول 1)

عکس

جدول 1- حجم ذخیره مجموعه Dchs

نخستین بار هنگامی که اندیشه بارگذاری منابع برخط در موردچین در اواخر 1999 در هایدلبرگ مطرح شد، به هیچ وجه وضعیت آن برای کسی روشن نبود این اندیشه هنوز توسعه نیافته بود و به عنوان بخشی از برنامه ای بزرگ تر برای ایجاد مرکز اروپایی منابع دیجیتال در مطالعات چینی، با هدف بهبود شرایط مربوط به تحقیقات چین و دسترسی به اطلاعات در اروپا، معرفی شد. این پروژه، شامل فعالیت هایی همچون خرید طیف وسیعی از پایگاه داده های تجاری تمام متن حمایت از توسعه پروژه های پایگاه داده دانشگاهی و توسعه امکانات بود ،همچنین توسعه برای یافتن بیماری ایدز در منابع چاپی و غیر چاپی در چین و فراهم آوردن امکان دسترسی آزاد به تمام متن منابع به طور گسترده را شامل می شد. این پروژه به مدت پنج سال در نظر گرفته شد و امکانات مالی آن برای بهبود وسایل سخت افزاری موجود و برای استخدام برخی کارکنان – به عنوان دستیاران دانشجویی – صرف شد.

دستور عمل اصلی این پروژه «حداکثر انعطاف پذیری با حداکثر پاسخگویی» بود. این امر به این معنی است که فضایی برای توسعه طرح های تفصیلی بسیاری از پروژه های فرعی و حتی ایده هایی برای پروژه های فرعی جدید مورد نیاز وجود دارد.

در سال 2000، بیشتر به جذب پروژه های فرعی برای اجرای برنامه ریزی پیشرفته و سازماندهی مجدد زیر ساخت های فناوری اطلاعات مؤسسه پرداخته شد، و برنامه ریزی واقعی برای بایگانی وب در سال 2001 آغاز گردید.

2-1.گام های اولیه

با نگاهی به زیر ساخت های موجود که بایگانی وب باید از طریق آن توسعه داده شود، محدودیت ها قابل مشاهده است. بایگانی دیجیتال برای مطالعات زبان چینی توسط کتابخانه مؤسسه به اجرا در آمد. در آن زمان ،کتابخانه با استفاده از فناوری اطلاعات با چهار سرور و نزدیک به 100 ایستگاه کاری [جریان کار] نظارت داشت. مسئول نگهداری این زیر ساخت های فناوری اطلاعات کتابداران با کمک یک یا دو دستیار دانشجویی پاره وقت بودند و در صورت نیاز از سوی بخش آی. سی. تی دانشگاه پشتیبانی می شدند.

ص: 221

برای پروژه بایگانی وب باید یک دستیار دیگر دانشجویی به صورت پاره وقت برای تهیه برخی امور (بارگذاری کردن بایگانی ایجاد ابر داده ها و مانند آن) استخدام می شد؛ زیرا کتابدار – علاوه بر مسئولیت های خود برای کتابخانه و محیط فناوری اطلاعات – باید کارهای مدیریت پروژه و توسعه سایر امور نیز در نظر داشته باشد در واقع سهم دستیار پروژه برای توسعه DACHS مهم است. مدرک تحصیلی بالای دستیار برای جلوگیری از تمرکز زدایی دانش در مورد چارچوب نظری پروژه ها اهمیت .دارد به خصوص اینکه وقتی کتابدار مؤسسه را جهت کار دیگری محل را ترک کند، اهمیت آن روشن خواهد شد.

البته برخی مسائل، با توجه به اندازه مؤسسه و امکانات آن باید در آغاز پروژه در نظر گرفته شود و به برخی سؤال ها نیز پاسخ داده شود مانند هدف دسترسی دراز مدت منابع بایگانی شده چیست؟ الزامات و نیازهای سخت افزاری و نرم افزاری برای ایجاد و نگهداری بایگانی چه چیزهایی است؟ چگونه انتخاب منابع باید به عنوان یک کار در حال انجام سازماندهی شود و چگونه باید اطلاعات ایجاد شده در دسترس قرار گیرد؟ و مهم تر از همه چه چیز دیگری نیاز است که برای یک برنامه ریزی مناسب در نظر گرفته شود و کجا در جست وجوی پاسخ باشیم؟

پاسخ به آخرین سؤال را می توان در سندی که در سال 2003 به چارچوب استاندارد بایگانی وب تبدیل شده پیدا کرد. 14721:2003 ISO که با عنوان OAIS (سیستم اطلاعات آرشیوی باز) شناخته می شود ثابت کرده است که از اهمیت حیاتی برخوردار است زیرا در آن پس زمینه های تئوری بسیار مهمی تعیین شده است و کمک می کند تا بسیاری از مسائل مهم بایگانی وب را در آن ببینیم.

این سند مفید یک اشکال عمده نیز دارد: همان طور که از نام آن پیداست، فقط چارچوبی برای راهنمایی های تئوری در مسائل مختلف است و پیاده سازی واقعی مهارت های کاربر را ارائه نمی دهد ،بنابراین لازم است جست و جوی دیگری برای درک چگونگی تبدیل چارچوب به عمل انجام شود.

اطلاعات بسیاری می توان در جاهایی مانند ،RLG PADI و مانند آن یافت اما مهم تر از آن مشارکت فعالانه در کارگاه ها و همایش هایی است که در برخورد با مسائل مرتبط با بایگانی وب ارائه می دهند.

2-2. توسعه پایدار سازمانی

یکی از مهم ترین سؤال هایی که کل پروژه با آن روبه روست چگونگی فراهم کردن توسعه پایدار مؤسسه برای بایگانی دراز مدت است. هیچ یک از سه عامل اساسی در این پرسش نمی توانند به صورت دراز مدت در نظر گرفته شوند تأمین مالی پروژه نمی تواند محدود شود و به پایان برسد؛ علاقه مندی های کاربران در سازمان می تواند تغییر کند و منجر به غفلت از پروژه شود؛ و حتی وضعیت دراز مدت سازمان به هیچ وجه تضمین شده نیست. بنابراین واضح است که سازم آن جایی با قابلیت اعتماد به مراتب کم تری از کتابخانه یا بایگانی ملی برای بایگانی دراز مدت است که در آن برخی مفاد قانونی می تواند مؤسسه را برای تحقق مسئولیت ها نسبت به مجموعه مجبور می کند.

بنابراین، راهبرد توسعه باید برای بقای بایگانی ،فنی در صورت توقف کار مؤسسه یا عدم توانایی آن برای حمایت از پروژه .باشد بقا را می توان به دو روش تعریف کرد: یا باید بایگانی را فعال نگه داشت بدین معنی که تمام فعالیت های انتخاب منابع برای داده های در دسترس ادامه داشته باشد؛ یا این که بایگانی باید حداقل در یک وضعیت رکودی حفظ شود، بدین معنی که با وجود این که منابع جدید به بایگانی اضافه نمی شوند حداقل دسترسی به آن چه که در حال حاضر موجود است تضمین شود.

دو راه برای این کار وجود دارد نخست و مهم تر از همه بایگانی باید ویژگی های اولیه یک منبع قابل اعتمادی را محقق کند، که به معنای پایبندی به استانداردهای اعلام شده است همان گونه که در مدل OAIS توصیف شد. در عمل استفاده از آن را برای مؤسسه های دیگر ممکن سازد – حالت ایده ال کتابخانه های ملی – البته تا زمانی که قادر به همسانی آن با سایر بایگانی ها نباشیم

راه دوم توسعه بایگانی است که برای آن تلاش شده است. اگر تعدادی از مؤسسه ها به طور فعال و تعاملی در این پروژه مشارکت داشته باشند در زمان توقف کار برای یکی از شرکا برای دیگری این امکان به وجود می آید که کار بایگانی را ادامه داده و از امکانات آن استفاده کند. در عمل، پایبندی به استانداردهای برقرار شده در این نوع کار ضروری به نظر می رسد برای این پروژه نیز تصمیم گرفته شد که به صورت مشارکتی کار شود [پروژه] در جست و جوی شرکایی است که امکان همکاری در این زمینه را داشته باشند و مسائل مربوط را حل کنند و [در نهایت] بایگانی به عنوان یک کمک اساسی به حوزه مطالعات چین وارد شده و شناخته شود.

3-2.سخت افزار

برای اطمینان از کارکرد مناسب DACHS در سطح محلی باید محیط سخت افزاری مناسبی را اندازی شود بعد از تدارک سخت افزاری برای سرور مناسب و ایستگاه های کاری اختصاص داده شده به امور روزمره مانند بارگذاری و مقاصد ،مدیریتی لازم است به امکانات پشتیبانی و امنیت و محافظت از این سرور در مقابل ویروس های کامپیوتری توجه شود.

مرکز کامپیوتر دانشگاه هایدلبرگ با استفاده از روش ذخیره سازی توزیع شده IBM ADSTAR، سیستم پشتیبان گیری مهمی را فراهم کرده است. با استفاده از این سیستم هر شب یک نسخه پشتیبان از کل بایگانی بر روی نوار های مغناطیسی در کامپیوتر مرکزی ذخیره می شود کپی های پشتیبانی از این نوارها نیز به طور مرتب در دانشگاه کالسروهه ذخیره می.شود بنابراین داده های بایگانی در سه مکان مختلف که امنیت مناسبی را ارائه می دهند نگهداری می شوند.

دستگاه منبع تغذیه اضطراری (UPS) نیز به عنوان یک سیستم RAID (آرایه افزونه دیسک های مستقل ارزان) (سطح (اول) جهت تأمین امنیت اولیه برای دسترسی بی وقفه نصب شد. برای محافظت در مقابل ویروس از نرم افزار ویروس یاب McAfee استفاده شد. از طریق تعاریفی که در این ویروس یاب وجود دارد، به صورت ساعتی از سرور McAfee استفاده می شود تمام داده های ورودی به طور مداوم چک شده و به طور خودکار فرآیندهای اسکن منظم از کل بایگانی انجام می پذیرد.

بدینوسیله می توانیم زیر ساخت های فناوری اطلاعات موجود در مؤسسه را که حمایت بخش ICT دانشگاه را نیز دارد در این بخش مورد استفاده قرار دهیم در تمامی موارد این تجهیزات باید از نظر دانشگاه ها استاندارد ،باشد به همین دلیل این تجهیزات باید در حوزه آی.سی.تی. مرکزی بوده و در آن جا نگهداری شود. این موضوع می تواند از لحاظ حرفه ای نوعی مزیت به حساب آید. از طرفی هم می تواند یک نقص باشد زیرا به نوعی می تواند محدودیت هایی بر روی سخت افزار و نرم افزار اعمال کند.

2-4. نرم افزار

برای شروع کار بایگانی دیجیتال برای پروژه مطالعات چینی به نرم افزار مناسب نیاز داریم که باید برخی شرایط این کار را داشته باشد.

اینترنت مجموعه عظیمی از داده های به هم پیوسته در فرمت ها و سیستم های کدگذاری مختلف است و بایگانی چنین داده هایی برای محافظت از آن ها به نحوی که محتوا و عملکرد آن به صورت اصلی باشد، قدری مشکل به نظر می رسد. از آن جا که کارکرد محتوای وب وابسته به نرم افزار مرورگر است، هیچ داده قابل اعتمادی برای محافظت از آن وجود ندارد با این حال چنان چه اساس آن محافظت شود، بسیاری از موارد فوق برآورده می شود به این معنی که اگر کل وبگاه بارگذاری شد، باید ساختار فایل اصلی دست نخورده نگه داشته شود البته برای نگهداری اسناد بارگذاری شده و ارتباط آن با باید ارتباطات مناسبی در نظر گرفته شود برای این منظور به نرم افزاری نیاز است که مناسب این کار بوده و قابلیت تنظیم را داشته باشد این نرم افزار باید توانایی اداره طیف گسترده ای از فرمت های مختلف را داشته و مقرون به صرفه نیز باشد بعد از آزمایش هایی که بر روی نرم افزارها صورت گرفت به این نتیجه رسیدیم که نرم افزار آفلاین اکسپلورر، تمام قابلیت های مورد نیاز را برای شروع کار دارد.

2-5. فراداده

مسئله ای که موجب دغدغه فکری شده ایجاد فراداده است از یک طرف فراداده اطلاعات لازم را برای مشاهده و توصیف محتوای سند ایجاد می کند و از طرفی هم موجب دسترسی کاربران به اطلاعات می شود. فراداده هایی مانند نویسنده عنوان و موضوع تا زمانی که جست وجوی متنی وجو نداشته باشد این نوع فراداده به کاربر کمک می کند از طرفی هم فراداده مسائل فنی را برای کمک به حفاظت درازمدت فراهم می کند؛ زیرا اطلاعات لازم را برای مدیریت مناسب را فراهم و بررسی را آن در آینده میسر می کند.

ایجاد فراداده پر هزینه است؛ با این حال با وجود فراداده نیمه خودکار، جمع آوری برخی داده ها باید به صورت دستی انجام شود .بنابراین بدیهی است که ایجاد فراداده برای صدها هزار سند با همان سرعتی که از اینترنت بارگذاری شده غیر ممکن است (شکل 1 را ببینید). حتی اگر این کار ممکن بود یافتن آن چه که دقیقاً این فراداده باید حاوی آن باشد، باز هم یافتن جزئیات و فرمت های آن، بسیار سخت است.

یکی از پرسش های مهمی که در این مورد وجود دارد این است که آیا وجود فراداده ضروری است یا بهتر است بگوییم: آیا صرفاً تکیه بر الگوریتم های جست وجو برای بازیابی تمام داده های مورد نیاز

برای دسترسی و نیز برای اهداف بلند مدت حافظت از داده ها امکان پذیر است؟ از دیدگاه کاربر می توان استدلال کرد که جست وجوی متن ،کامل در مقایسه با سر فصل های موضوعی خام یا عناوین حاوی فراداده های ناقص جهت پیدا کردن اسناد ابزار قابل اعتماد تری است از نظر فنی بسیاری از اطلاعات ضروری از جمله فرمت فایل سرعت بارگذاری رمز گذاری و غیره برای راحتی در بازیابی داده هاست یا حتی می تواند برای ساختار نام گذاری فایل باشد

عکس

برای دسترسی و نیز برای اهداف بلند مدت حافظت از داده‌ها امکان‌پذیر است؟ از دیدگاه کاربر می‌توان استدلال کرد که جست‌وجوی متن کامل، در مقایسه با سرفصل‌های موضوعی خام یا عناوین حاوی فراداده‌های ناقص جهت پیدا کردن اسناد ابزار قابل اعتمادتری است. از نظر فنی، بسیاری از اطلاعات ضروری از جمله فرمت فایل، سرعت بارگذاری، رمزگذاری، و غیره برای راحتی در بازیابی داده‌هاست یا حتی می‌تواند برای ساختار نام‌گذاری فایل باشد.

شکل ۱-۱۰. رابط کاربری جست‌وجوی فراداده DACHS که امکان جست‌وجوی پیشرفته در بایگانی را دارد.

در پایان این ایده رد شد. منابع دیجیتال نوعی اطلاعات هستند که زود از بین می‌روند. بنابراین امکان دارد تا در اسناد جداگانه‌ای نیز ذخیره شوند تا این منابع را با اطمینان و به‌صورت یکپارچه در آینده و به‌صورت طولانی مدت قابل دسترسی کرد. با توجه به نوع داده‌ها و اطلاعات، امکان نابودی آنها وجود دارد.

برای مدیریت و ادغام بهتر مواد بایگانی شده در موجودی کتابخانه‌های بزرگ، تصمیم گرفته شد تا فراداده به‌عنوان بخشی از کاتالوگ به‌صورت منظم ایجاد و تهیه شود. بنابراین، کاتالوگ با فراداده تطبیق داده شد تا بخش‌هایی برای مدیریت درست، تاریخچه مبدأ، سابقه مدیریت، نوع فایل، شناسه، و سایر موارد تعیین شود. بسته به پیچیدگی منابع، در حال حاضر سرگرم تهیه رکوردهای فراداده‌ای هستیم که یا بتواند تک‌فایل را توصیف کند که فقط متن را شامل می‌شود یا شامل مجموعه‌ای کامل از فایل‌ها مانند وبگاه، انجمن و یا روزنامه باشد.

۲-۶. سیاست‌گذاری و خط‌مشی مجموعه

کاملاً بدیهی است که هدف از بایگانی دیجیتال مطالعات چینی نمی‌تواند شامل تمامی موارد موجود در اینترنت باشد و از آن محافظت نماید. نه از لحاظ فنی چنین امکانی وجود دارد و نه مفید به‌نظر می‌رسد. به‌عنوان مؤسسه‌ای تحقیقاتی، علاقه‌مند به بخشی از اینترنت چین هستیم که منعکس کننده جنبه‌های

شکل 101 رابط کاربری جست و جوی فراداده DACHS که امکان جست وجوی پیشرفته در بایگانی را دارد.

در پایان این ایده رد شد منابع دیجیتال نوعی اطلاعات هستند که زود از بین می روند. بنابراین امکان دارد تا در اسناد جداگانه ای نیز ذخیره شوند تا این منابع را با اطمینان و به صورت یکپارچه در آینده و به صورت طولانی مدت قابل دسترسی کرد. با توجه به نوع داده ها و اطلاعات امکان نابودی آن ها وجود دارد.

برای مدیریت و ادغام بهتر مواد بایگانی شده در موجودی کتابخانه های بزرگ، تصمیم گرفته شد تا فراداده به عنوان بخشی از کاتالوگ به صورت منظم ایجاد و تهیه شود. بنابراین کاتالوگ با فراداده تطبیق داده شد تا بخش هایی برای مدیریت درست، تاریخچه مبدأ، سابقه مدیریت، نوع فایل، شناسه، و سایر موارد تعیین شود. بسته به پیچیدگی ،منابع در حال حاضر سرگرم تهیه رکوردهای فراداده ای هستیم که با بتواند تک فایل را توصیف کند که فقط متن را شامل می شود یا شامل مجموعه ای کامل از فایل ها مانند ،وبگاه انجمن و یا روزنامه باشد.

2-6.سیاست گذاری و خط مشی مجموعه

کاملاً بدیهی است که هدف از بایگانی دیجیتال مطالعات چینی نمی تواند شامل تمامی موارد موجود در اینترنت باشد و از آن محافظت نماید نه از لحاظ فنی چنین امکانی وجود دارد و نه مفید به نظر می رسد. به عنوان مؤسسه ای ،تحقیقاتی علاقه مند به بخشی از اینترنت چین هستیم که منعکس کننده جنبه های

خاصی از جامعه چینی است که موارد زودگذری هم هستند بنابراین فراهم کردن انتخاب آگاهانه منابع یک سرمایه است که در حال حاضر به ما کمک می کند و نیز به کاربران کمک می کند در آینده تا جهت شناسایی موادی که ما نیاز به آن داریم و مربوط به هدف ما می شود یاری می رساند. البته ارزش ها در گذر زمان تغییر خواهد کرد و نسل بعد ممکن است انتخاب های مختلفی داشته باشد. در این مورد، بایگانی اینترنت هنوز جایگزینی بسیار غنی را فراهم میکند که در آن گزینه های دیگری در دسترس خواهد بود و ترکیبی از دو رویکرد گزینشی و جامع است که کاربر در آینده با گسترده ترین آرایه از امکانات برای تحقیقات خود قادر به تأمین آن است.

با این حال همزمان میخواهیم خطر محدودیت بیش از حد را در انتخابمان کاهش دهیم. همانطور که در قبلا نیز گفته شد و در ادامه بحث خواهد شد DACHS در حال تبدیل شدن به پروژه تعاونی بزرگ تر است شرکای ،مختلف معیارهای مختلفی دارند و در نتیجه مقادیر مختلفی را در روند انتخاب خود اعمال می کنند این امر تنها می تواند محتوای بایگانی را غنی ،کند در حالی که در سیاست انتخاب آگاهانه منابع تغییری رخ نمی دهد

با این حال به منظور استفاده بهتر از منابع محدود و در دسترس باید راهبردهای هوشمندانه ای را برای توسعه مجموعه به کار گرفت. با توجه به تعداد زیاد سیاست های اینترنتی در مورد چین و سرعت زیاد آن و توسعه آن و همچنین ناپدید شدن این مطالب در شبکه چین وظیفه ساختن بایگانی منابع برخط که بتواند مهم و ارزشمند باشد کاری سخت و دلهره آور است.

برای رفع این مشکل شروع به ساخت یک شبکه اطلاعاتی از افراد (محققان بومی و شهروندان) کردیم که به طور فعال یا غیر فعال بخشی از این کار را بر عهده گیرند. این کار زمانی عملی می شود که با استفاده از دانشی باشد که برای شناسایی سیاست های خاص و متناسب با کار باشد با این حال این شبکه اطلاعاتی بزرگ می تواند خیلی متنوع تر از فرآیند انتخاب باشد.

معمولاً شبکه اطلاع رسانی نیز هدف دیگری را دنبال می کند. از آن جا که اعضای آن بخشی از فرهنگ اینترنت معاصر است آن ها می توانند اطلاعات زمینه ارزشمندی در مورد منابع بایگانی شده ارائه دهند تا آن جا که ممکن است این اطلاعات باید در فراداده و یا صفحات وب تخصیص داده شده به عنوان بخشی از مجموعه ایجاد و حفظ .گردد

از آن جا که چند ماه نیز بر روی تعدادی پروژه خاص مانند شعر معاصر، سارس (1) یا موضوع همجنس بازی در چین کار کرده ایم، محققان و دانشجویان کارشناس ارشد برای ایجاد آرشیوهای جامع بر روی این موضوع ها کار می کنند، از جمله این وب گاه ها، آن هایی هستند که در معرض خطر بوده و یا شامل موارد دیگر مانند عکس و پوستر می شود برای بایگانی این منابع می توان متن ساده ای را در زمان بایگانی به آن اضافه کرد این نوع عملکرد را برای درک بهتر این منابع که زود از بین می روند، به ویژه در آینده دور، ضروری می دانیم.

اما این کار صرفاً با تکیه بر آگاهان و دانشمندان انجام نمی شود. تأثیر رویدادهای خاص بین المللی مانند حمله تروریستی 11 سپتامبر و یا بازی های المپیک چین در سال 2008، غالباً باعث بحث های داغ بر

1- SARS

روی اینترنت می شود برای جلوگیری از شیوع افکار عمومی ما در حال کار بر روی سیاهه وارسی تالار گفتگو و روزنامه ها هستیم که نوع گزارش ها به آن مربوط می شود تا بتوان فاصه زمانی چند هفته قبل و بعد از چنین رویدادهایی را پوشش دهد.

مجموعه های مشابه دیگری هم توسط اشخاص ،خصوصی ،پژوهشگران گروه های پژوهشی و یا سایر مؤسسه های به DACHS اهدا و یا فروخته شده اند. گاهی اوقات برای کمک به ناشران وب گاه ها در معرض خطر برای حفظ محتوای وبگاه شان نزدیک شدیم (و یا با آن ها تماس گرفتیم). به هر حال، این مجموعه ها برای DACHS طراحی و ایجاد نشده اند (یا حتی ممکن است همه استانداردهای کیفیت مد نظر ما را دنبال نکنند) اما DACHS به آن ها کمک می کند تا این منابع در درازمدت در دسترس باشند.

2-7.مشارکت

ملاحظه کردیم که اجرای بایگانی تلاشی مشارکتی است و مشارکت راهبردی مهمی برای بقای آرشیوهای وب در مقیاس کوچک است مشارکت می تواند به نگهداری طولانی مدت از طریق عرضه خدمات به اعضا کمک کند. بنابراین، شرکا نیز اجازه توزیع کار را می دهند که این کار به معنای کاهش هزینه و امکان انتخاب گسترده تر از منابع بایگانی است استانداردهای سخت افزار تجربه و کیفیت نیز می تواند به طور وسیعی به اشتراک گذاشته شود و عملکرد کلی بایگانی را بهبود بخشد. مشارکت به عنوان بخشی از راهبرد سیاسی مهم .است پروژه بین الملل مشارکتی قطعاً ساده تر از شناخت جامعه علمی و کمپین موفق برای تأمین بودجه است.

شناخت مسائل DACHS چیزی است که توسعه دستور عمل ها برای مشارکت ها را ممکن می نماید این دستور عمل ها توسط دانشگاه لیدن ،هلند مورد بررسی قرار گرفت که در انتهای سال 2003 مشارکت آن عملی شد. هدف اصلی این بود که شرکا مستقل مانده و تا حد امکان هویت خود را حفظ کنند ولی برخی استانداردها و خدمات وجود دارد که باید به اشتراک گذاشته شود.

یکی از مسائل اصلی در همکاری ایجاد یافته های ایدز ،است از جمله راهنمای موضوعی پیونددار (یا جدول محتوا) و همچنین متن کامل و یکپارچه سازی امکانات جست وجو در فراداده و بایگانی آن مسائلی است که همکاری شرکای فعلی و آینده پروژه را می طلبد این امر نه تنها به دسترسی متمرکز به فرم صفحه اصلی به صورت اشتراکی باید دارد بلکه باید در مورد چگونگی گزینه های جست و جو برای رسیدن به هدف مطلوب نیز مذاکره شود جست و جوی متن کامل باید در تمام حوزه ها امکان پذیر باشد و برای فراداده یا یک فهرستگان لازم است که همه داده های فیزیکی با هم ذخیره شود که بتواند از طریق فراداده ها منابع مختلف حتی منابع محلی را جست وجو کند و آن ها را به شیوه ای منسجم ارائه دهد. در هر مورد به ایجاد استانداردهای مختلف جهت ایجاد فراداده نیاز داریم.

مسائل دیگری که باید مورد بحث قرار گیرند عبارت اند از سیاست محدود کردن دسترسی مشترک، مقررات تقسیم کار و امور روزمره ای که از دوباره کاری منابع بایگانی جلوگیری می کند.

نکته مهم دیگر برای همکاری این است که کدام یک از شرکا در آینده قادر هستند زیر ساخت بایگانی به صورت محلی ایجاد نمایند و یا اینکه از امکانات خود برای این پروژه استفاده کنند. از آن جا که

تمامی کارهای فوق هنوز در شرف انجام می باشد به اشتراک گذاشتن تجربه در این زمان غیر ممکن است.

بیش تر مسائل در فرآینـده مـذاکره و پیاده سازی واقعی برای امور مشارکتی پدیـدار خواهـد شد. در هر صورت، مشارکت برای پروژه های بایگانی وب در مقیاس کوچک در بسیاری از سطوح ضروری است و از امور مهم و بدیهی است.

3. درس های آموخته شده: جمع بندی

در حدود 4 سـال پس از اجرای DACHS، در حال حاضـر بسـیاری از پرسـش ها و مشـکلات هنوز حل نشـده باقی مانده و یا در حال برطرف شدن و در حال توسـعه است. با نگاهی به گذشـته، می توان گفت که مسائل کمی وجود دارد که لاینحل باقی مانده است که سعی خواهیم کرد آن ها به طور متفاوتی نگاه کنیم.

ضرورت امر در این ،زمینه تخصـیص سـمت ها برایپروژه است. علاوه بر روال کار روزانه باید به کار توسـعه و مدیریت پروژه های بایگانی وب نیز توجه کرد و آن را دست کم نگرفت. تصمیم به انحصاری کردن آن و سپردن همه چیز به یک کتابدار که در حال حاضر مجموعه ای از وظایف دیگر را انجام می دهد، قدری جای تأمل دارد.

در هر صورت نقش او در توسعه پروژه مزایای بسیاری دارد و مهم است سمت مدیریتی تخصیص داده شده برای DACHS بسیار مناسب بوده و منجر به اجرای موفق پروژه می.شود جا داشت تا بسیاری از مسائل خیلی زود تر و مؤثر تر به کار گرفته میشد تا تأثیر مثبتی در توسعه پروژه داشته باشد.

مسـئله دوم که – حـداقل از امروز – بایـد به طور ویژه ای به پرداخته شود و مسائل آن حل شود انتخاب نرم افزار جمع آوری داده و بایگانی می باشـد. در سال 2001 که DACHS شـروع به بایگانی وب کرد هنوز در مرحله ابتـدایی بود و بسـیاری از دست اندکاران بزرگ امروز تنها فقط شـروع به توسـعه و انتشار تلاش های خود کردنـد. اگر چه نرم افزار امروز ما که برای اهداف مورد نظر در آن زمان انتخاب کردیم بیش تر نیازهای ما را برآورده می کند ابزار های امروزی نیز برای این کار بسـیار مناسب هستند. لازم است ذکر گردد که یک انتخاب جدید باید روند استفاده و ایجاد فراداده را ساده کند.

مطالبی که عنوان شد نقطه نهایی بوده است که من می خواسـتم آن را ایجاد کنم. مسائل بسـیاری هم وجود دارد که بایگانی وب را مشـکل می کند و در برخی موارد به مسائلی بر می خورید که به دفعات سـراغ آن رفتید ولی خودتان نتوانستید آن را حل کنید بدون اینکه نگران نوع فرمت فایـل یـا توسـعه نرم افزار جمع آوری داده باشـیم این امر امکان پـذیر است – حتی ضـروری است– که بر تلاش دیگران تکیه کنیم. همکاری نه تنها در میان شرکای پروژه بایگانی است بلکه همکاری در سایر مؤسسه های در زمینه بایگانی وب، می تواند راه حل های مهم و ابزارهایی را فراهم کنند که شما به تنهایی قادر به ایجاد آن نخواهید بود.

4. منابع مفید

فهرست زیر تنها انتخاب بسـیار کوچکی از منابع است که برای کار مفیـد است اگر شـما وب سایت های زیر را مشاهده کنید منابع بسـیار بیش تری را خواهید یافت به خصوص این که این موضوع بسیار عالی با

عنوان دروازه دیجیتال PADI. نام دارد که همه منابع در ماه می 2006 قابل دسترس بود.

Websites 4.1

Council on Library and Information Resources (CLIR)

http://www.clir.org/

Electronic Resource Preservation and Access Network (erpaNet)

http://www.erpanet.org/

International Internet Preservation Consortium

http://netpreserve.org/

Internet Archive

http://www.archive.org/

Networked European Deposit Library (NedLib)

http://www.kb.nl/coop/nedlib/

PADI Preserving Access to Digital Information (National Library of Australia)

http://www.nla.gov.au/padi)

PADI: Web archiving

http://www.nla.gov.au/padi/topics/92.html

Mailing Lists 4.2

Archivists

http://groups.yahoo.com/group/archivists/

DigiCULT

http://www.digicult.info/pages/subscribe.php

DIGLIB - Digital Libraries Research mailing list (IFLA)

http://infoserv.inist.fr/wwsympa.fcgi/info/diglib/

OAIS Implementers (RLG)

http://lists2.rlg.org/cgi-bin/lyris.pl?enter=oais-implementers

PadiForum (National Library of Australia) http://www.nla.gov.au/padi/forum/ Web-Archive/ http://listes.cru.fr/wws/info/web-archive

Newsletters and Magazines 4.3

CLIR Issues

http://www.clir.org/pubs/issues/ DigiCULT Newsletter

http://www.digicult.info/pages/newsletter.php

D-Lib Magazine

http://www.dlib.org/DPC/PADI What is new in digital preservation

http://www.nla.gov.au/padi/qdiges

RLG DigiNews http://www.rlg.org/en/page.php?Page-ID=12081

چکیده

هدف این پژوهش بررسی قابلیت های قالب های یونی مارک و مارک 21 برای سازماندهی منابع اطلاعاتی وب و مقایسه آن ها با یکدیگر است تـا از این طریق بتوان به ارزیـابی این دو قـالب در تهیه پیشـینه هـای اطلاعـاتی برای منـابع اطلاعاتی وب دست یافت روش پژوهش تحلیل محتوا ،است به صورتی که ابتـدا به شـناسایی عناصـر توصـیف مرتبط با منابع اطلاعاتی وب در قالب کتاب شـناختی یونی مارک و مارک 21، و سـپس به تحلیل و مقایسه آن ها با یکـدیگر پرداخته شد. از یافته های این پژوهش می توان چنین دریافت که قالب مارک می توانـد به عنوان اسـتانداردی برای ذخیره و بازیابی منابع اطلاعاتی وب به کار رود با این حال مارک راه حلی برای چگونگی روزآمد نمودن و تعیین سـطح توصـیف منـابع اطلاعـاتی وب تهیه نکرده است. کلیـد واژه هـا منـابع اطلاعـاتی ،وب ،یونی مارک مارک ،21 قالب کتاب شناختی، سازماندهی

بررسی و مقایسهٔ قابلیت های قالبهای یونی مارک و مارک 21 برای سازماندهی منابع اطلاعاتی وب

اشاره

رقیه حجازی، (1) دکتر مرتضی کوکبی (2)

1-مقدمه و بیان مسئله

منـابع اطلاعاتی وب بر اساس ویراست 2005 قواعـد فهرست نویسـی انگلوامریکن، شامل منابع (داده ها و یا برنامه) کـد گذاری شده به منظور کـاربرد با دسـتگاه های رایانه ای است که نیاز به برقراری ارتباط با یک شـبکه رایانه ای دارد (Weitz 2006) این منابع محسوس و شامل محمل فیزیکی نیستند، و همچنین برای استفاده از آن ها نیاز به برقراری ارتباط با یک دستگاه رایانه (مانند شبکه) یا منابع ذخیره شده بر روی دیسک سخت یا دیگر ابزارهای ذخیره سازی است (Miller 2008).

با ظهور محمل های جدیـد ،اطلاعات از جمله منابع اطلاعات موجود در محیط وب نظام های سازمانـدهی اطلاعات نیازمند روش های نوینی در ذخیره و بازیابی اطلاعات شده اند. نمایه سازی موتورهای کاوش را نمی توان روش مناسبی در ذخیره و بازیابی اطلاعات دانست زیرا بـا تـوجه به گفته فتـاحی (1380) آشـفتگی اینترنت در حال حاضـر بیشتر ناشـی از پراکنـدگی و ناکارآمـدی روش های سازمانـدهی اطلاعـات در این محیـط .است بنـابراین برای سازمانـدهی اطلاعـات نیـاز به رویکردهـایی اسـتاندارد با توجه به ویژگی های محمل های اطلاعاتی است

1- کارشناس ارشد علوم کتابداری و اطلاع رسانی r.hejazi86@yahoo.com 2.

2- استاد کتابداری و اطلاع رسانی دانشگاه شهید چمران اهواز kokabi80@yahoo.com

نظام سازمانـدهی دانش برای سازگاری با تغییر ساختار منابع اطلاعاتی همواره دو رویکرد را برگزیـده است. نخست هماهنگی سازگاری و تطـابق نظـام هـای سـنتی با محیط و رسانه های جدیـد و دیگر طراحی و ایجاد نظام های جدیـد به منظور حـداکثر بهره وری از امکانات و قابلیت های محیط جدیـد (طاهری 1387). بـا توجه به اینکه در محیـط هـای کتابخانه ای حجم عظیمی از منابع اطلاعاتی با اسـتفاده از اسـتانداردهای موجود سازماندهی شده اند و تبدیل و تغییر در آن ها نیازمند صـرف وقت و هزینه است، به نظر می رسد استفاده از رویکرد نخست حداقل در محیط های کتابخانه ای مناسب تر باشد یکی از استانداردهای مورد استفاده در نظام های سنتی قالب مارک است قالب مارک از اسـتانداردهای مورد اسـتفاده برای ماشـین خوان کردن داده های کتابشـناختی است. قالب های مختلفی از مارک همچون مارک 21 و یونی مـارک وجود دارنـد. از سوی دیگر، بیـانیه کتابخانه کنگره برای چارچوب کتابشـناختی برای دوران دیجیتال (2011 در عمرانی 1390) به این نکته اشاره می کند که «استاندارد مارک مسئولیت تولید میلیون ها پیشینه کتاب شناختی در گوشه و کنار دنیا را بر عهده دارد و نیاز است که در تمام دورهٔ نقل و انتقال به پشتیبانی از مارک ادامه دهیم. علاوه بر این بیانیه عنوان می کنـد که سامانه ها و خـدمات مبتنی بر پیشـینه هـای مرتبـط بـا قـالب مـارک تـا سـال هـای زیادی بخش مهمی از زیر ساخت ها خواهنـد بود». بنابراین توجه به این قالب در سازماندهی اطلاعات حائز اهمیت است زیرا علاوه بر استفاده کنونی از این ،قالب الگوها و استانداردهای جایگزین آینده نیز سازگار با آن طراحی خواهند شد.

همـان طور که بیان ،شـد قالب های مختلفی از مارک موجود است که می توان قالب مارک 21 (به دلیل کاربرد آن در پیشـینه های کتاب شنـاختی کتابخانه کنگره آمریکا، کتابخانه بریتانیا، کتابخانه ملی کانادا و برخی کتابخانه های دیگر) و قالب یونی مارک (به دلیل بین المللی بودن و مبنای مارک ایران) را از مهم ترین آن ها دانست. هر دو قالب بر اساس قواعـد فهرست نویسـی انگلوامریکن و تأکید بر منابع چاپی تهیه شـده انـد. با ظهور محمل های اطلاعاتی ،جدیـد تغییراتی در قالب ها ایجاد شـد تا بتواننـد با محمل های جدید سازگار و در جهت سازمانـدهی آن ها کارآمد باشـند که از آن جمله می توان به افزودن فیلدها و فیلدهای فرعی ،نشان گرها و نویسه های موجود در برچسـب پیشینه اشاره کرد.

بـا توجه به ویژگی منـابع اطلاعاتی وب از یک سو و تفاوت های موجود میان قالب مارک 21 و یونی مارک این پرسـش پیش می آید که آیا قالب هایی همچون یونیمارک و مارک 21 توانایی سازماندهی منابع اطلاعاتی وب را دارند؟ علاوه بر این تفاوت های موجود در میان این دو قالب تأثیری بر کیفیت سازماندهی منابع اطلاعاتی وب دارد؟ پژوهش حاضـر سعی دارد با بررسی دو قالب یونی مارک و مارک 21، به بررسـی عناصـر تعریف شـده بر اساس خصوصـیات منابع اطلاعاتی ،وب و همچنین تبیین کارایی و قابلیت دو قالب در سازماندهی بهینه منابع اطلاعاتی وب بپردازد. بنابراین به ارزیابی توان و کارایی هر دو قالب در سازمانـدهی اطلاعات وب و مقایسه آن ها با یکدیگر پرداخته شد.

2-پرسش های پژوهش

این پژوهش در پی پاسخ به پرسش های زیر انجام شد:

1-1 هر یک از قالب های یونیمارک و مارک 21 شامل چه عناصری (اعم از فیلد فیلد، فرعی نشان گر و غیره) برای سازماندهی منابع اطلاعاتی وب هستند؟

2-1 آیا عناصر تعریف شده در دو قالب یونی مارک و مارک 21 نشان دهندهٔ خصوصیات منابع اطلاعاتی وب هستند؟

3-1 آیا تفاوتی میان عناصر تعریف شده در قالب یونی مارک و مارک 21 برای سازماندهی منابع اطلاعاتی وب وجود دارد؟

4-1 اگر پاسخ پرسش سوم مثبت است تفاوت های موجود چه تأثیری بر سازماندهی منابع اطلاعاتی وب می گذارد؟

3-پیشینه پژوهش

بررسی پیشینه های مرتبط با پژوهش نشان داد که پژوهش های صورت گرفته بر اساس دو موضوع کلی قابل تقسیم بندی هستند. موضوع نخست در رابطه با ماهیت منابع اطاعاتی وب و شناسایی تفاوت های آن ها با دیگر قالب های اطلاعات بود که از آن جمله می توان به پژوهش وارد (2001) (1) اشاره نمود. وی در این پژوهش به اهمیت سازماندهی منابع اینترنتی اشاره کرده و معتقد است که مهارت های فهرست نویسی برای سازماندهی منابع موجود بر روی وب مناسب است در این مقاله به برخی از فیلدهای مارک 21 که در سازماندهی منابع اینترنتی کارآمد هستند نیز اشاره شده است. همچنین حاجی زین العابدینی (1381) به بررسی و تحلیل آخرین فعالیت ها و تحقیقات در زمینه سازماندهی اطلاعات در اینترنت و ارائه یک الگوی مناسب برای فهرست نویسی منابع فارسی پرداخت. نتایج به دست آمده نشان داد که دو روش مهم برای سازماندهی اطلاعات در اینترنت وجود دارد. روش نخست ایجاد پیشینه های کتاب شناختی با استفاده از قواعد عام فهرست نویسی و قواعد خاص منابع اینترنتی و روش دوم ایجاد پیشینه های کتاب شناختی با استفاده از روش های ابر داده عنوان شده است. او به دلیل ویژگی های منابع اینترنتی فارسی روش دوم را برای منابع اینترنتی فارسی مناسب ندانسته است.

موضوع دیگر در رابطه با بررسی انواع فراداده ها در سازماندهی منابع اطلاعاتی وب بود که از آن جمله می توان به پژوهش فارد و ریگیو (2004) (2) اشاره نمود که به بررسی چند استاندارد فراداده از جمله قالب مارک برای مدیریت منابع الکترونیکی (که منابع اطلاعاتی وب زیر مجموعه ای از آن است) پرداختند تحلیل های این پژوهش گران مشخص می کند که در زمان انجام پژوهش هیچ استاندارد و الگوی فراداده ای توانایی نمایش پیچیدگی های منابع الکترونیکی را نداشته است. طاهری (1387) نیز به مقایسه کارایی طرح فراداده ای هسته دوبلین و قالب فراداده ای مارک 21 در سازماندهی منابع اطلاعاتی وب پرداخت نتایج پژوهش وی نشان داد که قالب مارک 21 برای ذخیره پردازش و مبادله اطلاعات محیط وب مناسب تر است. علاوه بر این پژوهش های دیگری همچون مرور قالب های ابر داده (Heer،

1- Ward

2- Fard and Riggio

1996)، کنترل مستند در زمینه کنترل کتاب شناختی در محیط الکترونیک (Gorman 2004)، بررسی و مقایسه قواعد فهرست نویسی انگلوامریکن و عناصر هسته دوبلین برای سازماندهی منابع اینترنتی (کوکبی و آخشیک 1385 و سازماندهی صفحات وب با استفاده از نظام های رده بندی کتابخانه (ملای مقدم و نعیم آبادی 1385) نیز با هدف بررسی و مقایسه دو رویکرد سنتی و نوین فهرست نویسی منابع اینترنتی صورت گرفته است

مطالعه پژوهش های مرتبط نشان داد که سازماندهی منابع وب با استفاده از استانداردها و قواعد فهرست نویسی مورد توجه بوده است. علاوه بر این پژوهش گران همواره در پی یافتن استانداردهای مناسب با ویژگی های منابع و بی در جهت سازماندهی هر چه بهتر آن ها بوده اند به همین جهت به ارزیابی و بررسی آن ها پرداخته اند.

4-روش شناسی پژوهش

روش این پژوهش تحلیل محتوا بود به صورتی که در ابتدا به شناسایی عناصر توصیف مرتبط با منابع اطلاعاتی وب در قالب کتاب شناختی یونی مارک و مارک 21 و سپس به تحلیل و مقایسه آن ها با یکدیگر پرداخته شد. برای انجام این پژوهش از قالب کتاب شناختی یونی مارک International Federation of 2008 (Library Associations and Institutions (IFLA و قالب کتاب شناختی مارک 21 (Library 2012 of Congress) استفاده شد. قابل ذکر است که شماری از عناصر دادهای موجود در قالب های کتاب شناختی یونی مارک و مارک 21 می توانند در پیشینه های منابع اطلاعاتی وب کاربرد داشته باشند، همان طور که در پیشینه های قالب های دیگر هم کاربرد دارند اما به دلیل تمرکز این پژوهش بر روی منابع اطلاعاتی ،وب تنها عناصری که به طور خاص به این نوع از منابع اختصاص داشتند، مورد بررسی قرار گرفتند.

5-تجزیه و تحلیل یافته ها

برای پاسخ به پرسش نخست پژوهش به شناسایی فیلدها فیلدهای فرعی نشان گرها و نویسه های موجود در برچسب پیشینه مرتبط با منابع اطلاعاتی وب پرداخته شد برای پاسخ به این پرسش، علاوه دست نامه کامل هر دو قالب دست نامه آموزشی میلر برای فهرست نویسی منابع اینترنتی (Miller 2008)، راهنمای یونی مارک برای منابع الکترونیکی (International Federation of Library Associations and 2000) و جدول های تبدیل یونی مارک به مارک 21 (International Federation Institutions (IFLA). 2001 of Library Associations and Institutions (IFLA مرجع در نظر گرفته شد. در جدول 1 تنها فیلدها فیلدهای فرعی و نشان گرهایی که با محتوای منابع اطلاعاتی وب (به طور خاص) سازگار بودند، نشان داده شده است علاوه بر این معادل هر یک از عناصر در دو قالب در یک سطر قرار گرفت.

عکس

جدول ۱: جدول تطبیقی عناصر مربوط به منابع اطلاعاتی وب در قالب‌های یونی‌مارک و مارک ۲۱

مارک ۲۱						یونی‌مارک					
برچسب پیشنه	نویسه 06: نوع پیشنه/ کد m: فایل رایانه‌ای			کد l شامل سیستم‌ها و خدمات پیوسته هم می‌شود		برچسب پیشنه	نویسه ششم: نوع رکورد/ کد l: فایل رایانه‌ای			کد l شامل سیستم‌ها و خدمات پیوسته هم می‌شود	
شماره فیلد	نام فیلد	نشانگرها		فیلدهای فرعی	توضیحات	شماره فیلد	نام فیلد	نشانگرها		فیلدهای فرعی	توضیحات
		اول	دوم					اول	دوم		
						4—	بلوک شناسه رابط	—	—	$b نام عام مواد	عبارت "پیوسته" در فیلد فرعی $b نشان می‌دهد که مدرک پیوند شده به شکل پیوسته می‌باشد.
245	وضعیت عنوان	—	—	$h: رسانه	عبارت "پیوسته" در فیلد فرعی $h نشان دهندهٔ منابع وبی است.	200	عنوان و نام پدیدآور	—	—	$b نام عام مواد	عبارت "پیوسته" در فیلد فرعی $b نشان دهندهٔ منابع وبی است
240	عنوان قراردادی	—	—	$h: رسانه		500	عنوان قراردادی	—	—	$b نام عام مواد	
243	عنوان مشترک قراردادی	—	—	$h: رسانه		501	عنوان مشترک قراردادی	—	—	$b نام عام مواد	
630	شناسه افزودهٔ موضوع- عنوان قراردادی	—	—	$g: اطلاعات متفرقه	عبارت "پیوسته" در فیلد فرعی $g نشان دهندهٔ منابع اطلاعاتی وب است	605	عنوان به منزله موضوع	—	—	$n اطلاعات متفرقه	عبارت "پیوسته" در فیلد فرعی $n نشان دهندهٔ منابع اطلاعاتی وب است
006	عناصر داده‌ای با طول ثابت- ویژگی مواد اضافی	—	—	نویسه 06: شکل مدرک نویسه 09: نوع فایل رایانه‌ای	کد o نشان دهندهٔ شکل پیوسته است کد j نشان دهندهٔ خدمات و سیستم‌های پیوسته است						
007	توصیف فیزیکی فیلدهای ثابت- اطلاعات عمومی	—	—	نویسه 01: تعیین مواد خاص	کد r نشان دهندهٔ دسترسی از راه دور است	135	فیلد داده-های کد شده: منابع الکترونیکی	—	—	$a: داده‌های کد شده برای منابع الکترونیکی/نویسه 1: تعیین مواد خاص/کد r	کد r نشان دهندهٔ سیستم‌های پیوسته است.
008	عناصر داده‌ای با طول ثابت	—	—	نویسه 26: نوع فایل رایانه‌ای نویسه 23: شکل مدرک	کد j نشان دهندهٔ خدمات و سیستم‌های پیوسته است کد o نشان دهندهٔ شکل آنلاین است	135	فیلد داده-های کد شده: منابع الکترونیکی	—	—	$a: داده‌های کد شده برای منابع الکترونیکی/نویسه 0: نوع منبع الکترونیکی/کد j	کد j نشان دهندهٔ خدمات و سرویس‌های پیوسته است

جدول 1 جدول تطبیقی عناصر مربوط به منابع اطلاعاتی وب در قالب های یونی مارک و مارک 21

ص: 235

عکس

256	مشخصات فایل رایانه‌ای	—	—	$a: مشخصات فایل رایانه‌ای	اطلاعاتی همچون نوع فایل و اندازه آن در این فیلد فرعی وارد می‌شود	230	منطقه خاص مواد: مشخصات منابع الکترونیکی	—	—	$a: برچسب و اندازهٔ فایل	اطلاعاتی همچون نوع فایل و اندازه آن در این فیلد فرعی وارد می‌شود
338	نوع حامل	—	—	$a: اصطلاح نوع حامل	بر اساس آر.دی.ای عبارت "منابع پیوسته" یکی از حامل‌های رایانه‌ای است.						
516	نوع فایل رایانه‌ای یا یادداشت داده	—	—	$a: نوع فایل رایانه‌ای یا یادداشت داده	عبارت "سیستم‌های پیوسته" در این فیلد فرعی نشان دهندهٔ نوع منبع الکترونیکی است.	336	یادداشت نوع منابع الکترونیکی	—	—	$a: متن یادداشت	عبارت "خدمات و سیستم‌های پیوسته" در این فیلد فرعی نشان دهندهٔ نوع منبع الکترونیکی است.
538	یادداشت جزئیات سیستم	—	—	$a: یادداشت جزئیات سیستم $i: متن نمایش $u: شناسگر متحدالشکل منبع $3: مواد معین	اطلاعات در مورد شکل دسترسی به منابع وبی در فیلد فرعی $a: آدرس دسترسی خودکار به مدرک در فیلد فرعی $u، و بخشی از مواد شرح داده شده در $3 وارد می‌شود	337	یادداشت سیستم مورد نیاز	—	—	$a: متن یادداشت $u: شناسگر متحدالشکل منبع	اطلاعات در مورد شکل دسترسی به منابع وبی در فیلد فرعی $a و آدرس دسترسی خودکار به مدرک در فیلد فرعی $u وارد می‌شود
753	جزئیات دسترسی سیستم به فایل رایانه‌ای	—	—	$a: ساختار و الگوی دستگاه‌ها $b: زبان‌های برنامه‌نویسی $c: سیستم عامل		626	جزئیات دسترسی فنی (منابع الکترونیکی	—	—	$a: ساختار و الگوی دستگاه‌ها $b: زبان‌های برنامه-نویسی $c: سیستم عامل	
856	دسترسی و تعیین محل الکترونیکی	روش-های دسترسی	روابط	$a نام میزبان؛ $b شماره دسترسی؛ $c فشرده‌سازی اطلاعات؛ $d مسیر؛ $f نام الکترونیک؛ $h پردازشگر درخواست؛ $i دستورالعمل؛ $j بیت در هر ثانیه؛ $k کلمه عبور؛ $l ورود به سیستم؛ $m برقراری تماس برای دریافت کمک؛ $n نام محل میزبان؛ $o سیستم عامل؛ $p درگاه؛ $q نوع قالب الکترونیکی؛ $r تنظیمات؛ $s اندازه فایل؛ $t تکرار پایانه‌ای؛ $u شناسگر متحدالشکل منبع؛ $v ساعاتی که روش دسته‌ای قابل دستیابی است؛ $w شماره کنترل پیشینه؛ $x یادداشت غیر عمومی؛ $y متن پیوند؛ $2 روش‌های دسترسی؛ $3 مواد مشخص شده؛ $6 پیوند؛ $8 پیوند فیلد و شماره توالی	856	دسترسی و تعیین محل الکترونیکی	روش-های دسترسی	تعریف نشده	$a نام میزبان؛ $b شماره دسترسی؛ $c اطلاعات فشرده؛ $d مسیر؛ $e تاریخ و ساعت دسترسی؛ $f نام الکترونیک؛ $h پردازشگر درخواست؛ $i دستورالعمل؛ $j بیت در هر ثانیه؛ $k کلمه عبور؛ $l ورود به سیستم؛ $m برقراری تماس برای دریافت کمک؛ $n نام محل میزبان در فیلد فرعی $a؛ $o سیستم عامل؛ $p درگاه؛ $q نوع قالب الکترونیکی؛ $r تنظیمات؛ $s اندازه فایل؛ $t تکرار پایانه‌ای؛ $u شناسگر متحدالشکل منبع؛ $v ساعاتی که روش دسته‌ای قابل دستیابی است؛ $w شماره کنترل پیشینه؛ $x یادداشت غیر عمومی؛ $y روش‌های دسترسی؛ $z یادداشت عمومی؛ $2 متن پیوند

همان طور که در جدول 1 مشاهـده می شود در هر دو قالب یکی از نویسه های برچسب پیشـینه می توانـد نشان دهندهٔ منابع الکترونیکی باشد و در توضیحات ذیل کد آمده که شامل منابع پیوسته نیز هست. «برچسب پیشینه در ابتدای هر پیشینه قرار می گیرد و حاوی داده های مربوط به پردازش آن پیشـینه است» (مارک ایران 1381 ، 28) و به طور غیر مسـتقیم برای کاربرد در تشخیص خود مدرک کتاب شناختی به کار می رود (International Federation of Library Associations and Institutions (IFLA). 2008 12).

همچنین نشان گرهای یکی از فیلدها (856) برای نمایش خصوصیات منابع وبی تعریف شده است. داده های جدول 1 نشان می دهد که در قالب مارک ،21 می توان در سیزده فیلد، و در قالب یونیمارک، می توان در دوازده فیلد اطلاعات مربوط به منابع و بی را وارد کرد.

برای پاسـخ به پرسـش دوم ،پژوهش از خصوصـیات تعریف شـده برای منـابع اطلاعاتی وب توسط هیری(1) (1996) اسـتفاده و عناصر تعریف شده در دو قالب یونی مارک و مارک 21 مقایسه شد که نتایج آن در جدول 2 قابل مشاهده است.

عکس

همان‌طور که در جدول ۱ مشاهده می‌شود، در هر دو قالب، یکی از نویسه‌های برچسب پیشینه می‌تواند نشان دهندهٔ منابع الکترونیکی باشد و در توضیحات ذیل کد آمده که شامل منابع پیوسته نیز هست. «برچسب پیشینه در ابتدای هر پیشینه قرار میگیرد و حاوی داده‌های مربوط به پردازش آن پیشینه است» (مارک ایران، ۱۳۸۱، ۲۸) و به‌طور غیرمستقیم برای کاربرد در تشخیص خود مدرک کتاب‌شناختی به‌کار می‌رود (International Federation of Library Associations and Institutions (IFLA). 2008. 12).

همچنین، نشانگرهای یکی از فیلدها (۸۵٦) برای نمایش خصوصیات منابع وبی تعریف شده است. داده‌های جدول ۱ نشان می‌دهد که در قالب مارک ۲۱، می‌توان در سیزده فیلد، و در قالب یونی‌مارک، می‌توان در دوازده فیلد اطلاعات مربوط به منابع وبی را وارد کرد.

برای پاسخ به پرسش دوم پژوهش، از خصوصیات تعریف شده برای منابع اطلاعاتی وب توسط هیری[1] (۱۹۹٦) استفاده، و عناصر تعریف شده در دو قالب یونی‌مارک و مارک ۲۱ مقایسه شد که نتایج آن در جدول ۲ قابل مشاهده است.

جدول ۲: مقایسه خصوصیات منابع اطلاعاتی وب با عناصر موجود در قالب‌های یونی‌مارک و مارک ۲۱

ردیف	ویژگیها	مارک ۲۱	یونیمارک
۱	اطلاعات مربوط به مکان‌های مختلف یک منبع	فیلد ۸۵۶ و قابلیت تکرارپذیری آن برای ثبت مکان‌های مختلف یک منبع	فیلد ۸۵۶ و قابلیت تکرارپذیری آن برای ثبت مکان‌های مختلف یک منبع
۲	شیوه (های) دسترسی به منبع	نشانگر اول در فیلد ۸۵۶ و قابلیت تکرارپذیری برای بیش از یک روش دسترسی فیلد فرعی $۲ در فیلد ۸۵۶ فیلد ۷۵۳	نشانگر اول در فیلد ۸۵۶ و قابلیت تکرارپذیری برای بیش از یک روش دسترسی فیلد فرعی $y در فیلد ۸۵۶ فیلد ۶۲۶
۳	قالب های مختلف نسخههای مربوط به یک منبع (PDF, HTML, XML, ...)	فیلد فرعی $q در فیلد ۸۵۶ فیلد فرعی $y در فیلد ۸۵۶ فیلد ۳۳۸	فیلد فرعی $q در فیلد ۸۵۶ فیلد فرعی $b در بلوک ۴---
۴	ثبت تغییرات ناشی از نبود ثبات	تکرارپذیر بودن فیلد ۸۵۶ در فیلدهای فرعی $a, $b, $c	تکرارپذیر بودن فیلد ۸۵۶ در فیلدهای فرعی $a, $b, $c
۵	روزآمد سازی منابع	————	————
۶	سطح توصیف	————	————
۷	اطلاعات مربوط به دسترسی و شرایط آن	فیلد ۵۳۸ فیلد ۲۵۶	فیلد ۳۳۷ فیلد ۲۳۰

داده‌های جدول ۲ نشان می‌دهد که برای ۵ مورد از خصوصیات بیان شده توسط هیری، حداقل یک عنصر در یونی‌مارک و مارک ۲۱ تعریف شده است. همچنین، برای نمایش دو مورد از خصوصیات نیز

1. Heery

جدول 2: مقایسه خصوصیات منابع اطلاعاتی وب با عناصر موجود در قالب های یونی مارک و مارک 21

داده های جدول 2 نشان می دهد که برای 5 مورد از خصوصیات بیان شده توسط هیری حداقل یک عنصر در یونی مارک 21 تعریف شده است. همچنین، برای نمایش دو مورد از خصوصیات نیز

عنصری در مارک 21 و یونی مارک یافت نشد.

برای پاسخ به پرسش سوم پژوهش با استفاده از داده های به دست آمده در جدول 1، به مقایسه عناصر تعریف شده در مارک 21 و یونیمارک پرداخته شد که نتایج این بررسی در جدول 3 نمایش داده شده است.

عکس

عنصری در مارک ۲۱ و یونی‌مارک یافت نشد.

برای پاسخ به پرسش سوم پژوهش، با استفاده از داده‌های به‌دست آمده در جدول ۱، به مقایسه عناصر تعریف شده در مارک ۲۱ و یونیمارک پرداخته شد که نتایج این بررسی در جدول ۳ نمایش داده شده است.

جدول ۳: مقایسه عناصر موجود در مارک ۲۱ و یونیمارک برای نمایش خصوصیات منابع اطلاعاتی وب

مارک ۲۱	یونیمارک
نمایش خصوصیات منابع اطلاعات وب در ۱۳ فیلد	نمایش خصوصیات منابع اطلاعات وب در ۱۲ فیلد
عدم نمایش شکل مدرک در فیلدهای شناسه رابط	نمایش شکل مدرک در فیلدهای شناسه رابط
تعیین نوع مدرک در فیلد داده‌ای (۰۰۸)	عدم تعیین نوع مدرک در فیلدهای داده‌ای
تعیین نوع حامل بر اساس استاندارد آر.دی.ای.	عدم تعیین نوع حامل بر اساس استاندارد آر.دی.ای.
در فیلد ۸۵۶ نشانگر دوم برای نمایش روابط تعریف شده است.	در فیلد ۸۵۶ نشانگر دوم تعریف نشده است.

داده‌های موجود در جدول ۳ نشان می‌دهد که قالب‌های یونی‌مارک و مارک ۲۱ برای نمایش خصوصیات منابع وبی در ۵ مقوله با یکدیگر متفاوت هستند. در قالب مارک ۲۱، در ٤ مقوله به نمایش خصوصیات و اختصاص عناصر جزئیتر در نمایش آنها توجه بیشتری در مقایسه با یونی‌مارک شده است. همین تفاوت در قالب یونی‌مارک در یک مقوله نسبت به مارک ۲۱ دیده می‌شود.

برای پاسخ به پرسش چهارم، به تجزیه و تحلیل نتایج به‌دست آمده در پرسش‌های اول و سوم پژوهش پرداخته شد که نتایج آن به قرار زیر است:

- فراوانی فیلدهای اختصاص داده شده به نمایش خصوصیات منابع وبی: با توجه به داده‌های موجود در جدول ۳، تفاوت مارک ۲۱ و یونی‌مارک در این مورد یک فیلد است. جدول ۱ نشان می‌دهد که فیلد فرعی $b در فیلدهای شناسه رابط یونی‌مارک نشان دهندهٔ قالب مدرک پیوندی است. با توجه به ویژگی بلوک شناسه رابط در قالب یونی‌مارک، با نمایش قالب پیوسته در این فیلدها، علاوه بر مشخص شدن قالب پیوستهٔ پیشینه‌های مرتبط با پیشینه اصلی، در هنگام بازیابی اطلاعات، قالب پیوستهٔ مرتبط با مدرک بازیابی شده نیز مشخص خواهد شد.

در مارک ۲۱، فیلد ۰۰٦ مربوط به نمایش ویژگی مواد اضافی است. تفاوت این فیلد با فیلد ۰۰۸ این است که این فیلد برای مواردی که نمی‌توان آنها را در فیلد ۰۰۸ وارد کرد، آورده می‌شود. فیلد ۰۰٦ بیشتر برای مواردی که دارای چند نوع از یک گروه خصوصیت هستند استفاده می شود. به نظر می‌رسد فیلد ۰۰٦ جایگزینی برای فیلد فرعی $b در بلوک --٤ یونی‌مارک باشد، اگرچه فیلد فرعی $b علاوه بر نمایش قالب‌های دیگر مدرک مربوط، نمایش دهندهٔ قالب تمام پیشینه‌های مرتبط (بر اساس انواع روابط موجود در میان آثار) است. علاوه بر این، در مارک ۲۱ بر اساس استاندارد آر.دی.ای، فیلد ۳۳۸ نشان دهندهٔ نوع حامل است که در یونی‌مارک تعریف نشده است.

- تعیین نوع مدرک در فیلد داده‌ای: در مارک ۲۱، علاوه بر تعیین نوع فایل رایانه‌ای در فیلد داده‌ای،

جدول 3: مقایسه عناصر موجود در مارک 21 و یونیمارک برای نمایش خصوصیات منابع اطلاعاتی وب

مارک 21

داده های موجود در جـدول 3 نشـان می دهـد که قالب های یونی مارک و مارک 21 برای نمـایش خصوصـیات منابع و بی در 5 مقوله با یکدیگر متفاوت هستند در قالب مارک ،21، در 4 مقوله به نمایش خصوصـیات و اختصاص عناصر جزئیتر در نمایش آن ها توجه بیشتری در مقایسه با یونی مارک شده است. همین تفاوت در قالب یونی مارک در یک مقوله نسبت به مارک 21 دیده می شود.

برای پاسخ به پرسـش چهارم به تجزیه و تحلیل نتایج به دست آمده در پرسـش های اول و سوم پژوهش پرداخته شد که نتایج آن به قرار زیر است:

– فراوانی فیلدهای اختصاص داده شده به نمایش خصوصیات منابع وبی: با توجه به داده های موجود در جدول 3 تفاوت مارک 21 و یونی مارک در این مورد یک فیلد است جدول 1 نشان می دهد که فیلد فرعی b در فیلدهای شـناسه رابط یونی مارک نشان دهندۀ قالب مدرک پیوندی است. با توجه به ویژگی بلوک شـناسه رابط در قالب یونی مارک با نمایش قالب پیوسته در این فیلدها، علاوه بر مشخص شدن قالب پیوستۀ پیشینه های مرتبط با پیشینه اصلی در هنگام بازیابی اطلاعات، قالب پیوسته مرتبط با مدرک بازیابی شده نیز مشخص خواهد شد.

در مارک ،21 فیلد 006 مربوط به نمایش ویژگی مواد اضافی است تفاوت این فیلد با فیلد 008 این است که این فیلد برای مواردی که نمی توان آن هـا را در فیلـد 008 وارد کرد آورده می شـود. فیلـد 006 بیشتر برای مواردی که دارای چنـد نوع از یک گروه خصوصـیت هسـتند اسـتفاده می شود. به نظر می رسـد فیلد 006 جایگزینی برای فیلـد فرعی b در بلوک – یونی مـارک ،باشـد اگر چه فیلـد فرعی b علاوه بر نمایش قالبهای دیگر مدرک مربوط نمایش دهندۀ قالب تمام پیشینه های مرتبط (بر اساس انواع روابط موجود در میان آثار) است. علاوه بر این در مارک 21 بر اساس استاندارد آر.دی.ای فیلد 338 نشان دهنده نوع حامل است که در یونی مارک تعریف نشده است.

– تعیین نوع مدرک در فیلد داده ای: در مارک 21، علاوه بر تعیین نوع فایل رایانه ای در فیلد داده ای

شکل مدرک نیز با کدی مشخص در نویسه 23 مشخص می شود باید توجه داشت که در مارک 21 مشخصات تمام انواع قالب ها در فیلد داده ای 008 وارد می شود، در صورتی که در یونی مارک بلوک --1به اطلاعات کد شده اختصاص یافته و برای انواع مدارک دارای فیلد خاصی است. فیلد 135 مختص داده های کد شده برای منابع الکترونیکی است به همین دلیل نیازی به تعریف کدی برای شکل مدرک نبوده است.

- تعریف نشان گر برای نمایش :روابط در مارک ،21 نشانگر دوم برای نمایش رابطه بین اطلاعات موجود در فیلد 856 و منبع توصیف شده در پیشینه است. این نشان گر ممکن است برای تولید نمایشی پیوسته یا نظم همگانی فیلدهای 856 استفاده شود (Network Development and MARC 2003 .Standards Office, Library of Congress) فیلد فرعی 3(1) نشان می دهد. به نظر می رسد نمایش روابط در فیلد 856 (حداقل برای منابع وبی) مناسب تر ،باشد زیرا با توجه به این که تمام مشخصات منبع الکترونیکی (مانند مسیر شناس گر متحد الشکل منبع، و غیره) در فیلد 856 وارد می شود می توان قالب وبی هر منبع را (در صورت وجود) در پیشینه اصلی منبع وارد کرد و دیگر نیازی به تهیه پیشینه جدید برای قالب و بی آن نیست.

6-بحث و نتیجه گیری افزایش روزافزون منابع اطلاعاتی وب و آشفتگی موجود در بازیابی اطلاعات جامع و مانع، لزوم سازماندهی منابع اطلاعاتی وب را پر رنگ تر می سازد از آن جایی که منابع وبی سازماندهی شده جزئی از نظام های ذخیره و بازیابی کتابخانه ای خواهند بود لازم است برای سازماندهی این گونه منابع استانداردهای موجود در سازماندهی اطلاعات کتابخانه ای مورد بررسی قرار گیرند.

نتایج پژوهش حاضر نشان داد که قالب های مارک 21 و یونی مارک عناصری را به نمایش خصوصیات منابع وبی اختصاص داده اند این عناصر خصوصیاتی همچون ،قالب نوع دسترسی، ملزومات دسترسی، و نشانی دسترسی را نمایش می دهند که با نتایج طاهری (1387) مبنی بر توانایی های مارک همخوانی داشت. علاوه بر این عناصر تعریف شده در این دو قالب توانایی نمایش سطح توصیف و روزآمدی را برای منابع اطلاعاتی وب .نداشتند نتایج این بخش از پژوهش با نتایج پژوهش وارد (2001) همخوانی نداشته و با نتایج پژوهش فارد و ریگو (2004) مبنی بر ناتوانی ابر داده ها در سازماندهی پیچیدگی های منابع الکترونیکی همخوانی دارد دو مورد بیان شده از مسائل اصلی در سازماندهی منابع و بی هستند زیرا همین خصوصیات منابع اطلاعاتی وب را از دیگر منابع اطلاعاتی متمایز و تصمیم گیری در چگونگی

1- نیز برای نمایش اطلاعات بیش تر در مورد وضعیتی است که در آن یک رابطه یک به یک وجود نداشته باشد. برای مثال نشان گر 1 نمایش گر وضعیتی است که مدرک توصیف شده در پیشینه کتاب شناختی الکترونیکی نیست اما یک نسخه الکترونیک از آن موجود و اطلاعات ثبت شده در فیلد 856 نشان دهنده نسخه الکترونیکی است (Library of Congress. 2003) همان طور که در بالا اشاره شد یونی مارک برای نمایش روابط از بلوک شناسه رابط استفاده می کند و قالب مدرک را در فیلد فرعی b

سازماندهی آن ها را چالش برانگیز نموده است. مقایسه دو قالب کتاب شناختی یونی مارک و مارک 21 نیز نشان داد که تفاوت چندانی در میان تعریف عناصر مربوط به قالب و بی وجود ندارد و تنها تفاوت هایی در شکل تعریف آن ها مشاهده شد به نظر می رسد در انتخاب یکی از دو ،قالب باید به وضعیت کنونی دو قالب در ایران میزان انطباق هر یک با استانداردها و الگوهای مطرح در سازماندهی اطلاعات وب، و همچنین وضعیت هر یک از آن ها در سازماندهی دیگر قالب های اطلاعاتی توجه شود.

از یافته های این پژوهش می توان چنین دریافت که قالب مارک می تواند به عنوان استانداردی برای ذخیره و بازیابی منابع اطلاعاتی وب به کار رود، با این حال مارک راه حلی برای چگونگی روزآمد نمودن و تعیین سطح توصیف منابع اطلاعاتی وب تهیه نکرده است پژوهشگران برای ویژگی سطح توصیف، استفاده از فیلدهای رابطه ای را پیشنهاد می کنند به این صورت که برای نمایش اجزای یک منبع وبی (مانند صفحه های یک وبگاه یا همان پیوندهای درونی) از فیلدهای رابطه ای کل و جزء و برای نمایش پیوندهای خارجی از فیلدهای رابطه ای که نشان دهنده روابطی همچون هم ارز هستند، استفاده شود. پژوهش گران برای ویژگی روزآمدی استفاده از فیلدی خاص برای ثبت تغییرات را پیشنهاد می دهند. این فیلد شامل فیلدهای فرعی باشد که هر یک نشان دهنده یکی از تغییرات احتمالی (تغییر در آدرس، عنوان و غیره) در منابع اطلاعاتی وب و تعیین نشانگرها برای نمایش نوع تغییر (روزآمد شده، حذف شده، جایگزین شده و غیره) باشد با توجه به این که پژوهش حاضر به بررسی قالب مارک برای سازماندهی منابع اطلاعاتی وب به طور کلی پرداخته است انجام پژوهش هایی برای بررسی قالب مارک برای انواع منابع اطلاعاتی وب و ارزیابی توانایی های آن ضروری به نظر می رسد.

منابع

حاجی زین العابدینی محسن 1381 بررسی مسائل فهرست نویسی منابع اینترنتی و ارائه دست نامه پیشنهادی برای کتابخانه های ایران پایان نامه کارشناسی ارشد دانشگاه علوم پزشکی ایران. دانشکده مدیریت و اطلاع رسانی پزشکی

طاهری، سید مهدی 1387. مقایسه کارآیی طرح فراداده ای هسته دوبلین و قالب فراداده ای مارک 21 در سازماندهی منابع اطلاعاتی شبکه جهانی وب فصلنامه کتابداری و اطلاع رسانی، 3(11): 139-158.

عمرانی ابراهیم 1390 چارچوب کتاب شناختی برای دوران دیجیتال خبرنامه انجمن کتابداری و اطلاع رسانی (27) 21-31

فتاحی، رحمت الله .1380 چالش های سازماندهی دانش در قرن بیستم فصلنامه کتاب، 12(4): 59-83 کمیته ملی مارک ایران 1381 مارک ایران تهران کتابخانه ملی جمهوری اسلامی ایران

کوکبی، مرتضی، سمیه سادات آخشیک 1385 سازماندهی منابع اینترنتی: قواعد فهرست نویسی انگلو-امریکن یا عناصر فرادادهای هسته دوبلین؟ سازماندهی اطلاعات: رویکردها و راهکارهای نوین (مجموعه) مقالات اولین همایش انجمن کتابداری و اطلاع رسانی ،ایران 16 و 17 اسفند 1385) (صص. 333 -344) تهران: کتابدار

ملاـی ،مقـدم ،گلناز محمـد نعیم آبادی 1385 سازمانـدهی صـفحات وب بـا اسـتفاده از نظـام های رده بنـدی کتابخانه ای سازمانـدهی اطلاعـات: رویکرد هـا و راهکـار هـای نوین مجموعه مقالاـت اولین همـایش انجمن کتابـداری و اطلاع رسانی ایران ، 16 و 17 اسفند 1385) (صص. 345-364). تهران: کتابدار.

Fard, S.E., A. Riggio. 2004. Medium or message? A new look at standards, structures, and schemata for managing electronic resources. Library Hi Tech, 22(2): 144-152.

Gorman, M. 2004. Authority Control in the Context of Bibliographic Control in the Electronic Environment. Cataloging Classification Quarterly, 38(3-4). Retrieved on September 19, 2011, from http://www.sba.unifi.it/ac/relazioni/gorman-eng.pdf.

Heery,R. 1996. Review of metadata formats. Program: Electronic Library and Information Systems, 30(4): 345-373.

International Federation of Library Associations and Institutions (IFLA). 2000. UNIMARC guideline no.6: electronic resources. Retrieved on September 20, 2012, from http:// archive.ifla.org/VI/3/p1996-1/guid6.htm.

International Federation of Library Associations and Institutions (IFLA). 2001. UNIMARC to MARC 21 conversion specifications. Retrieved on September 12, 2012, from http:// www.loc.gov/marc/unimarctomarc21.html.

International Federation of Library Associations and Institutions (IFLA). 2008. UNIMARC manual. München: K. G. Saur.

Library of Congress 2003. MARC 21 format for bibliographic data: 856: Electronic location and access. Retrieved on September 19, 2012, from http://www.loc.gov/marc/ bibliographic/bd856.html.

Library of Congress 2012. MARC 21. Retrieved on September 19, 2012, from www.loc. gov/marc/translations.html

Miller, S. 2008. Rules and tools for cataloging internet resources (trainee manual). Retrieved on September 29, 2012, from http://www.loc.gov/catworkshop/courses/catalogininternet/ pdf/ceig1-IM-FINAL.pdf.

Network Development and MARC Standards Office, Library of Congress 2003. MARC 21 formats: Guidelines for the use of field 856. Retrieved on September 29, 2012, from http://www.loc.gov/marc/856guide.html.

Ward, D. 2001. Internet resource cataloging: the SUNY Buffalo Libraries' response. OCLC Systems .Services, 17(1): 19–26

Weitz, J. 2006. Cataloging electronic resources: OCLC–MARC coding guidelines. Retrieved on September/ 20, 2012, from http://www.oclc.org/support/documentation/worldcat/ cataloging/electronicresources

چکیده

هدف پژوهش حاضر سنجش کیفیت محیط های رابط کاربر پایگاه های اطلاعاتی مجلات پیوسته تمام متن فارسی «سید»، «مگ ایران»، «نما متن»، «نورمگز»، «پژوهشگاه علوم و فناوری اطلاعات» و «مرکز منطقه ای» از طریق بررسی و مقایسه آن ها با معیارها و استانداردهای رعایت شده در سیاهه وارسی رابط کاربر پایگاه اطلاعاتی می باشد. روش پژوهش حاضر پیمایشی -توصیفی است گردآوری داده ها با استفاده از یک سیاهه وارسی محقّق ساخته و از طریق مشاهده ی مستقیم انجام گرفته است. یافته ها نشان داد که پایگاه اطلاعاتی «پژوهشگاه علوم و فناوری اطلاعات» در مجموع با رعایت 35 مؤلفه (55/55 درصد) از مؤلفه های سیاهه وارسی برترین محیط رابط کاربری را از میان پایگاه های اطلاعاتی پیوسته مجلات تمام متن فارسی دارا می باشد محیط های رابط کاربری پایگاه های اطلاعاتی نور ،مگز، مرکز منطقه ای و نمامتن به ترتیب با میزان رعایت 34 مؤلفه (53/96 درصد)، 32 مؤلفه (50/79 درصد) و 28 مؤلفه (44/44 درصد) رتبه های دوّم سوّم و پنجم را به خود اختصاص دادند همچنین محیط کاربری دو پایگاه مگ ایران و سید به طور مساوی با میزان رعایت 31 مؤلفه (49/20 درصد) در جایگاه چهارم قرار گرفتند نتایج پژوهش نشان داد که محیط رابط کاربری پایگاه های تمام متن فارسی در امکانات نوینی که طی سال های اخیر پدیدار شده اند دچار کمبود می باشند از این رو پیشنهاد می شود طراحان پایگاه های مذکور به طراحی مجدد پایگاه های خود پرداخته و این امکانات و قابلیت ها را به پایگاه های خود بی افزایند.

کلیدواژه: رابط کاربر پایگاه های اطلاعاتی پیوسته مجلات تمام متن مجلات فارسی

سنجش رابط کاربر پایگاه های اطلاعاتی پیوسته: مجلات تمام متن فارسی

اشاره

صدیقه جعفر زاده (1) معصومه ،پیروزفر، (2) عبدالحسین فرج پهلوی (3)

مقدّمه

امروزه با رشـد و گسترش شـبکه جهانی اینترنت پایگاه های اطلاعاتی الکترونیک به عنوان مهم ترین و پر کاربرد ترین ابزار فراهم کننـده ی دسترسـی به اطلاعات تبـدیل شده اند به طوری که پژوهشـگران و محققان می توانند اطلاعات مورد نیاز خود را در میان انبوه اطلاعات از طریق این محمـل های اطلاعاتی، به دقت و سـرعت جسـتجو و بازیابی کننـد نخستین نقطه برخورد کاربر با پایگاه های اطلاعاتی، محیط رابـط کـاربر است. در پایگاه های اطلاعاتی رابط کاربر عامل مهمی در تسـهیل دستیابی کاربران به اطلاعات مورد نیاز خود محسوب می شود از این رو میزبان ها و تولید کنندگان این پایگاه ها می کوشـند تا ضـمن در نظر گرفتن نیاز کاربران و با پیروی از اصول موجود عوامل و خصـیصه های ضروری را جهت طراحی رابط کاربر پایگاه های اطلاعاتی خود شناسایی کرده و از این طریق دسترسی مؤثر کاربران نهایی را به اطلاعات موجود تضمین کنند (مهراد و زاهدی 1386).

یکی از نمودهای عینی کیفیّت استفاده کاربر از پایگاه های اطلاعاتی به ویژه پایگاه های اطلاعاتی پیوسته

1- دانشجوی کارشناسی ارشد کتابداری و اطلاع رسانی دانشگاه شهید چمران sedighehjafarzadeh88@gmail.com

2- دانشجوی کارشناسی ارشد کتابداری و اطلاع رسانی دانشگاه شهید چمران piruzfar@gmail.com

3- عضو هیئت علمی و مدیر گروه کتابداری و اطلاع رسانی دانشگاه شهید چمران farajpahlou@gmail.com

را می توان در قالبی با عنوان رابط کاربر مشاهده کرد (یمین فیروز 1382). در واقع رابط کاربر نخستین نقطه ی برخورد کاربر با پایگاه های اطلاعـاتی است و به عنوان پلی ارتبـاطی بین انسـان و سامـانه ی اطلاعاتی عمل می کنـد. به همین دلیل مهم ترین هـدف از طراحی رابط کـاربر برآوردن رضـایت کـاربران و ایجـاد تعامـل بیش تر و بهتر بین کاربر و محیط های رایانه ای است (Shneiderman 1998 نقل در اعظمی و فتاحی 1388).

با توجه به این که رابط کاربر خوب باعث می شود کاربران مسیر خود را در پایگاه اطلاعاتی بهتر شناسایی کنند و تأثیر به سزایی در عملکرد آنان خواهد داشت (Bates 2002 ، نقل در علیجانی و دهقانی 1386)؛ بنابراین محیط رابط کاربر بایـد به کاربران کمک کند تا کلیدواژه ها و عبارت های مناسبی برای جسـتجوهای خود به کار ببرنـد منبع مورد نظر را از میان منابع اطلاعاتی انتخاب کنند، نتایج جسـتجو ها را درک و از چگونگی پیشرفت کارشان آگاه شوند تا بتوانند آسان تر و بهتر به اطلاعات مورد نیاز خود دست یابند (Hirst1999)

در کشور ما نیز در دو دهه ی اخیر تعدادی پایگاه اطلاعاتی پیوسته ی تمام متن طراحی شده است که پژوهش گران و محققان بسیاری برای انجـام کارهـای علمی و پژوهشـی و رفع نیازهـای اطلاعاتی خود به آن ها مراجعه می کننـد با توجه به نقش مهم رابط کاربر در دستیابی به محتوای پایگاه های اطلاعاتی و بالا بردن کیفیت خـدمات ،آنان محیط رابط این پایگاه ها بایـد به گونه ای طراحی شده باشد که دسترسـی منـاسب و مؤثر به اطلاعـات را برای کاربران تسـهیل بخشـد؛ به طوری که آنان بتواننـد با درک روشـنی از نحوه ی طراحی پایگاه اطلاعاتی ساختار اطلاعات و چگونگی ارائه اطلاعات به ،کاربران اطلاعات مورد نیاز خود را در پایگاه مورد نظر به سـرعت جستجو و بازیابی کنند. بنابراین بررسـی و مقایسه محیط رابط کاربر پایگاه های اطلاعاتی پیوسته مجلات تمام متن فارسی با معیارهای رعایت شده در پایگاه های اطلاعاتی معتبر پر کاربرد و اسـتاندارد جهانی امری ضروری است این امر ضمن نشان دادن فاصله بین وضع موجود تا وضع مطلوب سبب بهبود کیفیت ارائه ی اطلاعات در پایگاه های اطلاعاتی مجلات فارسی می شود.

بیان مسأله و ضرورت پژوهش

پایگاه های اطلاعات علمی پیوسـته مهم ترین و پرکاربرد ترین منابعی هستند که امروزه پژوهشگران جهت دسترسی به اطلاعات مورد نیاز خود به آن ها مراجعه می کننـد در ایران نیز در همین راسـتا طراحی پایگاه های اطلاعاتی الکترونیکی آغاز شـده و فعالیت های پراکنده ای توسط برخی سازمان ها در خصوص ارائه ی شـکل الکترونیکی مجلات به صورت پیوسته یا ناپیوسته انجام گرفته است. از بین پایگاه هایی که به طور پیون سـته شـکل الکترونیکی نشـریه های ایرانی را ارائه می دهند چهار پایگاه اطلاعاتی «سید» (1)، «مگ ایران»، «نما متن» و «نور مگز» جـامعیت و پوشـش موضوعی گسترده تری دارنـد. از طرفی، مـدت زمان بیش تری از طراحی آن ها می گـذرد بنابراین امکانات بالاتری نسـبت به سایر پایگاه ها دارند و برای کاربران شـناخته شده تر هسـتند (اسداللهی و نوکاریزی 1389) در پژوهش حاضر علاوه بر این چهار پایگاه که صرفاً پایگاه های اطلاعاتی نشریات می باشند به بررسی محیط رابط کاربری پایگاه های اطلاعاتی مجلات

1- منظور «پایگاه اطلاعات علمی جهاد دانشگاهی» می باشد.

فارسی «پژوهشگاه علوم و فناوری اطلاعات» و «مرکز منطقه ای» (1) که دارای پایگاه مجلات و پایگاه های گوناگون دیگری نیز می باشند پرداخته می شود.

با توجّه به استفاده گسترده کاربران از پایگاه های اطلاعاتی به عنوان منابع بازیابی اطلاعات و علی رغم تأکید متخصصان بر اهمیت رابط کاربر پایگاه ها و وب سایت ها پژوهش های انجام شده در این زمینه نشان می دهد که طراحان در پاره ای از موارد به دلیل در نظر نگرفتن شرایط و ویژگی های لازم از این مسئله غفلت کرده و در نتیجه مشکلاتی را در خصوص استفاده بهینه کاربران از خدمات خود به وجود آورده اند (Hansen 1998) همچنین بررسی ها نشان داد که به نظر می آید محیط رابط کاربری برخی پایگاه های اطلاعاتی پیوسته مجلات فارسی موجود در ایران در پاره ای موارد با استاندارد ها هماهنگی ندارند و دارای برخی کمبود ها می باشند از این رو وجود این مسئله پژوهش گران را به این امر واداشت تا بررسی و ارزیابی کنند که وضعیت محیط رابط کاربر پایگاه های اطلاعاتی پیوسته تمام متن فارسی در مقایسه با معیارهای رعایت شده در محیط رابط کاربری پایگاه های اطلاعاتی معتبر بین المللی چگونه است؟ و در این راستا نقاط قوت و ضعف این پایگاه ها کدامند؟ نتایج حاصل از این پژوهش می تواند سبب استفاده ،بهتر افزایش موفقیت در جستجو ها و انجام بهتر و سریع تر بازیابی اطلاعات شود.

هدف پژوهش

هدف پژوهش حاضر سنجش کیفیت محیط های رابط کاربر پایگاه های اطلاعاتی مجلات پیوسته تمام متن فارسی «پژوهشگاه علوم و فناوری اطلاعات»، «سید»، «مرکز منطقه ای»، «مگ ایران»، «نما متن» و «نور مگز» از طریق بررسی و مقایسه آن ها با معیارهای رعایت شده در محیط رابط کاربر پایگاه های اطلاعاتی مجلات پیوسته تمام متن معتبر و برجسته بین المللی می باشد. به منظور دستیابی به این هدف نقاط قوّت و ضعف موجود در محیط رابط کاربری این پایگاه ها مشخص می شود

با توجّه به اهداف فوق سعی می شود به پرسش های زیر پاسخ داده شود:

1. وضعیت محیط رابط کاربری هر یک از پایگاه های اطلاعاتی مورد بررسی از لحاظ مقوله جستجو چگونه است؟

2. وضعیت محیط رابط کاربری هر یک از پایگاه های اطلاعاتی مورد ،بررسی از لحاظ مقوله نمایش اطلاعات چگونه است؟

3. وضعیت محیط رابط کاربری هر یک از پایگاه های اطلاعاتی مورد بررسی از لحاظ مقوله خدمات چگونه است؟

4. وضعیت محیط رابط کاربری هر یک از پایگاه های اطلاعاتی مورد بررسی از لحاظ مقوله پیوندها چگونه است؟

5. وضعیت محیط رابط کاربری هر یک از پایگاه های اطلاعاتی مورد بررسی از لحاظ مقوله راهنمایی چگونه است؟

1- منظور «مرکز منطقه ای اطلاع رسانی علوم و فناوری» می باشد

6. کـدام یـک از مقوله هـای کلی رابـط کـاربری پایگـاه اطلاعاتی بیش ترین میزان رعایت را در پایگاه های اطلاعاتی مجلات الکترونیک پیوسته تمام متن فارسی دارد؟

7. کدام یک از پایگاه های اطلاعاتی مجلات الکترونیک پیوسته تمام متن فارسی بیش ترین انطباق را با معیارهای موجود در سیاهه وارسی محیط رابط کاربری پایگاه اطلاعاتی دارد؟

پیشینه ی پژوهش

سابین (2001) (1) رابـط کاربر میزبان های پایگاه های اطلاعاتی را بر اساس پنـج طبقه کلی عملکرد ارزیابی کرد و سـپس خصوصـیات بیش تری را که مربوط به هر طبقه بود به عنوان خصیصه های ضروری، مطلوب یا مورد نیاز مشخص کرد یافته های این مطالعه نشان داد که دایالوگ (2) با 32 امتیاز از مجموع 41 امتیاز بالاترین رتبه را دارد. او سـی ال سی (3)، اوید (4) با 27 امتیـاز در رتبه دوم پروکوئست و سیلور پلاتر (5) با 26 امتیاز در مقام سوم قرار دارند. نکسیس (6) با کسب 25 امتیاز در رتبه هفتم و ویلسون (7) با 20 امتیاز مقام آخر را کسب کرد. در پایان این پژوهش خصـیصه های سیسـتم ایده آل مشـخص شد بر این اساس دایالوگ به این سیستم بسیار نزدیک است اما هنوز فاقد تعدادی از خصیصه های مهم می باشد.

کراستینی (2004) (8) یک رابط کاربر گرافیکی جدید برای بازیابی مدارک به صورت سلسـله مراتبی ارائه کرده و سپس به نحوهی طراحی، اجرا و ارزیـابی رابـط کاربر پیشـنهادی پرداخته است. نتایـج این پژوهش حاکی از آن است که رابط کاربر پیشـنهادی ابزارهای قدرتمنـد و مؤثری را برای جستجوی مدارک، رهیابی فهرست بازیابی شده و تصحیح جستجو فراهم می آورد.

شاید بتوان نخستین تلاش در ارزیابی رابط کاربر پایگاه های اطلاعاتی کتاب شـناختی فارسی را پژوهش مهراد و زاهدی (1386) دانست. آنـان در پژوهش خود به بررسـی و مقـایسه محیـط رابـط کـاربر دو میزبـان داخلی (کتابخانه منطقه ای علوم و تکنولوژی (9) و پژوهشگاه اطلاعات و مدارک علمی ایران) با چهار میزبان خارجی (الزویر، امرالد، ابسکو و پروکوئست) ارائه دهنده ی پایگاه های اطلاعاتی به روش پیمایش تطبیقی پرداختنـد. با اسـتفاده از سـیاهه وارسـی جامع با پنـج خصـیصه (خصـیصه های کلی، جسـتجو، بازیابی، نمایش و کاربر پسـندی) تلاش شد تا ضـمن شـناخت ویژگی ها و نقاط قوت و ضعف هر یک رابط کاربر میزبان های داخلی و خارجی با یکدیگر مقایسه شونـد یـافته هـای این پژوهش نشـان داد در میزبـان هـای ،داخلی به ترتیب کتابخانه منطقه ای علوم و تکنولوژی و پژوهشـگاه اطلاعات و مدارک علمی ایران و در میزبان های خارجی به ترتیب ابسکو، پروکوئست امرالد و الزویر در پنج خصیصه مورد بررسی دارای

1- Sabin

2- Dialogweb

3- OCLC First Search

4- Ovid

5- Silver Platter

6- Nexis

7- WilsonWeb

8- Crastini

9- منظور «مرکز منطقه ای اطلاع رسانی علوم و فناوری» می باشد

بیش ترین ویژگی ها بودند.

اعظمی و فتاحی (1388) به تعیین همخوانی محیط رابط پایگاه های اطلاعاتی ابسکو، امرالد پروکوئست و ساینس دایرکت با عناصر رفتار اطلاع یابی مدل «الیس» (1) که عبارتند از: شروع پیوندیابی ،مرور تمایز ،بازبینی و استخراج پرداختند نتایج نشان داد که عناصر «شروع» «پیوندیابی» و «تمایز» تا حدودی به وسیله ی برخی از محیط های رابط کاربر پایگاه های مورد بررسی حمایت می شوند، اما دیگر عناصر رفتار اطلاع یابی (تورق بازنگری و استخراج) در ساختار رابط کاربر این پایگاه ها لحاظ نشده اند. به طور کلی میزان مطابقت و همخوانی رابط کاربر پایگاه های اطلاعاتی با عناصر رفتار اطلاع یابی الیس در حد متوسط است بنابراین استفاده از این عناصر در طراحی و ارزیابی محیط رابط کاربری می تواند تأثیر زیادی بر بهینه شدن محیط رابط پایگاههای اطلاعاتی و در نتیجه بر فرآیند جستجو و بازیابی داشته باشد.

انتظاریان و فتاحی (1389) به تحلیل تبیین و شناسایی نقاط قوت و ضعف عناصر و ویژگی های مهم در محیط رابط پایگاه های اطلاعاتی مقاله های الکترونیکی مرکز منطقه ای اطلاع رسانی علوم و فناوری و پژوهشگاه اطلاعات و مدارک علمی پرداختند آن ها در این بررسی میزان همخوانی محیط رابط پایگاه های مورد بررسی با 10 مؤلفه نیلسن (2)، مشکلات اساسی محیط رابط این پایگاه ها و نیز تفاوت بین میزان در ک کاربران متخصص و مبتدی را مورد سنجش قرار دادند. یافته ها نشان داد میزان همخوانی محیط رابط پایگاه مقاله های پژوهشگاه با 10 مؤلفه نیلسن به طور کلی در حد متوسط و در پایگاه مقاله های مرکز منطقه ای کمی بیش از حد متوسط است هر دو پایگاه در برخی از مؤلفه های مدل نیلسن دارای مشکلات اساسی هستند

اسداللهی و نوکاریزی (1389) طی پژوهشی به ارزیابی ساختار و محتوای پایگاه های اطلاعاتی الکترونیکی ،سید مگ ایران و نما متن پرداختند. میزان مطابقت پایگاه سید، مگ ایران و نمامتن با معیارهای ساختار به ترتیب 50/43 درصد، 54/70 درصد و 41/88 درصد برآورد شد. پایگاه نمامتن بر خلاف دو پایگاه سید و مگ ایران به دلیل استفاده از زبان نمایه سازی کنترل شده، روند مشخص و منسجمی برای نمایه سازی مجله ها داشت. نتایج حاصل از بررسی ویژگی های منحصر به فرد ساختاری محتوایی پایگاه ها مبین آن بود که پایگاه سید از این نظر در رتبه ی نخست قرار دارد. به طور کلی هر سه پایگاه می توانند با توجه بیش تر به معیارهای مطرح شده در سیاهه وارسی بر قابلیت های ساختاری خود بیفزایند در بخش محتوا نیز افزایش جامعیت و توسعه ی پوشش ،موضوعی استفاده از رویکرد ترکیبی یعنی زبان آزاد و کنترل شده، عدم گزینش در نمایه سازی افزایش پوشش گذشته نگر و کاهش دوره ی ،روز آمد سازی در هر سه پایگاه توصیه می شود.

جمع بندی پیشینه

مرور پیشنه ها نشان می دهد که اهمیت بازیابی اطلاعات باعث گردیده پژوهش هایی بر روی پایگاه های

1- Ellis's Information Seeking Behavior Model

2- Nielsen

اطلاعـاتی نشـریات الکترونیک انجام گیرد. این پژوهش ها ابعاد مختلف این پایگاه ها از جمله رفتار اطلاع یابی در این پایگاه ها ،گرافیک ساختـار و محتوای پایگاه ها و به ویژه رابط کاربری این پایگاه ها را مورد بحث قرار داده انـد اما نکته ای که حائز اهمیت است این است که تاکنون پژوهشـی که رابط کاربری تمام پایگاه های شـناخته شده مجلات فارسـی را به صورت یک جا با معیارهای رعایت شده در چندین پایگاه بین المللی مورد ارزیابی قرار دهد انجام نشده است. از این رو در این پژوهش رابط کاربری پایگاه های مجلات فارسی با معیارهای پایگاه بین المللی مورد بررسی قرار می گیرد.

روش شناسی پژوهش

روش پژوهش حاضـر پیمایشـی- توصـیفی است. جامعه ی مورد مطالعه در این پژوهش پایگاه های اطلاعاتی مجلات پیوسـته تمام متن فارسی «پژوهشگاه علوم و فناوری اطلاعات»، «سید»، «مرکز منطقه ای» «مگ ایران»، «نما متن» و «نور مگز» می باشد با توجّه به اهداف پژوهش ابزار گردآوری داده ها با اسـتفاده از یک سـیاهه وارسـی محقق ساخته است بـدین ترتیب که محتوای سـیاهه وارسـی از معیارهای رعـایت شـده در محیـط رابـط کاربر پایگاه های اطلاعاتی مجلات پیوسـته تمام متن معتبر و برجسـته دنیا از قبیل «ساینس دایرکت» (1)، «پروکوئست» (2)، «امرالد»(3)، «ابسکو» (4) و «اشپرینگر» (5) و همچنین بررسی متون مرتبط استخراج و طراحی شد. سپس روایی آن توسط متخصـصان این حوزه بررسی گردید آن گاه با مراجعه مستقیم به وب سایت پایگاه های اطلاعاتی مورد نظر بر اساس سیاهه وارسی به ارزیابی آن ها پرداخته شد. بر اساس معیارهای مطرح شده در سیاهه وارسی به ازای دارا بودن آن معیار عدد یک و در صورت نداشتن آن معیار عـدد صـفر منظور گردید برای تجزیه و تحلیل داده های گردآوری شده از طریق سـیاهه وارسـی نیز از آمار توصـیفی (فراوانی درصـد فراوانی و میانگین) استفاده شد.

یافته های پژوهش

در پاسخ به پرسـش اول پژوهش باید گفت که پایگاه های اطلاعاتی سـید و نور مگز به طور مساوی با میزان رعایت 11 مؤلفه (50 درصد) بیش ترین میزان رعایت معیارهای موجود در سیاهه وارسی را نسبت به پایگاه های اطلاعاتی دیگر داشته اند همچنین پایگاه های اطلاعاتی پژوهشـگاه علوم و فناوری اطلاعات و مگ ایران و مرکز منطقه ای به طور مساوی 10 مؤلفه (45/45 درصد) و پایگاه اطلاعاتی نمامتن 8 مؤلفه (36/36 درصـد) از مؤلفه های مربوط به مقوله ی جسـتجو را در خود رعایت کرده انـد. همچنین از مجموع 22 مؤلفه ی مربوط به مقوله ی جستجو مؤلفه های «جستجوی مجاورتی»، «ریشه سازی»، «توانایی اصلاح استراتژی جستجوی قبلی»، «محدود گرهای مناسب جهت اصلاح جستجو قابلیت جستجوی متن آزاد»، «عملکردهای اصطلاحنامه» و «عملکرد مقالات مرتبط» در محیط رابط کاربری هیچ کدام از پایگاه های

1- ScienceDirect

2- ProQuest

3- Emerald

4- Ebsco

5- Springer

مورد بررسـی رعایت نشده .است مؤلفه های «ذخیره جسـتجوها در حساب کاربری خود» و «تاریخچه جسـتجو» تنها در پایگاه اطلاعاتی پژوهشـگاه علوم و فناوری اطلاعات رعایت شده است. به طور کلی می توان گفت محیط رابط کاربری پایگاه های اطلاعاتی مورد بررسی در این پژوهش به طور میانگین 45/45 درصد از مؤلفه های مربوط به جستجو را رعایت کرده اند (جدول 10)

عکس

مورد بررسی رعایت نشده است. مؤلفه‌های «ذخیره جستجوها در حساب کاربری خود» و «تاریخچه جستجو» تنها در پایگاه اطلاعاتی پژوهشگاه علوم و فناوری اطلاعات رعایت شده است. به‌طور کلی می‌توان گفت محیط رابط کاربری پایگاه‌های اطلاعاتی مورد بررسی در این پژوهش به‌طور میانگین ٤٥/٤٥ درصد از مؤلفه‌های مربوط به جستجو را رعایت کرده‌اند (جدول ۱۰).

جدول ۱. وضعیت رابط کاربری پایگاه‌های اطلاعاتی از لحاظ مقوله جستجو

مؤلفه‌های جستجو / پایگاه‌های اطلاعاتی	مرکز منطقه‌ای	پژوهشگاه علوم و فناوری اطلاعات	نورمگز	مگ‌ایران	نمامتن	سید	میانگین
جمع امتیازها	۱۰	۱۰	۱۱	۱۰	۸	۱۱	---
درصد	٤٥/٤٥	٤٥/٤٥	٥٠	٤٥/٤٥	٣٦/٣٦	٥٠	٤٥/٤٥

در پاسخ به پرسش دوّم پژوهش می‌توان این‌طور بیان نمود که از لحاظ مقوله نمایش، محیط‌های رابط کاربری پایگاه‌پژوهشگاه علوم و فناوری اطلاعات با رعایت ۱۸ مؤلفه (۸۱/۸۱ درصد) بیشترین میزان رعایت و پایگاه مگ‌ایران با رعایت ۱۰ مؤلفه (٤٥/٤٥درصد) کمترین میزان رعایت مؤلفه‌های مربوط به مقوله نمایش را داشته‌اند. همچنین از میان مؤلفه‌های مربوط به مقوله نمایش اطلاعات مؤلفه‌ی «نمایش آمار تعداد دانلود هر مقاله» در رابط کاربری هیچ‌کدام از پایگاه‌ها رعایت نشده است. به‌طور کلّی یافته‌ها حاکی از آن است به‌طور میانگین پایگاه‌ها از لحاظ مقوله نمایش ٦٤/٣٨ درصد معیارها را رعایت کرده‌اند. نتایج در جدول ۲ قابل مشاهده است.

جدول ۲. وضعیت رابط کاربری پایگاه‌های اطلاعاتی از لحاظ مقوله نمایش اطلاعات

مؤلفه های نمایش / پایگاه‌های اطلاعاتی	مرکز منطقه‌ای	پژوهشگاه علوم و فناوری اطلاعات	نورمگز	مگ‌ایران	نمامتن	سید	میانگین
جمع امتیازها	۱۵	۱۸	۱٦	۱۰	۱۲	۱٤	---
درصد	٦٨/١٨	٨١/٨١	٧٢/٧٢	٤٥/٤٥	٥٤/٥٤	٦٣/٦٣	٦٨/٣٨

پاسخ به پرسش سوّم پژوهش: همان‌طور که از جدول ۳ مشخص است پایگاه‌های مگ‌ایران و نورمگز با رعایت ۵ مؤلفه (٦٢/٥٠ درصد)، در میزان هماهنگی با مؤلفه‌های مقوله خدمات در رتبه نخست قرار دارند. پایگاه نمامتن و پژوهشگاه علوم و فناوری اطلاعات نیز با رعایت ۳ مؤلفه (۳۷/۵۰ درصد)، پایگاه مرکز منطقه‌ای با رعایت ۲ مؤلفه (۲۵ درصد)، پایگاه سید به ترتیب در رتبه‌های دوّم ، سوّم و

جدول 1. وضعیت رابط کاربری پایگاه های اطلاعاتی از لحاظ مقوله جستجو

در پاسـخ به پرسـش دوم پژوهش می توان این طور بیـان نمود که از لحاظ مقوله نمایش محیط های رابط کاربری پایگاه پژوهشـگاه علوم و فناوری اطلاعات با رعایت 18 مؤلفه (81/81 درصد) بیش ترین میزان رعایت و پایگاه مگ ایران با رعایت 10 مؤلفه (45/45) درصد کم ترین میزان رعـایت مؤلفه هـای مربوط به مقوله نمایش را داشـته انـد همچنین از میان مؤلفه های مربوط به مقوله نمایش اطلاعات مؤلفه ی «نمایش آمار تعـداد دانلود هر مقاله در رابط کاربری هیـچ کـدام از پایگاه ها رعایت نشـده است. به طور کلی یافته ها حاکی از آن است به طور میانگین پایگاه ها از لحاظ مقوله نمایش 64/38 درصد معیارها را رعایت کرده اند. نتایج در جدول 2 قابل مشاهده است

جدول 2. وضعیت رابط کاربری پایگاه های اطلاعاتی از لحاظ مقوله نمایش اطلاعات

پاسخ به پرسش سوم پژوهش همان طور که از جدول 3 مشخص است پایگاههای مگ ایران و نورمگز با رعایت 5 مؤلفه (62/50 درصد) در میزان هماهنگی با مؤلفه های مقوله خدمات در رتبه نخست قرار دارند پایگاه نما متن و پژوهشگاه علوم و فناوری اطلاعات نیز با رعایت 3 مؤلفه (37/50 درصد)، پایگاه مرکز منطقه ای با رعایت 2 مؤلفه (25 درصد) پایگاه سید به ترتیب در رتبه های دوّم، سوّم و

چهارم قرار گرفتند همچنین از میان مؤلفه های مقوله خدمات مؤلفه های «امکان ساخت پرونده شخصی (شخصی سازی محتوی)» و سهولت خروج از سیستم به طور مساوی با فراوانی 4 و درصد فراوانی 80 بیشترین میزان رعایت را داشته اند مؤلفه های «استخراج اطلاعات کتابشناختی برای نرم افزارهای مدیریت ارجاعات» و «ارائه خدمات آگهی رسانی جاری (آلرت نشریات، آلرت جستجو)» در محیط های رابط کاربری هیچ کدام از پایگاه های اطلاعاتی مورد بررسی رعایت نشده است. میانگین امتیاز پایگاه ها در این مقوله نیز 37/49 درصد به دست آمد (جدول 3)

عکس

چهارم قرار گرفتند. همچنین از میان مؤلفه‌های مقوله خدمات، مؤلفه‌های «امکان ساخت پرونده شخصی (شخصی‌سازی محتوی)» و «سهولت خروج از سیستم» به‌طور مساوی با فراوانی ٤ و درصد فراوانی ٨٠ بیشترین میزان رعایت را داشته‌اند. مؤلفه‌های «استخراج اطلاعات کتابشناختی برای نرم‌افزارهای مدیریت ارجاعات» و «ارائه خدمات آگهی‌رسانی جاری (آلرت نشریات، آلرت جستجو)» در محیط‌های رابط کاربری هیچ‌کدام از پایگاه‌های اطلاعاتی مورد بررسی رعایت نشده است. میانگین امتیاز پایگاه‌ها در این مقوله نیز ٣٧/٤٩ درصد به‌دست آمد (جدول ٣).

جدول ٣. وضعیت رابط کاربری پایگاه‌های اطلاعاتی از لحاظ مقوله خدمات

مؤلفه‌های خدمات / پایگاه‌های اطلاعاتی	مرکز منطقه‌ای	پژوهشگاه علوم و فناوری اطلاعات	نورمگز	مگ‌ایران	نمامتن	سید	میانگین
جمع امتیازها	٢	٣	٥	٥	٣	٠	---
درصد	٢٥	٣٧/٥٠	٦٢/٥٠	٦٢/٥٠	٣٧/٥٠	٠	٣٧/٤٩

در پاسخ به سؤال چهارم می‌توان گفت پایگاه‌های مرکز منطقه‌ای، مگ ایران، نمامتن و سید هر یک با رعایت ٣ مقوله (٦٠ درصد) مشترکاً در رتبه نخست و پایگاه‌های پژوهشگاه علوم و فناوری اطلاعات و نورمگز به‌طور مساوی با رعایت ٢ مقوله (٤٠ درصد) در رتبه دوّم قرار دارند. پایگاه‌ها از لحاظ مقوله پیوندها به طور میانگین ٥٣/٣٣ درصد از مؤلفه‌ها رادر محیط رابط کاربری خود رعایت کردهاند. نتایج در جدول ٤ آمده است.

جدول ٤. وضعیت رابط کاربری پایگاه‌های اطلاعاتی از لحاظ مقوله پیوندها

مؤلفه های پیوندها / پایگاه‌های اطلاعاتی	مرکز منطقه‌ای	پژوهشگاه علوم و فناوری اطلاعات	نورمگز	مگ‌ایران	نمامتن	سید	میانگین
جمع امتیازها	٣	٢	٢	٣	٣	٣	---
درصد	٦٠	٤٠	٤٠	٦٠	٦٠	٦٠	٣٧/٤٩

پاسخ سؤال پنجم: همان‌گونه که از جدول ٥ مشاهده می‌شود پایگاه مرکز منطقه‌ای با رعایت ٤ مقوله (٦٦/٦٦ درصد) در رتبه نخست، پایگاه‌های نورمگز، مگ‌ایران و سید با رعایت ٣ مقوله مشترکاً در رتبه دوم، پایگاه پژوهشگاه علوم و فناوری اطلاعات با رعایت ٢ مقوله (٣٣/٣٣ درصد) در رتبه سوم و پایگاه نمامتن تنها با رعایت ١ مقوله (١٦/٦٦ درصد) در رتبه چهارم قرار دارد. همچنین از لحاظ مقوله راهنمایی مؤلفه‌های «پیام‌های اخطاری قابل درک» و «نقشه سایت» به ترتیب با درصدهای فراوانی ١٠٠ و

جدول 3. وضعیت رابط کاربری پایگاه های اطلاعاتی از لحاظ مقوله خدمات

در پاسخ به سؤال چهارم می توان گفت پایگاه های مرکز منطقه ای مگ ،ایران، نمامتن و سـید هر یک با رعایت مقوله (60 درصد) مشترکاً در رتبه نخست و پایگاه های پژوهشـگاه علوم و فناوری اطلاعات و نورمگز به طور مساوی با رعایت 2 مقوله (40 درصد) در رتبه دوّم قرار دارند پایگاه ها از لحاظ مقوله پیوندها به طور میانگین 53/33 درصـد از مؤلفه ها را در محیط رابط کاربری خود رعایت کرده اند. نتایج در جدول 4 آمده است.

جدول 4. وضعیت رابط کاربری پایگاه های اطلاعاتی از لحاظ مقوله پیوندها

پاسـخ سؤال پنجم همان گونه که از جدول 5 مشاهـده می شود پایگـاه مرکز منطقه ای با رعایت 4 مقوله (6666) درصد در رتبه نخست پایگـاه های نورمگز مگیران و سـید با رعایت 3 مقوله مشترکـاً در رتبه دوم پایگـاه پژوهشـگاه علوم و فنـاوری اطلاعات با رعایت 2 مقوله (33/33 درصـد) در رتبه سوم و پایگـاه نمامتن تنها با رعایت 1 مقوله (1666) (درصـد در رتبه چهارم قرار دارد هم چنین از لحاظ مقوله راهنمایی مؤلفه های پیام های اخطاری قابل درک و نقشه سایت به ترتیب با درصدهای فراوانی 100 و

بیش ترین و کم ترین میزان رعایت را در رابط کاربری پایگاه های مورد بررسی داشته اند. به طور میانگین 44/44 درصد مؤلفه های مربوط به مقوله راهنمایی توسط پایگاه های مورد بررسی رعایت شده اند.

عکس

• بیشترین و کمترین میزان رعایت را در رابط کاربری پایگاه‌های مورد بررسی داشته‌اند. به‌طور میانگین ٤٤/٤٤ درصد مؤلفه‌های مربوط به مقوله راهنمایی توسط پایگاه‌های مورد بررسی رعایت شده‌اند.

جدول ۵. وضعیت رابط کاربری پایگاه‌های اطلاعاتی از لحاظ مقوله راهنمایی

مؤلفه‌های راهنمایی / پایگاه‌های اطلاعاتی	مرکز منطقه‌ای	پژوهشگاه علوم و فناوری اطلاعات	نورمگز	مگ‌ایران	نمامتن	سید	میانگین
جمع امتیازها	٤	۲	۳	۳	۱	۳	---
درصد	٦٦/٦٦	۳۳/۳۳	۵۰	۵۰	۱٦/٦٦	۵۰	٤٤/٤٤

پاسخ به پرسش ششم پژوهش: همان‌گونه که از جداول ۱ تا ۵ ملاحظه می‌گردد بیشترین امتیاز و درصد فراوانی میزان رعایت مقوله‌های مورد بررسی در محیط رابط کاربری پایگاه‌های اطلاعاتی مرکز منطقه‌ای، پژوهشگاه علوم و فناوری اطلاعات، نورمگز و سید به ترتیب با امتیازهای ۱۵، ۱۸، ۱٦، ۱٤ و درصدهای ٦۸/۱۸، ۸۱/۸۱، ۷۲/۷۲ و ٦۳/٦۳ مربوط به مقوله ی نمایش و می‌باشد. همچنین بیشترین امتیاز و درصد فراوانی میزان رعایت مقوله‌های مورد بررسی در محیط رابط کاربری پایگاه‌های اطلاعاتی مگ‌ایران و نمامتن به‌طور مساوی با ۳ امتیاز و ٦۰ درصد مربوط به مقوله پیوندها است.

پاسخ به پرسش هفتم پژوهش: همان‌طور که در جدول ٦ مشاهده می‌شود محیط رابط کاربری پایگاه اطلاعاتی پژوهشگاه علوم و فناوری اطلاعات در مجموع با میزان رعایت ۳۵ مؤلفه (۵۵/۵۵ درصد) بیشترین میزان رعایت مقوله‌های موجود در سیاهه‌وارسی را داشته است و رتبه نخست را به خود اختصاص داده است. سپس محیط‌های رابط کاربری پایگاه‌های نورمگز، مرکز منطقه‌ای، سید و مگایران به‌طور مساوی و نمامتن در رتبه‌های دوم تا چهارم قرار داشتند.

جدول ٦. وضعیت کلی رابط کاربری پایگاه‌های اطلاعاتی

مؤلفه کلی / پایگاه‌های اطلاعاتی	جستجو	نمایش	خدمات	پیوندها	راهنمایی	جمع امتیازها	درصد
مرکز منطقه‌ای	۸	۱۵	۲	۳	٤	۳۲	۵۰/۷۹
پژوهشگاه علوم و فناوری اطلاعات	۱۰	۱٦	۲	۲	۳	۳۵	۵۵/۵۵
نورمگز	۱۱	۱٦	۵	۲	۳	۳٤	۵۳/۹٦
مگ‌ایران	۱۰	۱۰	۵	۳	۳	۳۱	٤۹/۲۰
نمامتن	۸	۱۲	۳	۳	۱	۲۷	٤۲/۸۵
سید	۱۱	۱٤	۰	۳	۳	۳۱	٤۹/۲۰

جدول 5. وضعیت رابط کاربری پایگاه های اطلاعاتی از لحاظ مقوله راهنمایی

پاسخ به پرسش ششم پژوهش: همان گونه که از جداول 1 تا 5 ملاحظه می گردد بیش ترین امتیاز و درصد فراوانی میزان رعایت مقوله های مورد بررسی در محیط رابط کاربری پایگاه های اطلاعاتی مرکز منطقه ای، پژوهشگاه علوم و فناوری اطلاعات نورمگز و سید به ترتیب با امتیازهای 15، 18، 16، 14 درصدهای 18/18، 81/81، 72/72 و 63/63 مربوط به مقوله ی نمایش و می باشد. همچنین بیشترین امتیاز و درصد فراوانی میزان رعایت مقوله های مورد بررسی در محیط رابط کاربری پایگاه های اطلاعاتی مگ ایران و نما متن به طور مساوی با 3 امتیاز و 60 درصد مربوط به مقوله پیوندها است.

پاسخ به پرسش هفتم پژوهش: همان طور که در جدول 6 مشاهده می شود محیط رابط کاربری پایگاه اطلاعاتی پژوهشگاه علوم و فناوری اطلاعات در مجموع با میزان رعایت 35 مؤلفه (55/55 درصد) بیش ترین میزان رعایت مقوله های موجود در سیاهه وارسی را داشته است و رتبه نخست را به خود اختصاص داده است. سپس محیط های رابط کاربری پایگاههای نورمگز مرکز منطقه ای، سید و مگایران به طور مساوی و نمامتن در رتبه های دوم تا چهارم قرار داشتند.

جدول 6. وضعیت کلی رابط کاربری پایگاه های اطلاعاتی

بحث و نتیجه گیری

این پژوهش با هدف سنجش رابط کاربر پایگاه های اطلاعاتی پیوسته مجلات تمام متن فارسی انجام شد. نتایج نشان می دهد که از لحاظ مقوله ی جستجو پایگاه ها در سطح مشابه ای می باشند و در حد متوسط (میانگین 45/45 درصد) معیارها را رعایت کرده اند. از لحاظ مقوله ی نمایش نیز پایگاه ها در وضعیت خوبی میانگین 64/38) قرار دارند از این لحاظ رابط کاربری پایگاه پژوهشگاه علوم و فناوری اطلاعات با میزان رعایت 81/81 درصد و پایگاه مگ ایران با میزان رعایت 45/45 درصد به ترتیب بیش ترین و کم ترین میزان هماهنگی را با مقوله های موجود در سیاهه وارسی داشته اند. رابط کاربری پایگاه ها در میزان رعایت مقوله خدمات کم ترین هماهنگی را نسبت به سایر مقوله ها داشته اند؛ به طوری که میانگین رعایت معیارهای این مقوله 37/49 درصد و در حد ضعیف محاسبه شده است. در ارزیابی میزان رعایت مؤلفه های مقوله پیوندها نیز رابط کاربری پایگاه ها در سطح متوسطی قرار داشتند و به طور میانگین 53/33 درصد معیارها را رعایت نموده اند از لحاظ میزان رعایت مؤلفه های مقوله راهنمایی به جز پایگاه نما متن که تنها 16/66 درصد معیارهای این مقوله را رعایت نموده است؛ رابط کاربری سایر پایگاه های مورد بررسی در وضعیت مشابه و متوسطی (میانگین 44/44 درصد) قرار داشتند.

همچنین نتایج دیگری که از این پژوهش حاصل شد این است که پایگاه اطلاعاتی «پژوهشگاه علوم و فناوری اطلاعات» با بیشترین میزان رعایت ،مؤلفه ها برترین محیط رابط کاربری را از بین محیط رابط کاربری پایگاه های اطلاعاتی پیوسته مجلات تمام متن فارسی دارا می باشد در نهایت می توان اظهار نمود که محیط رابط کاربری پایگاههای مورد بررسی از لحاظ میزان رعایت معیارهای موجود در سیاهه وارسی در حد «متوسطی» می باشند و این نتیجه یافته های پژوهش های انتظاریان و فتاحی (1389)، اعظمی و فتاحی (1388) و اسداللهی و نوکاریزی (1389) را تأیید می نماید به طور کلی بهتر است هر شش پایگاه نسبت به گنجاندن آن دسته از معیارهای مطرح شده در سیاهه وارسی که فاقد آن هستند، اقدام کنند.

در نهایت پژوهشگران بر اساس نتایجی که از ارزیابی رابط کاربری پایگاه ها به دست آمد به طرّاحان محیط رابط کاربری پایگاه های اطلاعاتی مجلات تمام متن فارسی پیشنهاد می دهند که جهت استفاده بهتر، افزایش موفقیت در جستجوها و انجام بهتر و سریعتر بازیابی اطلاعات از طریق محیط رابط کاربری و در نتیجه جذب کاربران بیشتر برخی نکات را که در محیط رابط کاربری آن ها به ندرت استفاده شده یا اصلاً استفاده نشده است رعایت کنند. این نکات عبارتند از: در مقوله جستجو به طراحی آیتم های «محدودگرهای مناسب»، «عملکرد اصلاح نامه»، «عملکرد مقالات مرتبط»، «عملکرد پیشنهاد کلیدواژه های مرتبط»، «ذخیره جستجوها در حساب کاربری» و «تاریخچه جستجو»؛ در زمینه نمایش اطلاعات به طراحی آیتم های «نشانه دار کردن نتایج» و «ارسال نتایج به ایمیل»؛ از لحاظ مقوله خدمات پایگاه ها «امکان استخراج اطلاعات کتابشناختی به نرم افزارهای مدیریت ارجاعات» «نشان دادن مقالات استناد کننده» و «خدمات آگاهی رسانی جاری»؛ در زمینه پیوندها نیز «امکان فراپیوند به عناوین مشابه» و «فراپیوند به دیگر مقالات یک نویسنده» بپردازند. بنابراین پیشنهاد می شود که طراحان پایگاه ها به طراحی مجدد پایگاه های خود پرداخته و به ویژه خدمات جدید معرفی شده در این پژوهش را در محیط رابط کاربری پایگاه های اطلاعاتی خود اعمال نمایند.

پیشنهاد هایی برای پژوهش های آینده:

پیشنهاد می شود پایگاه های مذکور را از نظر جنبه های خاص شامل «دامنه پوشش موضوعی و زمانی مجلات»، «نحوه سازمان دهی اطلاعات» و غیره مورد پژوهش قرار دهند همچنین مطالعاتی نیز بر پایگاه های تهیه شده در کشور جز پایگاه های مجلات از قبیل پایگاه های اطلاعات کتابشناختی پایگاه های مقالات کنفرانس ها انجام شود.

منابع

اسد اللهی، زهرا، محسن نو کاریزی. 1389. ارزیابی ساختار و محتوای پایگاه های اطلاعاتی الکترونیکی نشریات ایرانی کتابداری و اطلاع رسانی. 13(2)

http://www.aqlibrary.ir/index.php?module=TWArticlesfile=indexfunc=view- (دسترسی در pubarticlesdid=882pid=10.(91/5/3

اعظمی، محمّد رحمت الله فتاحی 1388. تطابق رابط گرافیکی کاربر پایگاه های اطلاعاتی با مدل رفتار اطلاع یابي الیس علوم و فناوری اطلاعات. 25 (2): 247-264.

انتظاریان، ناهید، رحمت الله فتاحی. 1389. مبانی طراحی رابط کاربر مبتنی بر شناخت ویژگی ها، ادراک و رفتار کاربران. کتابداری و اطلاع رسانی. 13 (2).

http://www.aqlibrary.ir/index.php?module=TWArticlesfile=indexfunc=view (دسترسی در pubarticlesdid=878pid=10.(91/4/24

علیجانی ،رحیم، لیلا دهقانی 1386 مقایسه ی رابط کاربر پایگاه های اطلاعاتی کتاب مدار بین المللی فصلنامه کتاب 72: 233-252

مهراد، جعفر، لیلاـ دهقانی 1385 معیارهای ارزیابی رابط های کاربر نسخه های مختلف پایگاه های اطلاعاتی مجله کتابداری بهار و تابستان 77-95 .

مهراد جعفر زهره زاهدی. 1386 بررسی و مقایسه رابط کاربر دو میزبان داخلی کتابخانه منطقه ای علوم و تکنولوژی و پژوهشگاه اطلاعات و مدارک علمی ایران با چهار میزبان خارجی Proquest Elsevier ،Ebsco، Emerald کتابداری و اطلاع رسانی 10 (3): 107-124.

یمین فیروز ، موسی 1382 ویژگی ها و عناصر تشکیل دهنده رابط کاربر در وب سایت ها فصلنامه کتاب .14:159-168

Cracstini, F. 2004. A Graphical user interface for the retrieval of hierarchically structured documents.Information Processing Management. 40 (2): 269-289.

Hansen, P. 1998. Evaluation of IR user interface implications for user interface design. www.) Hb.Se/bhs/ith/2-98/ph.htm (Available at 12/6/2012

Hirst, S.J. 1999. HyperLib Deliverable 2.1.1: The Use of Icons in a Multilingual OPAC Interface. Hyperlib .Electronic Document Store, University of Antwerp–University of Loughborough

http://lib.ua.ac.be/MAN/WP211/root.htm(Available at 8/5/2012)

Sabin–Kidiss, L. 2001.Assessing the functionality of web–based versions of traditional search engines ..Online. 25 (2): 18–24

چکیده

طبق قانون حق مؤلف کشور فرانسه مصوب اول اوت 2006، کتابخانه ملی فرانسه (BNF از این پس کتابخانه) وظیفه گردآوری و حفاظت اینترنت فرانسه را بر عهـده دارد. این کتابخانه «مـدل تلفیقی» را برای آرشـیو وب ایجاد کرده است. این مدل، متشـکل از خزش های فراگیر دامنه .fr و خزش های کانونی و واسپاری های الکترونیکی است.

کتابخانه ملی فرانسه، به برکت همکاری پژوهشی با آرشیو اینترنت از سال 2004 هر سال به اجرای چهار خزش فراگیر مبادرت کرده است. آخرین آن هـا، بـا ویژگی های متفاوت چشـمگیری ایجاد شـده است و از مهم ترین این ویژگی ها کاربرد فهرست همه جانبه ای از اسامی دامنه های fr ،بود که توسط آفنیک انجمن همکاری نام گذاری اینترنت ،فرانسه برای ثبت fr) بعد از امضای توافق نامه ای میان دو سازمان در سپتامبر 2007 به کتابخانه ملی فرانسه ارائه شد.

گزینش های ماهرانه قبل و حین ،خزش نقش تعیین کنندهای در شـکل آینده مجموعه خواهند داشت ،بنابراین تصمیم هایی که باید بر طبق ساختار قانونی و معنوی در مـدت انجام خزش اتخاذ شونـد عبارت اند از: برای بیان اف این مجموعه شامل 5 قرن سنت گذشته و اسپاری قانونی است. برای تعیین پیامـدها و نتایج راه حل های فنی موجود، قصد داریم نتایج آخرین خزش این کتابخانه را تحلیل و با برداشت های سال های گذشته مقایسه کنیم به علاوه این مطالعات سودمندی تلاش های ما را برای توصیف وب 2007 فرانسه تأیید می کند.

کلیـد واژه هـا: آرشـیو سـازی ،وب قانون و اسـپاری اینترنت کتابخانه ملی فرانسه بیان اف، کنسرسـیوم بین المللی حفاظت ،اینترنت IIPC آرشیو اینترنت میراث دیجیتال

قانون واسپاری وب فرانسه: راهبردهایی برای گردآوری دامنه ملی

اشاره

راهبردهایی برای گردآوری دامنه ملی (1)

نوشته: فرانس لاس فارگوس کلمنت کیوری، برت وندلاند (2)

ترجمه: سودابه نوذری (3)

1. قلمرو فرانسه

1,1تعریف حوزه قانون و اسپاری

در اول اوت 2006، قانون حق مؤلف جدید در مجلس فرانسه به رأی گذاشته شد. مفاد این قانون، که کتابخانه مدت ها انتظارش را می کشید، قانون واسپاری را به اینترنت کشاند. قانون واسپاری، به هر ناشر تکلیف می کند نسخه هایی از تولیداتش را به کتابخانه ملی ارسال کند این ،قانون که نخستین بار در 1537 برای منابع چاپی تصویب شده بود با گذشت قرن ها به اشکال متفاوت تولیدات فرهنگی انتشاراتی جدید از منقوشات گرفته تا نرم افزارها و بازی های ویدئویی تعلق گرفت به موازات گسترش شبکه جهانی وب، به عنوان مکانی مطلوب برای ایجاد دانش و اطلاعات لازم بود برای نهادهای حافظ میراث ملی فرانسه، چارچوبی قانونی برای سازماندهی منابع مورد حفاظت شان ارائه شود.

قانون فوق درباره قلمرو اینترنت فرانسه شفاف نیست و برای روشن شدن این مسئله در آینده ای

1- Legal deposit of the French Web: harvesting strategies for a national domain

2- France Lasfargues, Clé ment Oury, and Bert Wendland

3- عضو هیئت علمی سازمان اسناد و کتابخانه ملی ایران

نزدیک صدور فرمانی انتظار می رود در این اثنا ،کتابخانه خط مشی خود را درباره دامنه ملی برخط به پیش می برد که احتمالاً با ،فرمان زمانی که تأیید شود سازگار باشد ،کتابخانه به منظور پالایش حوزه سیاست آرشیوسازی وب خود پنج قرن و اسپاری قانونی و تجربه جدیدتر چالش های فنی گردآوری وب را با یکدیگر تلفیق کرده است بنابراین روش ما هم عملی است (یعنی موافق با ابزار اکتشافی برداشت حجمی) و هم سازگار (با قانون واسپاری گذشته فرانسه).

این سنت بر اساس سه معیار قرار داشته است:

– انتشارات: اسناد گردآوری شده برای یک مخاطب و در خدمت یک هدف عمومی هستند، آن ها نباید در حد حوزه های خصوصی یا ارتباطات درون شرکتی تنزل کنند؛

– رسانه: همه ترتیبات قانون واسپاری قبلی بر اساس موجودیت فیزیکی یک رسانه قرار داشتند: منابع ،چاپی متون موسیقی (1) تصاویر نوارهای صدا دیسکت و جز آن؛ و

– قلمرو: اسناد باید در محدوده مرزهای قلمرو ملی منتشر یا توزیع می شدند.

به طور خلاصه، قانون واسپاری گذشته برای تمام انتشارات موجود در رسانه های تولید شده یا توزیع شده در قلمرو کشور فرانسه قابل اجرا بوده است.

متأسفانه هیچ یک از این معیارها برای گردآوری [منابع] وبی به راحتی قابل اجرا نیست، زیرا وب گاه ها:

– به ترکیب و ادغام ارتباطات شخصی و عمومی و ترکیب مبتکرانه آن ها تمایل دارند؛

– به معنای دقیقه کلمه، رسانه نیستند، بلکه بیش تر سکویی هستند که سایر رسانه ها تمایل به انتقال به آن ها دارند (می توان کتاب، تصاویر، فیلم، متون موسیقی، و جز آن را در وب یافت)؛ و

– به سهولت، به یک سرزمین خاص – دست کم در مقیاسی وسیع – محدود نمی شوند؛

به علاوه، دیگر نمی توان زبان فرانسه را معیار انتخاب قرار داد زیرا قانون و اسپاری بدون توجه به زبان انتشاراتی اعمال می شود: مجموعه های واسپاری شده به کتابخانه ملی فرانسه شامل اقلامی به زبان های خارجی می شود که در کشور فرانسه منتشر چاپ یا توزیع شده باشند به همین ترتیب، نمی توانیم از راهبرد قانون و اسپاری خود انتظار تمرکز بر موضوع نویسنده یا سطوح انتشاراتی خاص داشته باشیم جنبه مهم این قانون آن است که مجموعه ها باید منعکس کننده جامعه و فرهنگ فرانسه به هر شکل و صرف نظر از ارزش علمی یا محبوبیت آن ها باشند.

اگر مجبور به انتخاب هستیم منظور بیش تر تهیه نمونه هایی از منابع است تا انتخاب نسل های بعدی تصمیم می گیرند که چه چیزی ارزشمند است نه کتابخانه ملی در قفسه های کتابخانه، نویسندگان ناشناخته و مجله های مستهجن در کنار متفکران بزرگ قرار گرفته اند، انتظار داریم این فلسفه در مورد منابع و بی هم اعمال شود برداشت ،فل های فرصت بزرگی را برای عرضه این رویکرد در مقیاس وب فراهم می.کند قانون و اسپاری درباره محتوا و قالب – یا محمل – است به این مفهوم که کتابخانه ملی فرانسه به ساخت مجموعه هایی که

منعکس کننده گرایش هایی از نمونه های انتشاراتی است توجه می کند تلاش می کنیم طیف گسترده اشـیا و قالب هایی را برای ارائه فراهم کنیم که به طور عملی توسط افراد

ص: 256

---

scores -1

بهره مند از اطلاعات تولید می شود.

در نتیجه، برای دامنه ملی خود نیازمند تعریفی بودیم که با وجود انعطاف پذیری و سادگی اجرا منعکس کننده «روح» این گذشته باشد تعریف یک «کانون» فرانسه در عین داشتن قابلیت انعطاف پذیری، تنها راه تسهیل شیوه های گردآوری اکتشافی در مقیاس کلان بود ثابت شده است ،زبان، مکان جغرافیایی اسامی یا موضوع برای تعیین تفاوت در مقیاس کلان بسیار بی ربط و چالش برانگیز هستند، بنابراین، آن چه به عنوان نقطه آغاز احتمالی برای اکتشاف دامنه ملی باقی ماند استفاده از دامنه سطح بالای ملی (تی. ال دی) Top Level domain (TLD)) (1) بود این دامنه به سبب تهیه فهرست آغازین اصلی برای اکتشاف با سایر راهبردها تلفیق شده است. از این رو خط مشی واسپاری برخط ما حین انتظار آشکار شدن در حکم صادر شده، به شکل زیر تعریف شد. ما «فرانسه» را چنین در نظر گرفتیم:

– به عنوان یک اصل هر وبگاهی که در دامنه سطح بالای .fr یا هر دامنه سطح بالای مشابه دیگر، با ارجاع به قلمرو رسمی فرانسه ثبت می شود (برای ،نمونه re برای جزیره فرانسوی لاری نیو (2))؛

– هر وب گاهی (شاید خارج از فرانسه) که تولید کننده اش در سرزمین فرانسه قرار دارد (معمولاً این مسئله را می توان در وب گاهی یا با استفاده از Domain Name System (3) تعیین کرد)؛

– هر وبگاهی (شاید خارج از فرانسه) که بتواند برای نمایش محتواهای تولید شده در سرزمین فرانسه تأیید شده باشد (این معیار آخر چالش برانگیزتر از آن است که بررسی شود اما مجالی را برای تفسیر و گفت و گو میان کتابخانه ملی فرانسه و تولیدکنندگان وب به وجود می آورد).

البته انتظار نمی رود هیچ یک از این معیارها قبل از این که ما فهرست هسته (4) را بسازیم، به طور کامل تأمین شوند (این امر مانع اکتشاف می شود و در بررسی وب گاه ها، قبل از خزش (5)، ایجاد اخلال می کند، که به سادگی قابل اندازه گیری نیست) .با وجود ،این آن ها قصد شان خدمت در چارچوبی قانونی و فرهنگی است به منظور:

– تعریف و توصیف شیوه کلی سیاست آرشیو سازی وب ملی برای مردم صاحبان محتوا، و کتابداران درگیر در پروژه؛

– تبیین وظایف و دستورالعمل های مورد نیاز برای خود تا وقت نظارت بر خزش ها به یاد داشته باشیم؛ تعیین عناصر عینی تصمیم گیری وقتی ما نیازمندیم بدانیم آیا وب گاهی به طور کامل در حوزه کار ماست یا خیر این امر به ویژه زمانی که وب مستر یک وب گاه خاص از کتابخانه ملی می خواهد کار خزش را متوقف کند یا حتی هنگام رویارویی با یک دعوی ،قانونی سودمند است: اگر وب گاه در هیچ یک از معیارهای فوق ،نگنجد باید از خزش های آینده و از مجموعه کنار گذاشته شود.

بنابراین، رویکرد دامنه ملی مصالح های را میان قانون واسپاری گذشته و ویژگی های چالش برانگیز

1- fr

2- La Ré union

3- DNS

4- seed list (فهرست یو. آر.ال).

5- crawl

وب نشان می دهـد ،همچنین مصالح های میان رویکرد کاملاً باز و نامعقولانه ای که به احتمال می توانست ما را طوری هدایت کند تا کل شبکه جهانی وب را به عنوان فرانسه بالقوه در نظر بگیریم، یا بر عکس، رویکردی محـدود کننـده، که فقط به دامنه سـطح بالای fr تنزل می یافت در حالی که معروف است این دامنه تنها بخش محدودی از وب گاه های فرانسوی را شامل می شود در یک کلام، هدف این رویکرد، تأمین تمرکز و انعطاف پذیری است.

2.1.در حال حاضر کجا هستیم

کتابخـانه اجـازه دارد راه هـای متفـاوتی را برای گردآوری اینترنت فرانسه به کـار ببرد: «مؤسـسه های قیم ممکن است منابع اینترنت را با به کارگیری فنون خودکار یا با تنظیم موافقت نامه های خاص و فرآینـدهای واسـپاری با همکاری تولیدکنندکان گردآوری کنند» (ماده 41 II) به همین منظور کتابخانه ملی فرانسه مدلی تلفیقی متشکل از سه راهبردی زیر را تعریف کرده است.

– برداشت فله ای اینـترنت فرانسه هـدف گردآوری دامنه fr ، دست کم به صورت سالانـه است. این خزش هـای فراگیر (1) امکان آرشیو تصـاویر وب فرانسه را به کتابخـانه می دهنـد. این رویکرد، در مقـایسه با هزینه های ماشـینی و نیروی انسانی برداشت [و حجم] اطلاعات بازیابی ،شـده باصـرفه تر است. با وجود ،این به خاطر محدودیت منابع و ابزاری با چنین خزش گر (2) هایی امکان ندارد بتوان وب عمیق (وبگاه های بسیار بزرگ پایگاه های اطلاعاتی و جز آن) را گردآوری کرد.

– خزش های کانونی (3) برای وب گاه ها به تعداد محدود این سایت ها در داخل یا خارج از کتابخانه ملی ،فرانسه با همکاری شبکه ای از کتابـداران و پژوهش گران کشف شـده انـد خزش گر هـای کانونی به وبگاه های بزرگ و به وب گاه های غالباً در حال اصـلاح، اختصاص دارند.

– واسپاری های الکترونیکی ویژه شمار محدودی انتشارات الکترونیکی

کتابخانه، منتظر تصویب قانونی که فنون خزش را بررسـی کند نشد پروژه آرشـیو سازی وب در 1999 آغاز شد. نخستین خزش گر کانونی رویداد محور حدود سال 2002 آزمایش شد در آن زمان کتابخانه نزدیک به دو میلیون وب گاه وابسته به انتخابات کشور (انتخابات ریاست جمهوری و مجلس) را گردآوری کرد این کار برای اروپا و برای انتخابات محلی دو سال تمدیـد شـد کتابخانه ملی فرانسه 1162 وب گاه را گردآوری کرد (4)

با وجود این، تجهیزات فنی (سـخت افزاری و نرم افزاری) مهارت ها و تجربه لازم برای تحقق خزش گر های بزرگ مقیاس اینترنت ،فرانسه در کتابخـانه ملی هنوز کـافی نبود این دلیل چرایی موافقت نامه همکاری کتابخانه ملی فرانسه با آرشـیو اینترنت (IA)، بنیاد غیرانتفاعی در آرشیوسازی شبکه جهانی وب، از 1996 است. در نوامبر 2004 این دو مؤسسه موافقت نامه تحقیقاتی را با نام «پروژه پژوهشی: انتخاب

---

broad crawls –1

crawler –2

Focused crawls –3

4- خزش گرها برای انتخابات 2002 و 2004 استفاده شدند برای اطلاعات بیش تر درباره این دو خزش گر به منبع 19 مراجعه کنید.

دامنه ملی برای آرشیوسازی وب» امضا کردند هدف این پروژه تعیین شیوه ها و ابزارهایی برای استفاده در یک خزش دامنه ملی وب بود موافقت نامه تصریح کرد برای خزش های فراگیر، لازم است بررسی این ابزارها و شیوه ها توسط آرشیو اینترنت (IA) انجام و داده هایی که طی این خزش ها گردآوری می شد به قفسه های ذخیره سازی بی.ان.اف تحویل شود.

نخستین خزش فراگیر در چارچوب این موافقت ،نامه در پایان سال 2004 اتفاق افتاد. از آن، در سال های 2005 2006 و 2007 سه خزش فراگیر fr انجام شد (این آخرین خزش که تا سال 2009 ادامه خواهد ،داشت مرهون بسط موافقت نامه تحقیقاتی است).

خزش های با مقیاس کوچک تر نیز - مستقیم یا غیر مستقیم - توسط کتابخانه ملی فرانسه اجرا شده .است دو خزش کانونی در شمار محدودی وب گاه (در حدود 4000 عدد) توسط IA برای پرژه تحقیقاتی انجام شد [15] از 2007، این وب گاه ها توسط بی.ان.اف با استفاده از امکانات خود، برداشت می شوند. در همان سال سایر خزش های کانونی موضوعی یا رویداد محور به اجرا درآمدند، مانند وب گاه های مربوط به انتخابات ملی 2007 فرانسه.

به این ترتیب بی. ان. اف تاکنون - مستقیم یا به لطف همکاریش با IA-چهار خزش فراگیر به علاوه حجم فراونی خزش های کانونی به اجرا در آورده است ،دیگر آرشیوسازی ،وب یک پروژه در کتابخانه ملی فرانسه نیست بلکه فعالیتی روزمره و واحدی دائمی در اداره قانون واسپاری (1) این کتابخانه است. از آن جا که این امر بهترین رویکرد در مواجه با چالش گردآوری حجم روزافزون اشیای دیجیتال در وب است برداشت فله ای هنوز اولویت نخست به شمار می آید.

با وجود این برداشت فله ای به معنای برداشت کورکورانه نیست حتی وقتی روبات ها سعی دارند حداکثر فایل های وبی را کشف کنند خود را به رعایت قوانین و تنظیمات ملزم می کنند [20]. تصمیمات فنی ،قبل حین و بعد از خزش نقش قطعی در نتیجه برداشت دارد.

این مقاله راهبردهایی که بی.ان.اف با همکاری آرشیو اینترنت (IA) برای اجرای خزش های بزرگ مقیاس ارائه کرده است توصیف میکند این خزش ها می تواند با دیدگاه دامنه ملی فرانسه سازگار باشد.

2. طراحی خزش

1.2. هدف چیست؟

به نظر می رسد، اجرای خزش فراگیر در حدود صدها میلیون فایل با طیفی از مشکلات فنی اجتناب ناپذیر همراه باشد. با این حال، نخستین سؤالی که قبل از آغاز یک خزش باید پاسخ داده شود، سؤالی فنی نیست: هدف این خزش گر چیست؟ پاسخ ها بسته به اینکه از یک شرکت بزرگ نهادی پژوهشی، یا مؤسسه ای میراثی باشد متفاوت خواهد بود. اگر لازم است داده ها در مدت طولانی نمایه سازی (توسط موتور جست و جو)، تحلیل (به طور مثال برای شناسایی دامنه های وبی) و یا آرشیو و نمایش داده شوند، خزش فراگیر شیوه یکسانی را اجرا نمی کند.

1- Legal Deposit Department

به عبارت دیگر توجه به محدودیت های فنی، هنگام تعریف اهداف خزش فراگیر ضروری است. این نکته دلیل لزوم گفت و گوی دائمی میان کتابداران (که وظیفه تعریف خط مشی مجموعه سازی را بر عهده دارند) و مهندسان (که وظیفه اجرای خزش ها را بر عهده دارند) را به خوبی بیان می کند متدولوژی ای که در این بخش توصیف می شود نقطه تلاقی دغدغه های مربوط به این دو گروه است

خزش های کتابخانه در چارچوب قانون و اسپاری به اجرا درآمده اند سودمندی این چارچوب تنها برای پرداختن به مسائل حفاظت مالکیت معنوی نیست بلکه آرشیوسازی وب را به عنوان وظیفه ای مستمر مطرح می کند. مطابقت خط مشی مجموعه سازی در آرشیوهای وب، با قالب های قدیمی ،انتشارات امری ضروری است.

کشف خودکار وب گاه ها با روبات راه کاری است برای سازگاری با ویژگی «غیر تبعیض آمیزانه» قانون واسپاری ،فرانسه تا هم «بهترین» (انتشارات ادبی علمی) و هم «بدترین» (از آگهی ها گرفته تا پورنوگرافی) انتشارات فرانسوی را گردآوری کند با وجود ،این حتی روبات ها در معرض پیش داوری هستند. ساختار فراپیوندی ،وب نخست به کشف و گردآوری پر استناد ترین وب گاه منتهی شد اما با روبات های آرشیوسازی ما وب گاه های کم طرفدار فراموش نمی شوند برای اجتناب از این پیش داوری BNF در سپتامبر 2007، موافقت نامه ای را با انجمن همکاری نامگذاری اینترنت فرانسه (آفنیک) (1)، سازمان مسئول دامنه های fr و re. امضا کرد. طبق شرایط ،موافقت نامه این سازمان باید هر دو سال یک بار فهرست کاملی از اسامی دامنه ثبت شده موجود در دامنه های .fr و re را ارائه دهد (هم اکنون بیش از یک میلیون اسامی دامنه دارد) به عبارت دیگر BNF باید اعتبار این اطلاعات ارزشمند را تضمین کند.

بنابراین، هدف یک خزش دامنه ،بزرگ گردآوری نمونه ای جامع از دامنه ملی و ارائه تصویری از تولیدات [فکری] فرانسه در زمان این گردآوری است در بیش تر موارد نمونه به عنوان یک تصویر در نظر گرفته می شود- راه کاری برای حفظ و تثبیت یک فضای در حال تحول از آن جا که گردآوری همه چیز ممکن نیست به بهای [از دست دادن] یکی برداشت چند سند از هر وب گاه را به گردآوری کل چند وب گاه ترجیح می دهیم

به این سبب خزش های عمیق تری را در مهم ترین وب گاه ها تجربه نکردیم مانند کتابخانه ملی استرالیا که با گذاشتن اولویت بالا در مورد وب گاه های دولتی و دانشگاهی این امر را انجام داد فهرستی از این وب گاه ها توسط کتابداران منتشر شده است [17]) این کار لازم هم نبود زیرا خزش های فراگیر ما با همراهی خزش های کانونی به صورت کامل به اجرا در می آیند.

از سوی دیگر، احتمال تأثیر سریع وب از تحولات فناورانه وجود دارد. قالب های انتشاراتی نوین پدیدار می شوند و در عرض چند ماه گسترش می یابند خزش فراگیر باید بازتاب دهنده این تحولات باشد به عنوان نهاد قانون واسپاری یکی از اهداف ،ما روشن کردن قالب های نوین انتشاراتی است و در نتیجه کسب اطمینان از اینکه روبات توانایی برداشت این اسناد را داشته باشد به این سبب، در سال 2006 ، به وب نوشت ها و وب گاه های شخصی توجه بیش تری داشتیم و در سال 2007 بر ویدیو ها تأکید

1- Association française pour le nommage Internet en cooperation

داریم (در ادامه می آید).

طبق اهـداف از پیش تعریف شـده در آینـده احتمـال دارد مجموعه قبـل و در زمان خزش شـکل گیرد. قبل از آغاز کار ،برداشت درباره دو عامل مهم که سهم بزرگی در ساخت مجموعه های آینده دارند تصمیم گیری شد طراحی فهرست هسته، و تنظیمات خزش.

2.2فهرست هسته

خزش گر وظایفش را با فهرستی از یو. آر. ال آغاز می کند که فهرست هسته نامیده می شود. هسته ها، در هایی برای دستیابی به وب گاه ها به شمار می آیند؛ به این سبب کیفیت برداشت تا حد زیادی به کیفیت این فهرست بستگی دارد.

از آن جا که همکار ما در اجرای خزش فراگیر برای چهار سال همان [آرشـیو اینترنت (IA)] باقی ماند، امکان غنی کردن تدریجی فهرست هسته وجود داشت منابع مختلف هسته ها سال به سال افزایش می یافتند:

– 2004: فهرست هسته [یو.آر.ال]، برای نخستین خزش فراگیر از استخراج دامنه های fr. آخرین خزش الکسا به وجود آمد. (1)

– 2005: هسـته های حاصل از استخراج دامنه های fr آخرین خزش گر و میزبـان آلکسا، هنگام خزش فراگیر قبلی کتابخانه ملی فرانسـه کشف شدند هدف فرصت دادن به خزش گر برای پیش رفتن و کشف میزبان های جدید بود

– 2006: فهرست هسته به شیوه سال گذشته آغاز شد.

– 2007: به لطف امضـای توافق نـامه با آفنیک فهرست جامع اسامی دامنه fr و re. به عنوان فهرست هسـته به کـار رفت برای اطمینان از تداوم و سازگاری با خزش های فراگیر ،قبلی این فهرست با استخراج هایی از خزش آلکسا و از خزش های میزبان قبلی ادغام شد.

آفنیک شامل:

– 890064 دامنه fr

– 1516 دامنه re، و

– 753/ 21 اسم دامنه سطح دوم دامنه سطح دوم (2،3،3) را ببیند.

با وجود این اسـتفاده از فهرست ،آفنیک به عنوان فهرست هسـته ،ساده امکان پـذیر نبود این فهرست شامل اسامی دامنه می شد نه یو آر ال آن ها به عبارتی پشت نام هر دامنه یک وب گاه وجود نداشت

بنابراین در برخورد با فهرست ،آفنیک چند تحلیل مورد پردازش قرار گرفتند.

1– اینترنت آلکسـا شـرکتی انتفاعی است که در وب می خزد تا داده هایی را برای یک مرورگر toolbar (plug–in) به همراه پیشنهاد داده

هایی درباره سایت های مورد بررسی فراهم کند و بر اساس داده های گردآمده از دیگر کاربران ، صفحه های مرتبطی را که ممکن است مورد علاقه آن ها باشند، توصیه می کند این شرکت از ،1996 آرشیوهای خزش خود را در اختیار آرشیوهای اینترنتی قرار داده است [16]

نخستین آن ها کمّی کردن تعداد دامنه هایی بود که هنوز فعال بودند و باید به طور عملی، به عنوان هسته، مورد استفاده قرار می گرفتند. این بررسی سنگین توسط آرشیو اینترنت (IA) انجام شده است. برای کسب اطمینان از برخط بودن پاسخ آن ها هر دامنه با دو نشانی مختلف بررسی می شد (یعنی بررسی 1/780/128 یو.آر.ال):

http://www.domainname.fr. و http://domainname.fr

عکس

نخستین آنها، کمّی کردن تعداد دامنه‌هایی بود که هنوز فعال بودند و باید به‌طور عملی، به‌عنوان هسته، مورد استفاده قرار می‌گرفتند. این بررسی سنگین، توسط آرشیو اینترنت (IA) انجام شده است. برای کسب اطمینان از برخط بودن پاسخ آنها، هر دامنه با دو نشانی مختلف بررسی می‌شد (یعنی بررسی ۱/۷۸۰/۱۲۸ یو.آر.ال):

http://www.domainname.fr. و http://domainname.fr

هر دو صورت دامنه دارای دی.ان.اس است	۷۹ درصد
یک صورت دامنه دارای دی.ان.اس است	۱۴ درصد
هیچ یک از دو صورت دی.ان.اس ندارد	۷ درصد

شکل ۱. پاسخ به وجود دی.ان.اس[1] در اسامی دامنه آفنیک

براساس نتایج، فهرستی از یو.آر.الهای معتبر به دست آمد، تا آنجاکه، اگر نسخه www پاسخ نمی‌داد، از «domainname.fr»، یا http://www.domainname.fr یا http://domainname.fr استفاده می‌شد.

هسته های بدون دی.ان.اس در فهرست هسته به‌طور تصادفی وارد می‌شدند.

ما اسامی دامنه‌های (و به‌ویژه در مجموعه خزش فراگیر ۲۰۰۶) موجود در فهرست آفنیک را که در آرشیوهای وب نیز موجود بودند، بررسی کردیم. فهمیدن این نکته که فقط ۳۰ درصد از دامنه‌های آفنیک در مجموعه ما وجود داشت، بسیار حائز توجه بود.

این امر شاید به افزایش چشمگیر اندازه دامنه .fr مربوط باشد. به یمن ساده شدن پی در پی قوانین مختص .fr، از سال ۲۰۰٤، .fr، به‌طورمداوم افزایش یافته است. عمده‌ترین آنها، اجازه ایجاد دامنه سطح بالا برای اشخاص، در سال ۲۰۰٦ بود (تا این تاریخ، تنها اداره‌ها و انجمن‌های خصوصی مجوز ایجاد دامنه .fr را داشتند): با گذشت یکسال، .fr، بیش از ٦۳ درصد افزایش یافت. این امر باید مربوط به ایجاد تصور مثبت از ccTLD در کاربران اینترنت باشد. اشخاص ۳۰ درصد دامنه‌های .fr ثبت شده، و ۵۰ درصد ثبت‌نام‌های جدید را تشکیل می‌دهند [۲].

همچنین، وجود پدیده فوق، در فهرست آفنیک، ممکن است تاحدی به خاطر حضور اسامی دامنه-هایی باشد که یا پیوندی به وبگاه‌های دیگر ندارند و یا یا پیوندشان ضعیف است، و توسط روبات‌ها در سال‌های گذشته کشف نشده بودند.

این رشد فوق‌العاده، تا حد زیادی گسترش چشمگیر فهرست هسته را از ۲۰۰٦ تا ۲۰۰۷ توصیف می‌کند.

---

۱. DNS [سروری که دامنه های آدرس سایت‌ها را پشتیبانی می کند، مترجم].

شکل 1. پاسخ به وجود دی. ان. اس (1) در اسامی دامنه آفنیک

بر اساس نتایج، فهرستی از یو.آر.ال های معتبر به دست آمد تا آن جا که اگر نسخه www پاسخ نمی داد، از «domainname.fr» یا http://www.domainname.fr یا http://domainname.fr استفاده می شد.

هسته های بدون دی. ان. اس در فهرست هسته به طور تصادفی وارد می شدند.

ما اسامی دامنه های (و به ویژه در مجموعه خزش فراگیر 2006) موجود در فهرست آفنیک را که در آرشیوهای وب نیز موجود بودند بررسی کردیم فهمیدن این نکته که فقط 30 درصد از دامنه های آفنیک در مجموعه ما وجود داشت بسیار حائز توجه بود

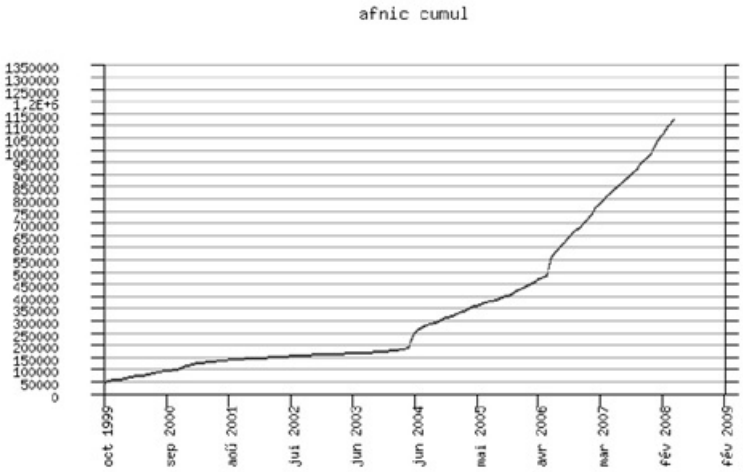
این امر شاید به افزایش چشمگیر اندازه دامنه .fr مربوط باشد به یمن ساده شدن پی در پی قوانین مختص fr. از سال 2004 fr. به طور مداوم افزایش یافته است. عمده ترین ،آن ها اجازه ایجاد دامنه سطح ، بالا برای اشخاص در سال 2006 بود (تا این ،تاریخ تنها اداره ها و انجمن های خصوصی مجوز ایجاد دامنه fr را داشتند): با گذشت یکسال fr. بیش از 63 درصد افزایش یافت این امر باید مربوط به ایجاد تصور مثبت از ccTLD در کاربران اینترنت باشد. اشخاص 30 درصد دامنه های fr ثبت شده و 50 درصد ثبت نام های جدید را تشکیل می دهند [2].

همچنین، وجود پدیده فوق در فهرست ،آفنیک ممکن است تا حدی به خاطر حضور اسامی دامنه هایی باشد که یا پیوندی به وب گاه های دیگر ندارند و یا یا پیوندشان ضعیف است و توسط روبات ها در سال های گذشته کشف نشده بودند.

این رشد فوق العاده تا حد زیادی گسترش چشم گیر فهرست هسته را از 2006 تا 2007 توصیف می کند.

1- DNS [سروری که دامنه های آدرس سایت ها را پشتیبانی می کند مترجم].

عکس



شکل ۲. تغییر حجم دامنه fr. از ۲۰۰۶ تا ۲۰۰۷[۱]

	۲۰۰۶	۲۰۰۷
هسته‌های استخراج شده از خزش‌های آلکسا	۲۰۷/۰۴۶	۵۸[۱]/۲۲۴
هسته‌های استخراج شده از خزش‌های قبلی	۴۲۷/۴۷۶	۲/۲۹۵/۸۹۰
فهرست آفنیک	–	۸۹۰/۰۶۴
تعداد کل پس از کاهش تکراری‌ها	۵۶۲/۶۳۷	۲/۸۸۸/۷۲۳

شکل ۳. تغییر اندازه فهرست هسته از ۲۰۰۶ تا ۲۰۰۷

۲ .۳. تنظیمات خزشگر

برداشت‌ها توسط هریتریکس[۲] انجام شده بود؛ خزشگر معتبر آرشیوی وب منبع باز توسط آرشیو اینترنت (IA) با کمک‌هایی از اعضای کنسرسیوم آی آی پی سی[۳] (به‌ویژه کتابخانه‌های ملی کشورهای حوزه شمال اروپا)[۴] [۱۱ و ۲۱][۵]. این خزشگر، در کتابخانه ملی فرانسه برای خزش‌های فراگیر، از ۲۰۰۴ و برای خزش‌های کانونی از ۲۰۰۶، مورد استفاده قرار گرفته است.

---

۱ . قابل دسترس در: http://www.afnic.fr/actu/stats/evolution (دیده شده: ۱۶مارس ۲۰۰۸).

2. Heritrix

3. IIPC

4. Nordic national libraries

۵ . کنسرسیوم بین المللی حفاظت اینترنت در ۲۰۰۳ توسط ۱۲ سازمان (آرشیو اینترنت و چند کتابخانه ملی) برای یافتن راه حل های مشترک به‌منظور آرشیوسای و اطمینان از دستیابی بلندمدت به انتشارات الکترونیکی در وب تأسیس شد. از ۲۰۰۷، عضویت برای سازمان های جدید باز است. برای اطلاعات بیشتر در باره اهداف و فعالیت های کنسرسیوم منبع [۱۳] را ببینید.

afnic cumu1

شکل 2. تغییر حجم دامنه fr. از 2006 تا 2007 (1)

شکل3. تغییر اندازه فهرست هسته از 2006 تا 2007

3.2. تنظیمات خزش گر

برداشت هـا توسط هریتریکس (2) انجـام شـده بود؛ خزش گر معتبر آرشـیوی وب منبع باز توسط آرشـیو اینترنت (IA) بـا کمک هایی از اعضای کنسرسیوم آی. آی. پی. سی (3) (به ویژه کتابخانه های ملی کشورهای حوزه شمال اروپا) (4) [11 و 21] (5) این خزش گر در کتابخانه ملی فرانسه برای خزش های فراگیر از 2004 و برای خزش های کانونی از 2006 مورد استفاده قرار گرفته است.

1- قابل دسترس در http://www.afnic.fr/actu/stats/evolution (دیده شده: 16 مارس 2008)

2- Heritrix

3- IIPC

4- Nordic national libraries

5- کنسرسـیوم بین المللی حفاظت اینترنت در 2003 توسط 12) سازمان (آرشـیو اینترنت و چنـد کتابخانه ملی) برای یافتن راه حل های مشترک به منظور آرشیوسای و اطمینان از دستیابی بلند مدت به انتشارات الکترونیکی در وب تأسیس شد از 2007 عضویت برای سازمان های جدید باز است. برای اطلاعات بیش تر درباره اهداف و فعالیت های کنسرسیوم منبع [13] را ببینید.

بـا قـابلیت تنظیم ،بالا هریتریکس تغییر مقادیر زیادی از تنظیمات شامل ،دامنه، اولویت های خزش، فیلتر ها، آداب روبات ها و جز آن را مقدور می سازد.

تنظیمات فوق به خزشـگر فرمان می دهـد چه چیزی را و چگونه بایـد برداشت کنـد. زیرا تنظیمات تأثیر زیادی در مجموعه سازی دارنـد و روبات لازم است مطابق با اهداف خزش گر قالب بندی شود.

1.3.2.حوزه

حوزه تعیین شـده برای خزشـگر تعریف می کنـد که کدام یک از یو. آر.ال های کشف شده باید در برداشت وارد و کدام یک کنار گذاشـته شود.

از این رو، حوزه خزش کتابخانه ملی ،فرانسه شامل هر وب گاهی می شود که در هر نام دامنه متعلق باشد به:

– دامنه سطح بالای fr؛

– دامنه سطح بالای re؛ و

– هر دامنه دیگری که خزش گر بـا آن مواجه شود زیرا از یک دامنه .fr یا از یک re تغییر مسیر داده است)طی بررسی خزش، 398,548 تغییر مسیر از اسامی دامنه آفنیک مورد توجه قرار گرفتند). به سبب تغییر مسیر از http://yahoo.fr به http://fr.yahoo.com لازم است خزش گر هر چیزی را که بخشـی از http://fr.yahoo.com است گزینش کنـد اگر چه در چنین موردی، خزش گر بایـد در همان میزبان باقی بماند به این ترتیب خزش گر هر چیزی را که بخشی از http://fr.yahoo.com است، بر می گزیند، نه از http://de.yahoo.com یا حتی http://fr.news.yahoo.com .

از یـک سو این حوزه بسـیار بزرگ به نظر می رسـد – شامل بیش از یک میلیون دامنه – و از دیگر سو محـدود است زیرا وب فرانسه تنها در دامنه های fir میزبانی نمی شود به نقل از گزارش منتشـر شده ،آفنیک کمتر از 30 درصد وب گاه های فرانسه در fir میزبانی می شود [2] این رقم بـا تحلیـل مجموعه سایت های انتخابات ،2007 مورد تأییـد قرار گرفته است. تنهـا 36 درصـد یو.آر. ال هـای این مجموعه برای میزبانی در دامنه fr. پذیرفته شده اند حتی اگر از قبل چنین تعیین کنیم که وب گاه های سیاسـی نماینده تام الاختیار کل وب فرانسه نیستند به طور مثال کاربرد وسیع org. کل توسط وب گاه های احزاب و اتحادیه های تجاری به بازنمون بیش از حد این دامنه سطح بالا منجر می شود(باید بدانیم fir در بخش عمده وب گاه های فرانسه استفاده نمی شود.

عکس

TLD	URLs	درصد
.fr	22938947	36.11
.com	18373574	28.93
.org	15955225	25.12
.net	3690655	5.81
.de	882656	1.39
.info	733634	1.16
.eu	257769	257769
.tv	110944	0.17
.us	106753	0.17
.re	99012	0.16
سایرTLDs	368148	0.58

شکل ٤. تعداد یو.آر.الهای هر دامنه سطح بالا (TLD)، خزش‌های کانونی انتخابات ۲۰۰۷

با وجود این، هدف ما برداشت دامنه‌های دیگری غیراز fr. و re. از طریق دنبال کردن پیوندهای تغییر مسیر داده شده بود. این موضوع، راهکاری برای اختیار کردن حوزه‌ای تغییرپذیر است. شاید رویکرد اکتشافی‌تر به گردآوری وبگاه‌های خارجی‌تر منتهی شود: این امر مشکلات قانونی (کتابخانه ملی فرانسه را به خارج از چارچوب قانون واسپاری می برد) و اقتصادی را پدید می آورد: گردآوری وبگاه‌های غیرمرتبط (مطابق با مأموریت ما) باعث اشغال فضایی می شود که از آن می‌توان برای برداشت سایت‌های معتبر فرانسه استفاده بهتری کرد. بنابراین، تأکید بر fr. یک انتخاب واقع بینانه بود. به علاوه، از آنجا که fr. به‌طور چشمگیری در حال رشد است، می‌توان امیدوار بود که این دامنه، بخش بزرگ‌تری از وب فرانسه را سال به سال نمایندگی کند. از دیگر سو، ممکن است در آینده، به‌طور مثال با استفاده از مراجعه به دی.ان.اس خودکار (محل جغرافیایی)، راه‌های نوینی را برای کشف سایت‌های فرانسوی خارج از دامنه fr. بررسی کنیم.

۲ .۳. ۲ . اولویت‌های خزش

هر URL که معلوم شود در حوزه خزش است، در «صف» خزشگر قرار می‌گیرد، یعنی در فهرست فایل‌های در انتظار خزش از آنجاکه ممکن است روبات نتواند همه یو.آر.الهایی که زمان اجرای یک خزش فراگیر در سر راهش می‌یابد، برداشت کند، برای خزشگر، مدیریت این صف و تعیین اولویت‌های خزش شدن مسائل اساسی هستند.

نخستین تصمیم مهم، انتخاب میان رویکرد «به ازای هر دامنه»[1] و «به ازای هرمیزبان»[2] بود. با رویکردِ به ازای هر میزبان، که در خزش‌های فراگیر قبلی استفاده شده بود، URLها در صف در حال انتظارخزش شدن در هر میزبان دسته‌بندی می‌شدند، و هر میزبان به‌صورت جداگانه تلقی می‌شد. این دسته‌بندی، به

1. per-domain
2. per-host

شکل 4. تعداد یو آر.ال های هر دامنه سطح بالا (TLD) خزش های کانونی انتخابات 2007

بـا وجود این هـدف ما برداشت دامنه های دیگری غیر از fr و re از طریق دنبال کردن پیونـد های تغییر مسـیر داده شده بود این موضوع راه کـاری برای اختیـار کردن حوزه ای تغییر پـذیر است. شایـد رویکرد اکتشـافی تر به گردآوری وب گاه های خارجی تر منتهی شود: این امر مشـکلات قانونی (کتابخانه ملی فرانسه را به خارج از چارچوب قانون واسپاری می برد) و اقتصادی را پدید می آورد گردآوری وب گاه های

غیر مرتبـط (مطابق با مأموریت ما) باعث اشـغال فضایی می شود که از آن می توان برای برداشت سایت های معتبر فرانسه اسـتفاده بهتری کرد بنابراین تأکید بر fr یـک انتخاب واقع بینانه بود. به علاوه، از آن جا که fr. به طور چشم گیری در حال رشـد است می توان امیدوار بود که این دامنه، بخش بزرگ تری از وب فرانسه را سـال به سـال نماینـدگی کنـد. از دیگر سو ممکن است در آینـده به طور مثال با اسـتفاده از مراجعه به دی.ان.اس خودکار (محل جغرافیایی) راه های نوینی را برای کشف سایت های فرانسوی خارج از دامنه fir بررسی کنیم

2.3.2. اولویت های خزش

هر URL که معلوم شود در حوزه خزش است، در «صف» خزشـگر قرار می گیرد، یعنی در فهرست فایـل هـای در انتظار خزش از آن جا که ممکن است روبـات نتوانـد همه یو. آر. ال هایی که زمان اجرای یک خزش فراگیر در سـر راهش می یابـد برداشت کنـد برای خزش گر مدیریت این صف و تعیین اولویت های خزش شدن مسائل اساسی هستند.

نخستین تصـمیم مهم انتخـاب میان رویکرد «به ازای هر دامنه» (1) و «به ازای هر میزبـان» (2) بود. بـا رویکردِ به ازای هر میزبان که در خزش های فراگیر قبلی استفاده شده بود URL ها در صف در حال انتظار خزش شدن در هر میزبان دسـته بندی می شدند و هر میزبان به صورت جداگانه تلقی می شد این دسته بندی به

1- per-domain

2- per-host

خزش وب گاه های بیش تری با چندین میزبان منجر می شد. وب نوشت های میزبانی شده در سکوهای تجاری یا صفحه های ،شخصی به عنوان میزبان های متمایز مقادیر زیادی از فضای مجموعه را اشغال می کنند به این ترتیب با استفاده از رویکردِ به ازای هر دامنه با این وب گاه ها به عنوان هویتی مستقل برخورد می شود.

برای خزش سال 2007 تصمیم گرفتیم رویکرد به ازای هر دامنه را اتخاذ کنیم. علت اصلی این امر تبعیت از فهرست آفنیک ،بود که فقط بر دامنه ها نظارت می.کند همچنین به عنوان یک نهاد و اسپاری می خواستیم به هر دامنه «شانس / فرصت» یکسانی برای برداشت شدن بدهیم. به علاوه، از آن جا که فهرست هسته 2007 بزرگ تر از فهرست های قبلی بود و مقادیر داده هایی که باید توسط آرشیو اینترنت (IA) بازیابی می شد افزایش نمی یافت (قسمت 5،2 را ببنید) بیم داشتیم ،که عمق خزش کم تری را کسب کنیم. بنابراین به وب گاه های تجاری یا سازمانی که اغلب چندین میزبان دارند و با خزش های کانونی بهتر برداشت می شوند، اهمیت زیادی بدهیم

در آغاز خزش لازم بود اسامی بزرگی مانند orange.fr skyblog.fr free.fr همچون سایر دامنه ها مدیریت شوند. تصمیم گیری برای اقدام به رفتار خاص با این وب گاه ها به تعویق افتاد: در طول خزش، هنگام تحلیل گزارش های ارسالی توسط مهندسان آرشیو اینترنت (IA) امکان انتخاب های مرتبط تر به نظر آسان تر می نمود.

همچنین برای جلوگیری از نمایندگی بیش از حد وب گاه های بزرگ برای هر سایت بیش ترین «بودجه» را تعیین کردیم برای روبات از یک دامنه برداشت بیش از URL 10,000 ممنوع بود. معنای این ،محدودیت این نبود که خزش گر وقتی به این مقدار می رسید به طور کامل متوقف شود؛ اگر کل بودجه خزش صرف نشده بود روبات اجازه داشت مقدار بیش تری برداشت کند

سرانجام برای اطمینان از اینکه روبات منابع کافی برای برداشت کل فهرست هسته در اختیار داشته ،باشد، سطح «میزان دوباره پر کردن» (1) را پایین تر انتخاب کردیم زمان اتصال به هسته هر «ریسمانِ» روبات (2) دستور برداشت صد URL که با دامنه مشابه را دریافت می کرد و سپس به هسته دیگر می رفت. پس از اتمام برداشت صد URL نخست هر هسته روبات به همان هسته باز می گشت.

این، رویکرد متوسطی میان رویکرد نخست پهنا (3) با هدف گردآوری وب گاه های مختلف تا جای ممکن و رویکرد آرشیو سازی ناب (4) است که در آن خزش عمیق یک سایت قبل از [آغاز] برداشت سایت بعدی انجام می گیرد

تصمیم گیری های فوق خطری را برای کتابخانه به همراه خواهد داشت در حقیقت، مجموعه فراهم شده نسبت به مجموعه های سال های قبل بسیار متفاوت است. با وجود این، برای ما بسیار اهمیت داشت که برای کشف راه های جدید در برآوردنِ ضرورت های وظیفه واسپاری قانونی خود، این شیوه گردآوری را تجربه کنیم.

1- replenishamount

2- thread" of the robot"

3- breadth-first

4- pure archiving approach

3.3.2.سایر تنظیمات

نقض رویکرد در هر دامنه این بود که: برای روبـات ممکن نبود به طور خودکـار دامنه های سـطح دوم (اس .ال.دی) (1) را کشف کند در مورد کار ما دامنه های سطح ،دوم بخش های فرعیِ تخصصی fr هستند. این دامنه ها یا «اسامی بخش سطح دوم» (2) مختص شناسایی یک صـنعت یا یک بخش متعارف (3) (مانند aeroport.fr برای فرودگـاه ها یا gouv.fr برای وب گاه های دولتی هسـتند یا «دامنه های توصیف سطح دوم» (4) که برای شناسایی یک فعالیت یا عنوان چند نوع، اختصاص داده شده اند (به طور مثال، asso.fr. برای فدراسیون ها یا .tm.fr برای دارندگان نشان های تجاری)

بـدون هیچ تنظیم خاصـی وب گاه های متفاوت که در دامنه های سـطح دوم مشابه میزبانی شده اند، از نظر روبات ها به عنوان یک هویت محسوب شده و بودجه یکسانی را دریافت نمی کنند برای اجتناب از این مسـئله، فهرست کلی دامنه های میزبانی شده در دامنه های سطح دوم را بـا دسـتور برخورد بـا این وب گـاه هـا به عنوان وب گاه های منفرد به روبات دادیم (این فهرست توسط آفنیک نیز برای کتابخانه ملی فرانسه تهیه شده است).

به محتوایی که که در صـفحه ای جای داده شده بود اما در دامنه ای متفاوت از خود آن صفحه میزبانی می،شد توجه خاصی مبذول .شد به روبات اجازه داده شد از سه جهش انتقالی حداکثر (5) پیروی کند که پیوندهایی جاسازی شده در صـفحه های وبی هستند. انجام این امر برای برداشت تعدادی زیادی فایل ،ویدئویی که هدف مهم این خزش بود ضرورت داشت.

در واقع ما سـعی داشتیم راه حل های بیشتری را برای برداشت فایل های ویـدئویی بیابیم. تعـداد ویـدئو در وب به طور مداوم رو به افزایش است و این مسـئله برای روبات هایی که آن ها را گردآوری می کنند چالش برانگیز است. مشکلات برداشت فایل های ویدئویی به اندازه و شیوه پخش آن هـا مربوط می شود (برای تحلیـل مشـکلات گردآوری رسـانه هـای کنونی، منـابع [4] و [7] را ببنیـد). کمبود فایـل های ویدئویی، به سـبب وجود محدودیت های فنی در خزش گر، با هدف ما از آرشیو سازی یعنی تهیه تصویری «بازنمون» از وب تناقض دارد ما تصـمیم گرفتیم تلاش های خود را بر دو سـکوی اصـلی پخش ویدئو مورد اسـتفاده (6 netsurfers) بر اینترنت فرانسه متمرکز سـازیم یوتوب [7 YouTube) و معـادل فرانسـوی آن دیلیموشن [8 Dailymotion) ما از مهندسان آرشـیو اینترنت خواستیم مهندسـی فنی دیلیموشن را مطالعه کنند (آن ها قبلا یوتوب را به شـکل بسـیار خوبی بررسـی کرده بودند) و نسخه های هریتریکس بهبود بخشند تا خزش گرها امکان گردآوری ویدئو فایل های این وب گاه ها را داشته باشند.

1- Second Level Domains (SLD)

2- second level sector names

3- regulated sector

4- second level descriptive domains

5- max-transhops

6- موجسواران

7- 25]

عکس

	۲۰۰۴	۲۰۰۵	۲۰۰۶	۲۰۰۷
قلمرو	میزبان	میزبان	میزبان	دامنه
جهش ـ حداکثر	---	۱۰۰	۱۰۰	۱۰۰
جهش انتقالی حداکثر	---	۳	۳	۳
مقدار بازذخیره (در URLها)	---	۱۰۰۰	۵۰۰	۱۰۰
بودجه (در URL)	---	۲۰۰۰۰۰	۲۰۰۰۰۰	۱۰۰۰۰
عامل تأخیر	---	۵	۴	۴
حداقل تأخیر (میان دو درخواست برای یک میزبان، به میلی ثانیه)	---	۵۰۰۰	۵۰۰۰	۵۰۰۰
حداکثر تأخیر (میان دو درخواست برای یک میزبان، به میلی ثانیه)	---	۱۰۰۰۰	۱۰۰۰۰	۱۰۰۰۰
تغییر مسیر	مناسب	مناسب	مناسب	مناسب
حداکثر اندازه URLهای بارگذاری شده	---	100 Mo	100 Mo	100 Mo

شکل ۵. تنظیمات خزشگر، از ۲۰۰٤ تا ۲۰۰۷

۲ .٤. پروتکل حذف روبات‌ها

قانون حق مؤلف ۲۰۰٦، به کتابخانه ملی فرانسه، اجازه سرپیچی از پروتکل حذف روبات‌ها (آر.ای.پی)[1] را می‌دهد: «[تولیدکنندگان یا ناشران] نباید در وبگاه‌های خود، رمزگذاری کنند یا مانع دستیابی مؤسسه‌های مسئول برای برداشت شوند[2]». به همین سبب، این کتابخانه، هنگام اجرای خزش‌های کانونی داخلی[3]، معمولاً به robots.txt توجه نمی‌کرد. در واقع، اغلب، حذف روبات‌ها برای جلوگیری از خزشگرهای گردآورنده صفحه‌هایی که قرار نیست نمایه‌سازی شود، توسط مدیران وب به‌کار می‌رود: راهنماهای تصویری یا صفحه های CSS. اما، این اسناد می‌توانند برای خزشگرهای آرشیوسازی مهم باشند، زیرا برای نمایش شکل اصلی صفحه‌های آرشیوی وب در آینده ضرورت دارند.

با وجود این، کتابخانه تصمیم گرفت از قوانین robots.txt برای خزش fr. سال ۲۰۰۷ خود اطاعت کند، همانطور که برای خزش‌های فراگیر قبلی نیز انجام می‌داد. تجارب خزش کانونی نشان داده بود که مدیران وب وقتی می‌فهمیدند که یک روبات، با شکستن قوانین robots.txt مربوط، مشغول خزش در سایت آنهاست، خوششان نمی‌آمد. به‌طور مثال، یک بلاگر فرانسوی که مورد خزش کتابخانه ملی فرانسه قرار گرفته بود، نامه‌های الکترونیکی خشونت‌آمیزی به کتابخانه فرستاد و یک تله ـ خزشگر برای پایین آوردن سرعت روبات کتابخانه ایجاد کرد، و پیامی در پست وبلاگش گذاشت تا دیگر مدیران وب را

1. robots exclusion protocol (REP)

۲ . همچنین این جمله به این معناست که بی ان اف، به‌طور مثال زمان دستیابی به داده هایی که در وبگاه رایگان نیست، به صورت قانونی حق درخواست رمز عبور و کدهای لازم را از مالک وبگاه برای خزش سایت او دارد.

۳ in-house . یا درون سازمانی

شکل 5. تنظیمات خزش گر، از 2004 تا 2007

4.2. پروتکل حذف روبات ها

قانون حق مؤلف 2006، به کتابخانه ملی فرانسه اجازه سرپیچی از پروتکل حذف روبات ها (آر.ای.پی) (1) را می دهد «[تولیدکنندگان یا

ناشـران] نبایـد در وب گاه های خود رمزگـذاری کنند یا مانع دستیابی مؤسـسه های مسـئول برای برداشت شوند». (2) به همین سـبب این کتابخـانه هنگـام اجرای خزش های کانونی داخلی (3)، معمولاً ـ به robots.txt توجه نمی کرد. در واقع ،اغلب، حـذف روبـات هـا برای جلوگیری از خزش گر های گردآورنده صفحه هایی که قرار نیست نمایه سازی شود، توسط مدیران وب به کار می رود: راهنماهای تصویری یا صـفحه های CSS ،اما این اسـناد می توانند برای خزش گر های آرشـیو سازی مهم باشـند زیرا برای نمایش شـکل اصـلی صـفحه های آرشیوی وب در آینده ضرورت دارند.

با وجود این کتابخانه تصـمیم گرفت از قوانین robots.txt برای خزش .fr سال 2007 خود اطاعت کنـد، همان طور که برای خزش های فراگیر قبلی نیز انجـام می داد. تجـارب خزش کـانونی نشـان داده بود که مـدیران وب وقتی می فهمیدنـد که یک ،روبات با شکسـتن قوانین robots.txt مربوط، مشغول خزش در سایت آن هاست خوش شان نمی آمد به طور مثال یک بلاگر فرانسوی که مورد خزش کتابخانه ملی فرانسه قرار گرفته بود نـامه هـای الکترونیکی خشونت آمیزی به کتابخـانه فرسـتاد و یـک تله – خزش گر برای پـایین آوردن سـرعت روبـات کتابخانه ایجاد کرد و پیامی در پست وبلاگش گذاشت تا دیگر مدیران وب را

1- robots exclusion protocol (REP)

2- همچنین این جمله به این معنـاست که بی ان اف، به طور مثـال زمان دستیابی به داده هایی که در وبگاه رایگان نیست به صورت قانونی حق درخواست رمز عبور و کدهای لازم را از مالک وبگاه برای خزش سایت او دارد.

3- in-house . یا درون سازمانی

به گذاشتن تله – خزش گر تشویق کند (1)! در واقع صفحه های و بی محدود شده توسط فایل های robots. txt اغلب تله های خزش گر را در خود جای می دهند به علاوه، گاهی هدف قوانین حذف جلوگیری خزش گرها برای جست و جوی URL های است که می توانستند موجب اضافه بار در وب گاه ها شوند (به طور ،مثال پیشنهادهایی به یک گروه بحث (2)) . وقتی امکان نظارت بر برداشت هر وب گاه شخصی وجود داشته ،باشد مسائل فنی و مشکلات تولیدکنندگان وبگاه به راحتی در طول خزش کانونی مدیریت خواهند شد. اما کتابخانه تمایلی برای مدیریت آن ها در مقیاس کلان نداشت و نمی خواست مهندسان آرشیو اینترنت زیر فشار مخالفت های مدیران وب عصبانی غرق شوند (به هر حال، سیاست IA احترام به حذف های robots.txt است).

اغلب مؤسسه های مجری خزش های فراگیر در وب به ویژه برداشت دامنه ملی سیاست احترام به حذف روبات ها را انتخاب کردند .[،17 10] Nearchive.dk مرکز مجازی (کتابخانه سلطنتی و کتابخانه ایالتی و دانشگاهی) با وظیفه آرشیوسازی دامنه دانمارک ترجیح می دهد از robots.txt چشم پوشی کند، زیرا این روبات بیش تر برای سایت های واقعا مهم شبکه مورد استفاده قرار می گیرد [3] (3) در واقع وب گاه ها با میزبانی محتوای بسیار ارزشمند مانند روزنامه ها یا – سایت های احزاب سیاسی رایج ترین وب گاه هایی هستند که از محدودیت های robots.txt استفاده می کنند [24].

5.2. برنامه ریزی با همکاری آرشیو اینترنت (IA)

برای خزش گر فراگیر سال 2007 آرشیو اینترنت و کتابخانه ملی ،فرانسه بر اندازه تعیین شده مجموعه توافق دارند به نظر می رسید حدود 300,000,000 URL برای هماهنگ کردن نیازهای خزش منطقی باشد. در صورت نیاز IA می توانست برای 10–15 درصد افزایش تصمیم بگیرد.

هر دو نهاد با سازماندهی روز به روز خزش موافق هستند حوزه و تنظیمات اصلی برای برداشت باید میان IA و کتابخانه ملی فرانسه به بحث گذاشته شود و کتابخانه درباره آن ها تصمیم گیری کند. مهندسان IA لازم است بر کار ماشین های خزش نظارت داشته باشند و هفته ای دو بار گزارش هایی به کتابخانه ملی ارسال کنند (در ادامه می آید) برداشت باید قبل از پایان سال خاتمه یابد و تا ماه های نخست سال 2008، داده ها (برای دسترسی ماشین های Wayback و NutchWAX) نمایه سازی شوند؛ و پس از آن، آرشیوها به قفسه های ذخیره خود در کتابخانه ارسال شوند (4) دو مهندس IA باید برای کمک به کتابخانه برای نصب ،قفسه ها مشاوره به تیم BNF و اطمینان از کیفیت مجموعه به پاریس بیایند.

1– کتابخانه به سرعت به اعتراض های بلاگرها پاسخ داد بعد از تبادل چندای میل با کتابخانه بلاگر با کاربرد قانون واسپاری وب آشنا شد و تصمیم گرفت تله اش را جمع کند

2– forum

3– توجه کنید که این تصمیم راهنمای متخصص آرشیو وب دانمارک برای انتخاب قوانین مؤدبانه محدود کننده روبات است تا مانع اضافه بار مورد درخواست سرور ها و پرونده های دعوی بالقوه شود.

4– این قفسه های Petaboxes ، دارای ظرفیت بالا ـ هزینه پایین سخت افزار ذخیره قدرت پایین ،هستند به فناوری های Capricorn اختصاص دارند. http://www.capricorn-tech.com [تاریخ دسترسی: 17 آوریل 2008]

3.خزش گر در حال کار

1.3.آزمایش خزش ها

قبل از اقدام به ،خزش باید به منظور بررسی روبات و عکس العمل های ماشین برای پیش بینی مشکلات آزمایش خزش ها انجام شود. این ،وظـایف همراه بـا نظارت بر روبات ها در طول خزش (3/2 را ببنیـد) و توصـیف مجموعه بعـد از برداشت، عملاً مهم ترین بخش فرآیند تضمین کیفیت برای یک خزش هستند (بخش 4 را ببینید). متدولوژی ،فوق برای چهار خزش فراگیر ما مورد استفاده قرار گرفت.

همـان طور که قبلاً توضـیح دادیم وظیفه اصـلی، قبل از آغاز خزش ادغام فهرست های هسـته متفاوت و بررسـی آن ها بود دیگر گام مهم اجرای «آزمایش خزش» بود: برای تجزیه و کشف حجم بالای URL در فهرست هسته روباتی راه اندازی شد اجرای خزش از آغاز تا پایان مورد نظر نیست، بلکه خزش در زمان کافی برای شناسایی اسامی دامنه غیر مرتبط یا URL های خطرناک به اجرا در می آید.

به طور مثال URL های ارسال کننـده کد «خطای 404» حـذف شدند به علاوه اسامی دامنه هایی که به شـمار کمی میزبان عمومی (در اکثر موارد ثبت کنندگان یا مالکان تصـرف کننده دامنه (1)) تغییر مسـیر داده بودند نیز از خزش حذف شدند. همچنین، اسامی دامنه برای کشت دامنه (یعنی با هدف بالا بردن رتبه یک وب گاه، با استفاده از اسامی چند دامنه ای برای یک IP واحد) شناسایی شدند.

2.3.ارتباط کتابخانه ملی فرانسه و آرشیو اینترنت در طول خزش

خزش فراگیر در 11 اکتبر 2007 راه اندازی شد و در 29 نوامبر به طور کامـل پایان یافت (یعنی بعـد از «خزش تکه ای») در زمان خزش رابطه IA/BNF بر اساس تحلیل «گزارش یافته ها» (2) بود. این گزارش، همه دامنه هـای موجود در صف را با نشان دادن URL هایی که قبلاً برای هر یک از آن ها برداشت شـده، مقـدار بودجه صـرف شده و URL های خزش شـده فهرست می کنـد. هدف از این کار تکمیل تخصص سنتی مهندسان IA در این زمینه بـا دانش تیمی هسـته های فرانسوی BNF است. این کـار در BNF توسط یک کتابـدار و یک مهندس هدایت گردید.

با افزایش بودجه به دامنه هایی که بالای URL 10,000 داشتند توجه خاصی مبذول شد. صرف نظر از اینکه داده ها مرتبط بودند یا نه این تعداد را به عنوان آسـتانه ای برای آن چه که باید بررسـی میشد تعیین کردیم. در تعریف ما داده های غیر مرتبط اسـناد با ارزش علمی اندک نیسـتند بلکه فایل های اضافی ای هسـتند که در اثر خصـیصه های مربوط به وب گاه های پاتولوژی ایجاد شـده انـد به طور مثال، تله های روبات (به خاطر ایجاد تقویم یا نسخه های جاوا) شمار نامحدودی URL برای صفحه های وبی یکسان ایجاد کردند. وب گاه های میرور (3) نیز مشـکل سـاز هسـتند زیرا چنـدین دامنه محتوای یکسان را با اسامی مختلف میزبانی می کننـد در اغلب موارد این دامنه ها حـذف شدند.

1- بیش از 100000 اسامی دامنه به تنها سه وب گاه - دو ثبت نام کننده و یک متصرف دامنه تغییر مسیر داده بودند.

2- frontier report

3- Mirror websites

با استفاده از کشت دامنه وب گاه های اصلی را شناسایی و حذف کردیم. بررسی های برخط نیز برای تعیین مطابقت رشته های حروف با تقویم ها پردازش شد تا آن ها را فیلتر و از ایجاد تله های روباتی جلوگیری کند.

هدف این کنترل ،هفتگی اطمینان از کسب 100 درصد کیفیت خزش نبود – به هر حال یک خزش فراگیر بود. هدف مدیریت بزرگ ترین صف ها به بهترین شکل ممکن و جلوگیری از هدر رفتن زمان و منابع زیادی توسط خزش گر بود.

3.3. «خزش تکه ای»

*«خزش تکه ای» (1)

بعد از سه هفته خزش مداوم شبانه روزی مهندسان IA تصمیم گرفتند خزش را متوقف کنند. آن ها به تحلیل QA در داده های بازیابی شده پرداختند تا دامنه هایی را مشخص کند که روبات به چهارمین سطح عمق آن ها (یعنی سه جهش نسبت به صفحه هسته) نرسیده بود این دامنه ها بار دیگر (با عنوان «خزش تکه ای») هر زمان که ممکن بود خزش شدند.

خزش تکه ای دیگری برای بازیابی فایل های ویدئویی شناخته شده در طول مدت برداشت راه اندازی شد اما بنا به دلایل فنی بارگذاری نشد (مشکلات فایل های میزبانی شده در دامنه های مختلف و جز آن).

4. پیامدهای خزش

تحلیل های زیادی درباره مجموعه برداشت شده صورت گرفت مهندسان ،IA که برای کمک به ما در زمان نصب قفسه ها به پاریس آمده بودند بسیار کارآمد بودند. نخستین هدف این تحلیل ها کنترل کیفیت داده های دریافتی بود همچنین مایل بودیم مجموعه ها را در مقیاس وسیع شناسایی کنیم: طیف اسناد موجود در قفسه ها در عمل چه بود شکل و عمق وب گاه های برداشت شده چگونه بود و جز آن. از این رو، هدف، تعیین کمیت و کیفیت مجموعه سال 2007 ما بود. سرانجام تحلیل نتایج خزش فراگیر 2007 ، و مقایسه این نتایج با خزش های قبلی (به ویژه خزش سال 2006) برای تعیین تأثیر تنظیمات جدید خزش و تصمیم گیری درباره این که آیا این خزش ها در جهت وظیفه واسپاری قانونی ما قرار دارند، ضرورت داشت.

1.4. اشکال اصلی

عکس

با استفاده از کشت دامنه، ویگاه‌های اصلی را شناسایی و حذف کردیم. بررسی‌های برخط نیز برای تعیین مطابقت رشته‌های حروف با تقویم‌ها، پردازش شد تا آنها را فیلتر و از ایجاد تله‌های روباتی جلوگیری کند.

هدف این کنترل هفتگی، اطمینان از کسب ۱۰۰ درصد کیفیت خزش نبود ـ به هرحال یک خزش فراگیر بود. هدف، مدیریت بزرگترین صف‌ها به بهترین شکل ممکن و جلوگیری از هدر رفتن زمان و منابع زیادی توسط خزش‌گر بود.

۳. ۳. «خزش تکه‌ای»[1]

بعد از سه هفته خزش مداوم شبانه‌روزی، مهندسان IA تصمیم گرفتند خزش را متوقف کنند. آنها، به تحلیل QA در داده‌های بازیابی شده پرداختند، تا دامنه‌هایی را مشخص کند که روبات، به چهارمین سطح عمق آنها (یعنی، سه جهش نسبت به صفحه هسته) نرسیده بود. این دامنه‌ها، باردیگر (با عنوان «خزش تکهای»)، هر زمان که ممکن بود خزش شدند.

خزش تکه‌ای دیگری برای بازیابی فایل‌های ویدئویی شناخته شده در طول مدت برداشت راه‌اندازی شد، اما بنا به دلایل فنی بارگذاری نشد (مشکلات فایل‌های میزبانی شده در دامنه‌های مختلف و جز آن).

٤.پیامدهای خزش

تحلیل‌های زیادی در باره مجموعه برداشت شده صورت گرفت. مهندسان IA، که برای کمک به ما در زمان نصب قفسه‌ها به پاریس آمده بودند، بسیار کارآمد بودند. نخستین هدف این تحلیل‌ها، کنترل کیفیت داده‌های دریافتی بود. همچنین، مایل بودیم مجموعه‌ها را در مقیاس وسیع شناسایی کنیم: طیف اسناد موجود در قفسه‌ها، در عمل، چه بود، شکل و عمق ویگاه‌های برداشت شده چگونه بود، و جز آن. از این رو، هدف، تعیین کمیت و کیفیت مجموعه سال ۲۰۰۷ ما بود. سرانجام، تحلیل نتایج خزش فراگیر ۲۰۰۷، و مقایسه این نتایج با خزش‌های قبلی (به‌ویژه خزش سال ۲۰۰٦)، برای تعیین تأثیر تنظیمات جدید خزش و تصمیم‌گیری در باره اینکه آیا این خزش‌ها در جهت وظیفه واسپاری قانونی ما قرار دارند، ضرورت داشت.

٤ .١. اشکال اصلی

تعداد	۲۰۰۶	۲۰۰۷
URLها	۲۷۱۶۹۷۴۵۶	۳۳۷۳۲۲۲۰۰
میزبان ها	۲۹۲۸۳۶۴	۱۵۸۹۴۵۸
دامنه ها	۳۸۲۵۴۰	۱۰۶۲۳۱۷
(از دامنه های fr.)	۱۳۱۱۳۶	۷۹۱۹۴۰

1. patch-crawl

patch-crawl -1

عکس

URL های هر دامنه	۷۱۰	۳۱۸
یو.آر.ال‌های هر میزبان	۹۳	۲۱۲
فایل‌های ARC یکتا	73073	91745
اندازه فشرده داده‌های یکتا (در Tb)	7,2	8,8

شکل ٦. اشکال اصلی خزش فراگیر fr. ۲۰۰۷

به‌خاطر افزایش چشـــمگیر فهرست هسته، رشـــد تعداد دامنه‌های برداشت شده پیش‌بینی می‌شد. همکاری با آفنیک (انجمن نامگذاری اینترنت فرانسه برای ثبت fr.) در سال ۲۰۰٦، به کتابخانه اجازه داد دامنه‌های fr. گردآوری شده را در شش مرحله کشف و برداشت کند.

به عبارت دیگر، در سال ۲۰۰۷ نسبت به ۲۰۰٦، با وجود افزایش یو.آر.ال‌های برداشت شده توسط IA، میزبان‌های کمتری خزش شدند. این امر، احتمال دارد به‌خاطر رویکردِ در ازای هر دامنه، مربوط به آخرین خزش فراگیر ما باشد.

شمار یو.آر.ال‌ها در ازای هر دامنه یا در ازای هر میزبان راهکاری ساده برای ارزشیابی «عمق متوسط» یک خزش است. این شکل، تفاوت‌های معنی‌دار فراوانی میان وبگاه‌ها را پنهان می‌کند: شکل در بخش ٦.٤ با جزئیات بیشتر بررسی شده است.

٤.٢. توزیع هر سرآیند[1]

گزارش سرآیند در باره خزش فراگیر سال ۲۰۰۷، حدود ١٦٠٤ نوع مختلف را نشان می‌دهد. تعجب‌آور نبود که یک سرآیند text/html، تنها دو سوم از فایل‌های برداشت شده را نمایش می‌دهد. به علاوه، ۹۷ درصد URLهای بارگذاری شده یکی از پنج نوع سرآیند پراستفاده: HTML، JPEG، GIF، PNG و PDF، را دارند. اگر فردی ســـرآیندهای اسناد برداشت شده طی خزش فراگیر ۲۰۰۷ را نگاه کند، تصور می‌کند وب فرانسه ۲۰۰۷ را بیشتر شامل متن و تصویر است.

با وجود این، باید در باره شـــکل‌ها بسیار محتاط باشیم. لازم است محدودیت‌های فنی روبات‌ها را نیز در نظر بگیریم؛ زیرا حتی اگر عملکرد خزشگر به‌طور مداوم پیشرفت کند، قادر به تجزیه و گردآوری قالب‌های مختلف فایل که در وب می‌یابد، نیست. قالب‌های فایلی پیچیده، بدون نمایش بوده یا به‌راحتی در مجموعه غایب هستند.

دلیل مناسب دیگر برای رعایت احتیاط این است که اطلاعات سرآیندِ مورد نیازِ محاسبه، همانی است که توسط سرور فرستاده می‌شود. در واقع، این اطلاعات قابل اعتماد نیست. گاهی، حتی سرور سرآیندی می‌فرستد که وجود ندارد (با کمال تعجب، «نرم افزار/ X- چیزی»[2] را در مجموعه خود یافتیم). به جز ١٦٠٤ سرآیند گوناگون، ١٤٠٠، سرآیند با کمتر از ۵۰۰ فایل مربوط هستند- می‌توان چنین استنباط کرد

1. MIME type
2. application/x-something

شکل .6 اشکال اصلی خزش فراگیر fr 2007

به خاطر افزایش چشم گیر فهرست ،هسته رشد تعداد دامنه های برداشت شده پیش بینی می شد. همکاری با آفنیک (انجمن نام گذاری اینترنت فرانسه برای ثبت fr) در سال 2006، به کتابخانه اجازه داد دامنه های .fr گردآوری شده را در شش مرحله کشف و برداشت کند.

به عبارت دیگر در سال 2007 نسبت به 2006 با وجود افزایش یو. آر.ال های برداشت شـده توسط IA میزبان های کمتری خزش شدند این ،امر احتمال دارد به خاطر رویکردِ در ازای هر دامنه مربوط به آخرین خزش فراگیر ما باشد.

شمـار .یو. آر.ال هـا در ازای هر دامنه یـا در ازای هر میزبـان راه کاری ساده برای ارزشـیابی «عمق متوسط» یک خزش است این ،شـکل تفاوت های معنی دار فراوانی میان وب گاه ها را پنهان می کند: شکل در بخش 4. 6 با جزئیات بیشتر بررسی شده است.

2.4. توزیع هر سرآیند

*توزیع هر سرآیند (1)

گزارش سرآیند درباره خزش فراگیر سال 2007، حدود 1604 نوع مختلف را نشان می دهد. تعجب آور نبود که یک سـرآیند text/html تنهـا دو سوم از فایل های برداشت شـده را نمایش می دهـد. به علاوه، 97 درصد URL های بارگـذاری شده یکی از پنج نوع سـرآیند پر استفاده HTML، JPEG، GIF، PNG و PDF را دارند اگر فردی سـرآیندهای اسناد برداشت شده طی خزش فراگیر 2007 را نگاه کند، تصور می کند وب فرانسه 2007 را بیشتر شامل متن و تصویر است.

بـا وجود ،این بایـد دربـاره شـکل ها بسـیار محتاط باشـیم. لازم است محـدودیت های فنی روبات ها را نیز در نظر بگیریم؛ زیرا حتی اگر عملکرد خزش گر به طور مداوم پیشرفت کند قادر به تجزیه و گردآوری قالب های مختلف فایل که در وب می یابد نیست قالب های فایلی ،پیچیده بدون نمایش بوده یا به راحتی در مجموعه غایب هستند.

دلیل مناسب دیگر برای رعایت احتیاط این است که اطلاعات سرآیند مورد نیاز محاسبه، همانی است که توسط سرور فرستاده می شود در واقع این اطلاعات قابل اعتماد نیست گاهی، حتی سـرور سرآیندی می فرستد که وجود ندارد (با کمال تعجب «نرم افزار/ X-چیزی» (2) را در مجموعه خود یـافتیم). به جز 1604 سرآیند گوناگون 1400، سـرآیند با کمتر از 500 فایل مربوط هسـتند- می توان چنین استنباط کرد

1- MIME type

2- application/x-something

که سرآیند ها به طرز نامناسبی تعیین شده اند.

به نظر می رسد اهمیت این مسئله سال به سال بیشتر شود. یو.آر.ال های خزش فراگیر 2004، دارای 554 سرایند گوناگون هستند؛ این رقم به 1024 در سال 2006 و به 1604 در 2007 تغییر یافت.

عکس

که سرآیندها به طرز نامناسبی تعیین شده‌اند.

به‌نظر می‌رسد اهمیت این مسئله سال به سال بیشتر شود. یو.آر.ال‌های خزش فراگیر ۲۰۰٤، دارای ٥٥٤ سرایند گوناگون هستند؛ این رقم به ۱۰۲٤ درسال ۲۰۰٦ و به ۱٦۰٤ در ۲۰۰۷ تغییر یافت.

MIME-type	یو.آر.الها	درصد
متن/ html	229257942	67.96
تصویر/jpeg	64222287	19.04
تصویر/gif	25376262	7.52
تصویر/png	3955885	1.17
نرم افزار/pdf	3955463	1.17
متن/ ساده	۲۲۵۶۷۵۹	۰٫۶۷
برنامه/ فلش -shockwave-x	۱۵۹۴۳۴۲	۰٫۴۷
متن/css	1432809	۰٫۴۲
برنامه/javascript-x	۱۴۱۵۲۳۰	۰٫۴۲
برنامه/XML	1083991	0.32
سایر	2771213	0.82

شکل ۷. ده رتبه نخست سرایند خزش فراگیر [۱] ۲۰۰۷

اما اگر نمی‌توانیم به سرآیند یک فایل منفرد اطمینان کاملی داشته باشیم، توزیع فراگیر تعیین شده برای صدها میلیون سند به احتمال زیاد قابل اعتماد است. تحول سرآیند می‌تواند به‌عنوان راهی برای تحلیل تغییرات و تمایلات در مقیاسی وسیع، دیده شود. به‌طور مثال می‌توان، از سال ۲۰۰٤ تا ۲۰۰۷، کاهش در استفاده از قالب GIF را، به نفع JPEG و فورمت باز PNG [۲] را مشاهده کرد (میزان تصاویر GIF در عرض این چهارسال تقریباً نصف شده است). بر همین اساس، میزان اسناد XML پنج برابر شده است. حتی وقتی به حجم اسناد برداشت شده نگاه می‌کنیم رشد فایل‌های XML در وب وضوح بیشتری می‌یابد: از ۸۸/۰۰۰ در سال ۲۰۰٤ به یک میلیون در ۲۰۰۷. شاید این رشد، تا اندازه‌ای، به خاطر کاربرد فزاینده آر.اس. اس خوان‌ها [۳] باشد (سرآیند صحیح برای یک فایل آر.اس. اس، «application/rss» است، اما غالبا «application/xml» یا حتی «text/xml» به جای آن به کار می روند).

---

۱. توجه کنید که این اشکال گاهی به چند سرآیند دسته بندی می شوند؛ به‌طور مثال، تعدادی اسناد JPEG معین با افزایش شماری اسناد دارای سرایند «image/jpeg»، «Image/jpeg» یا "image/JPEG".

۲. توجه کنید که نرخ های مشابه ی برای دامنه au. مشاهده شده اند (استرالیا، ۲۰۰۴ تا ۲۰۰۵)؛ درصد تصاویر GIF نصف شده است (از ۱۰ تا ۵ درصد). با وجود این، تصاویر png در استرالیا (۰٫۵۶ درصد) نسبت به فرانسه (۱٫۱۷ درصد) کمتر مورد استفاده قرارگرفته اند [۱۸].

3. RSS feeds

شکل 7. ده رتبه نخست سرایند خزش فراگیر 2007 (1)

اما اگر نمی توانیم به سـرآیند یک فایل منفرد اطمینان کاملی داشـته باشیم توزیع فراگیر تعیین شده برای صد ها میلیون سند به احتمال زیاد قابل اعتماد .است تحول سرآیند می تواند به عنوان راهی برای تحلیل تغییرات و تمایلات در مقیاسی وسیع دیده .شود به طور مثال می توان از سال 2004 تا 2007 کاهش در استفاده از قالب GIF را به نفع JPEG و فورمت باز PNG (2) را مشاهده کرد (میزان تصاویر GIF در عرض این چهار سال تقریباً نصف شده است). بر همین اساس میزان اسناد XML پنج برابر شده است. حتی وقتی به حجم اسناد برداشت شده نگاه می کنیم رشد فایل های XML در وب وضوح بیش تری می یابد: از 88/000 در سال 2004 به یک میلیون در 2007 شاید این ،رشـد تـا ،انـدازه ای بـه خـاطر کـاربرد فزاینـده .آر.اس. اس خـوان هـا (3) باشـد (سـرآیند صـحیح برای یـک فایـل آر.اس.اس، «application/rss» است، اما غالبا «application/xml» یا حتی «text/xml» به جای آن به کار می روند).

1- توجه کنید که این اشـکال گاهی به چند سـرآیند دسته بندی می شوند به طور مثال تعدادی اسناد JPEG معین با افزایش شماری اسناد دارای سرایند «image/jpeg" ، "Image/jpeg " یا "image/JPEG ".

2- توجه کنید که نرخ های مشابهی برای دامنه .au مشاهده شده اند (استرالیا 2004 تا 2005): درصد تصاویر GIF نصف شده است (از 10 تا 5 درصـد). با وجود ،این تصاویر png در استرالیا (0,56 درصد) نسبت به فرانسه (1,17 درصد) کمتر مورد اسـتفاده قرار گرفته اند [18]

3- RSS feeds

قـالب فـایلی دیگر رو به افزایش در مجموعه های خزش های فراگیر ، نرم افزار فلش شاک ویو (1) است. این رشد می تواند دو علت داشته باشد:محبوبیت در حال افزایش قالب فلش در وب توانایی بیش تر خزنده هریتریکس برای برداشت این نوع فایل

عکس

قالب فایلی دیگر رو به افزایش در مجموعه‌های خزش‌های فراگیر ، نرم‌افزار فلش شاک ویو[1] است. این رشـــد می‌تواند دو علت داشته باشـــد: محبوبیت در حال افزایش قالب فلش در وب، توانایی بیشتر خزنده هریتریکس برای برداشت این نوع فایل.

MIME Type evolution	2004	2005	2006	2007
text/html	68.11	67.22	70.15	67.96
image/jpeg	14.04	15.79	15.13	19.04
image/gif	12.70	11.09	8.05	7.52
application/pdf	1.36	1.39	1.19	1.17
image/png	0.79	0.73	0.87	1.17
text/plain	1.0833	1.19	1.01	0.67
application/x-shockwave-flash	0.2488	0.34	0.35	0.47
application/xml	0.07	0.16	0.50	0.32

جدول ۸. تحول در چند سرآیند از ۲۰۰٤ تا ۲۰۰۷

از منظر حفاظت بلندمدت، این اطلاعات بسیار ارزشمند است و قالبی را که ـ در مقیاس ملی و نیز بین‌المللی، باید تلاش‌های خود را بر آن متمرکزسازیم، ارائه می‌کند. کاربرد رو به افزایش قالب‌های باز، مانند PNG یا XML ، اخبار خوبی از این نقطه نظر است.

٤ .۳. فایل‌های ویدئویی

افزایش چهار قالب ویدئویی پراستفاده (Windows media video، Quicktime، Flash video و MPEG video) منجر به ایجاد حدود ٤۰۰۰۰ فایل ویدئویی برداشت شده در ۲۰۰٤ گردید، در مقابل ۱۲۰۰۰۰ فایل در چهار سال بعد (یعنی ۰/۰٤ درصد مجموعه). ما متوجه کاهش قالب ویدئویی MPEG در مقابل فلش شـــده‌ایم. در ۲۰۰٦ خزشـــگر هریتریکس فقط صد فایل ویدئویی فلش را برداشت کرده بود، یک سال بعد، نسخه مجوزدار خزشگر ما، برای برداشت محتوای میزبانی شده در سکوهای پخش ویدئویی به اجرا درآمد: هریتریکس سی هزار سند را گرد آورد. اراده ما برای تمرکز بر برداشت فایل‌های ویدئویی به هدفش رسیده بود: اگرچه اغلب آرشیوهای ما «ضعف‌هایی»[2] در ویدئو داشته است، در ۲۰۰۷، نمونه ویدئویی حجیم‌تری را به دست آوردیم.

1. application/x-shockwave-flash
2. holes

جدول 8. تحول در چند سرآیند از 2004 تا 2007

از منظر حفاظت بلندمـدت این اطلاعات بسـیار ارزشـمند است و قالبی را که – در مقیاس ملی و نیز بین المللی باید تلاش های خود را بر آن متمرکز سازیم ارائه می کند کاربرد رو به افزایش قالب های باز مانند PNG یا XML ، اخبار خوبی از این نقطه نظر است

4.3. فایل های ویدوئویی

افزایش چهار قالب ویـدئویی پر اسـتفاده (Flash video Quicktime Windows media video و MPEG video) منجر به ایجاد حدود 40000 فایل ویدئویی برداشت شده در 2004 ،گردید در مقابل 120000 فایل در چهار سال بعد (یعنی 0/04 درصد مجموعه) ما متوجه کاهش قالب ویدئویی MPEG در مقابل فلش شده ایم. در 2006 خزش گر هریتریکس فقط صد فایل ویدئویی فلش را برداشت کرده بود، یک سال بعـد، نسـخه مجوزدار خزش گر ما، برای برداشت محتوای میزبانی شـده در سـکو های پخش ویـدئویی به اجرا درآمد هریتریکس سـی هزار سـند را گرد آورد اراده ما برای تمرکز بر برداشت فایل های ویدئویی به هدفش رسیده بود اگر چه اغلب آرشیوهای ما «ضعف هایی» (2) در ویدئو داشته است، در 2007، نمونه ویدئویی حجیم تری را به دست آوردیم

---

application/x-shockwave-flash -1

holes -2

عکس

MIME-type	2004	2005	2006	2007
ویدئو/x-ms-wmv	4 408	7 705	33 936	39 218
ویدئو/quicktime	22 020	26 687	39 073	36 294
درخواست/x-fiv	0	0	104	31 556
ویدئو/mpeg	11 408	17 304	28 413	14 992
جمع	39 840	53 701	103 532	124 067

شکل ۹. تحول سرآیند فایل‌های ویدئویی از ۲۰۰٤ تا ۲۰۰۷

٤. ٤. توزیع هر TLD

دیگر کشف سه چهارم اسناد خزش شده متعلق به دامنه سطح بالای .fr تعجب‌آور نبود (می‌توان به این رقم، دامنه .re را که توسط آفنیک مدیریت می‌شود، افزود). با وجود این، شکل ۱۰ تأیید می‌کند که چنین چیزی با تنظیماتی که ما به کار گرفتیم، از فهرست هسته a.fr آغاز و تا مرزهای آن ادامه دادیم، ممکن است. دو نوع دامنه سطح بالا (تی.ال.دی.) ارائه شده در مجموعه نیز دامنه‌های سطح بالای کدهای عمومی و ملی[1] هستند. اغلب سایت‌های میزبانی شده تحت دامنه‌های سطح بالای عمومی (.com، .net، .org، .info) به احتمال، توسط مدیران وب فرانسوی تولید شده‌اند. کد کشوری مربوط به رتبه‌بندی دامنه‌های سطح بالا در شکل ۱۰ یا متعلق به کشورهای فرانسوی زبان (بلژیک و سویس) است یا به همسایگانی که (آلمان و انگلیس) فرانسه با آنها ارتباط تجاری عمده دارد مربوط می‌شود: این پدیده برای اسپانیا نیز مورد توجه قرار گرفت [۵]. دامنه .eu ویژه اروپاست.

TLD	تعداد یو.آر.الها	درصد
fr	259 869 452	77.12
com	59 843 624	17.76
net	4 951 932	1.47
org	3 171 196	0.94
de	2 808 359	0.83
eu	993 456	0.29
info	900 544	0.27
be	660834	0.20
ch	461 021	0.14
uk	434 315	0.13
re	381 746	0.11
Other TLDs	2 471 064	0.73

جدول ۱۰. تعداد یو.آر.الهای هر TLD، خزش‌گر فراگیر ۲۰۰۷

این ارقام، کاملاً با داده‌های مجموعه‌های قبلی، از ۲۰۰٤ تا ۲۰۰٦ مشابه است. به سقوط بالای .biz (که

1. general and country codes

شکل 9. تحول سرآیند فایل های ویدئویی از 2004 تا 2007

4.4.توزیع هر TLD

دیگر کشف سه چهارم اسناد خزش شده متعلق به دامنه سطح بالای .fr تعجب آور نبود (می توان به این رقم دامنه .re را که توسط آفنیک

مدیریت می شود افزود). با وجود ،این شکل 10 تأیید می کند که چنین چیزی با تنظیماتی که ما به کار گرفتیم، از فهرست هسته a.fr آغاز و تا مرزهای آن ادامه دادیم، ممکن .است دو نوع دامنه سطح بالا (تی.ال.دی.) ارائه شده در مجموعه نیز دامنه های سطح بالای کدهای عمومی و ملی (1) هستند اغلب سایت های میزبانی شده تحت دامنه های سطح بالای عمومی (org .net .com. info) به احتمال، توسط مدیران وب فرانسوی تولید شده اند کد کشوری مربوط به رتبه بندی دامنه های سطح بالا در شکل 10 یا متعلق به کشورهای فرانسوی زبان (بلژیک و سوییس) است یا به همسایگانی که (آلمان و انگلیس) فرانسه با آن ها ارتباط تجاری عمده دارد مربوط می شود: این پدیده برای اسپانیا نیز مورد توجه قرار گرفت .[5] دامنه .eu ویژه اروپاست

جدول 10. تعداد یو. آر. ال های هر ،TLD، خزش گر فراگیر 2007

این ارقام، کاملاً با داده های مجموعه های قبلی از 2004 تا 2006 مشابه است. به سقوط بالای biz. (که

---

general and country codes -1

سال ها قبل در 10 رتبه نخست دامنه سطح بالا جای داشت)، و ظهور ناگهانی eu. توجه کنید.

با وجود این، اگر به تعداد دامنه های میزبانی شده در یک دامنه سطح بالای خاص نگاه کنیم توزیع های متفاوت را متوجه می شویم.

عکس

سال‌ها قبل در ۱۰ رتبه نخست دامنه سطح بالا جای داشت)، و ظهور ناگهانی eu. توجه کنید.

با وجود این، اگر به تعداد دامنه‌های میزبانی شده در یک دامنه سطح بالای خاص نگاه کنیم، توزیع‌های متفاوت را متوجه می‌شویم.

TLD	دامنه های ۲۰۰۵(درصد)	دامنه های ۲۰۰۶(درصد)	دامنه های ۲۰۰۷ (درصد)
com	44.59	42.78	17.10
fr	26.86	34.28	74.55
net	5.37	5.95	2.06
org	4.39	4.69	1.46

جدول ۱۱. درصد دامنه هر تی.ال.دی (منحصر به com، fr، net. ، و org.)، از ۲۰۰۵ تا ۲۰۰۷.

مجموعه‌های ۲۰۰۵ و ۲۰۰٦، که با تنظیمات مشابه (فهرست هسته و تنظیمات خزش) شکل گرفته‌اند برتری بیشتر دامنه‌های سطح بالای عمومی را بر fr. نشان می‌دهند. این ارقام می‌تواند توسط حجم زیاد دامنه‌های غیرموجود در فهرست هسته، که توسط روبات لمس شده بود توصیف شود (یعنی جایی که یک یو.آر.ال. پیوند یافته به یک یو.آر.ال درون حوزه، مورد خزش قرار گرفته است). برخی پیوندها به دامنه‌های com. یا net. در صفحه‌های fr. قابل دسترس بودند، و به همین دلیل تا حدودی گردآوری شدند. به خاطر افزایش چشمگیر اندازه فهرست هسته[1]، میزان زیادی از دامنه‌های قابل دسترس در مجموعه ۲۰۰۵ و ۲۰۰٦ ـ اما نه در خزش فراگیر ۲۰۰۷ـ عرضه می‌شوند.

Robots.txt.۵. ٤

در سال ۲۰۰۷، پروتکل حذف روبات‌ها[2] مانع ما در آرشیوسازی ۱۵میلیون فایل بود، یعنی، ٤،۵ درصدِ فایل‌های کشف شده - و به وضوح از آرشیوسازی تمام اسنادی که روبات می‌توانست از آغاز تشکیل این فایل‌ها کشف کرده باشد جلوگیری کرد. این ارقام درکل، زمانی که فایل‌های robots.txt بارگذاری ٦ درصد اسناد کشف شده را متوقف کردند نسبت به خزش فراگیر ۲۰۰٦ بسیار پایین‌تر هستند. تفسیر این میزان مشکل است، زیرا آنها با آخرین مطالعات این پروتکل، که کاربرد در حال رشد robots.txt را شناسایی می‌کند، مغایرت دارند [به‌طور مثال منبع ۲۳].

زمان تحلیل آنچه خزش نشده بود، استفاده از سرآیند فایل‌ها امکان نداشت، زیرا سرور آنها را ارسال نکرده بود. با وجود این، می‌توانیم از پسوند فایل URLهای خواسته شده استفاده کنیم. تقریباً ٤۰ درصد این فایل‌ها تصویر هستند[3]. ممکن است برخی مدیران وب می‌خواستند مانع خزش روبات در این فایل‌ها

۱ . توجه کنید برخی دامنه های برداشت شده com. یا net. بین ۲۰۰۶ و ۲۰۰۷ تفاوت خیلی زیادی با یکدیگر ندارند. ۱۶۳۶۳۲ دامنه com. در ۲۰۰۶ نسبت به ۱۸۱۶۲۶ در ۲۰۰۷؛ ۲۲۷۴۶ دامنه net. در ۲۰۰۶ نسبت به ۲۱۸۵۳ در ۲۰۰۷.

2. robots Exclusion Protocol

۳ .jpg. (۲۶درصد)، .gif (۸ درصد)، .JGP (۲ درصد)، .png (۲ درصد). به علاوه توجه کنید که ۱۰۴،۱۰۹ فایل (۱ درصد) از فایل های CSS بودند.

جدول 11. درصد دامنه هر تی.ال.دی (منحصر به com. net fr. ، و org.)، از 2005 تا 2007.

مجموعه های 2005 و 2006، که با تنظیمات مشابه (فهرست هسته و تنظیمات خزش) شکل گرفته اند برتری بیش تر دامنه های سطح بالای عمومی را بر fr نشان می دهند این ارقام می تواند توسط حجم زیاد دامنه های غیر موجود در فهرست ،هسته که توسط روبات لمس شده بود توصیف شود (یعنی جایی که یک یو.آر.ال. پیوند یافته به یک یو.آر.ال درون حوزه مورد خزش قرار گرفته است). برخی پیوندها به دامنه های .com یا .net در صفحه های fr قابل دسترس بودند و به همین دلیل تا حدودی گردآوری شدند به خاطر افزایش چشمگیر اندازه فهرست هسته (1) میزان زیادی از دامنه های قابل دسترس در مجموعه 2005 و 2006 – اما نه در خزش فراگیر 2007 عرضه می شوند.

Robots.txt.5.4

در سال 2007، پروتکل حذف روبات ها (2) مانع ما در آرشیو سازی 15 میلیون فایل بود، یعنی، 4,5 درصدِ فایل های کشف شده – و به وضوح از آرشیوسازی تمام اسنادی که روبات می توانست از آغاز تشکیل این فایل ها کشف کرده باشد جلوگیری کرد این ارقام در کل زمانی که فایل های robots.txt بارگذاری 6 درصد اسناد کشف شده را متوقف کردند نسبت به خزش فراگیر 2006 بسیار پایین تر هستند. تفسیر این میزان مشکل است زیرا آن ها با آخرین مطالعات این پروتکل که کاربرد در حال رشد robots.txt را شناسایی می کند مغایرت دارند [به طور مثال منبع 23].

زمان تحلیل آن چه خزش نشده بود استفاده از سرآیند فایل ها امکان نداشت زیرا سرور آن ها را ارسال نکرده بود. با وجود این می توانیم از پسوند فایل URL های خواسته شده استفاده کنیم. تقریباً 40 درصد این فایل ها تصویر هستند (3) ممکن است برخی مدیران وب می خواستند مانع خزش روبات در این فایل ها

1- توجه کنید برخی دامنه های برداشت شده com. یا net بین 2006 و 2007 تفاوت خیلی زیادی با یکدیگر ندارند. 163632 دامنه com. در 2006 نسبت به 181626 در 2007؛ 22746 دامنه .net در 2006 نسبت به 21853 در 2007

2- robots Exclusion Protocol

3- 26) jpg درصد) 8)، gif درصد) 2) JGP درصد) 2) png درصد) به علاوه توجه کنید که 104,109 فایل (1 درصد) از فایل های CSS بودند.

شوند، زیرا آن ها توسط موتور های جست و جو نمایه سازی نشده بودند این فرض با کمک شیوه ای که روبات های ما این فایل ها را کشف کردند مورد تأیید قرار گرفته است: نیمی از آن ها (7/378/578)، هنگام پیگیری یک پیوند جاسازی شده یافت شدند.

بنابراین، تمکین از robots.txt مانع برداشت بسیار مرتبط داده هایی می شود که برای روبات های موتورهای جست وجو، غیر ضروری اما برای روبـات آرشـیوسازی ما بسـیار مفیـد است. در برخی مواقع، پروتکل حـذف روبات ها (آر.ای.پی) از دسترسـی ما به کل یک سایت جلوگیری می کرد بیش از 150/000 URL که به عنوان هسته هایی که به وسیله robots.txt محافظت می شدند به کار رفتند.

6.4عمق خزش

برای تعیین عمق خزش، ممکن است تعداد یو. آر. ال ها برای هر دامنه .fr را محاسبه کنیم.

عکس

شـــوند، زیرا آنها توسط، موتورهای جست‌وجو نمایه‌سازی نشده بودند. این فرض با کمک شیوه‌ای که روبات‌های ما این فایل ها را کشف کردند مورد تأیید قرار گرفته است: نیمی از آنها (۷/۳۷۸/۵۷۸)، هنگام پیگیری یک پیوند جاسازی شده، یافت شدند.

بنابراین، تمکین از robots.txt مانع برداشـــت بســـیار مرتبط داده‌هایی می‌شود که برای روبات‌های موتورهای جست‌وجو، غیرضروری، اما برای روبات آرشیوسازی ما بسیار مفید است. در برخی مواقع، پروتکل حذف روبات‌ها (آر.ای.پی) از دسترسی ما به کل یک سایت جلوگیری می‌کرد: بیش از ۱۵۰/۰۰۰ URL که به‌عنوان هسته‌هایی که به وسیله robots.txt محافظت می‌شدند به کار رفتند.

٤ .٦ عمق خزش

برای تعیین عمق خزش، ممکن است تعداد یو.آر.ال‌ها برای هر دامنه .fr را محاسبه کنیم.

تعداد دامنه ها	تعداد یو.آر.ال.ها
498777	10<
146356	10-100
103370	100-1000
43101	1000-10000
334	10000>

جدول ۱۲. تعداد یو.آر.ال ها برای هردامنه .fr ، خزش فراگیر ۲۰۰۷

تقریباً ۵۰ درصد دامنه‌های برداشت شده شامل ۱۰ یو.آر.ال یا کمتر هستند. این امر می تواند به سبب عدم دسترسی به سرور دائمی در طول خزش باشد. با وجود این، ما چندین آزمایش «دستی» انجام دادیم، یعنی برای کشف آنچه به‌صورت برخط قابل دسترس بود، در یک دو جین نام دامنه با آستانه ۱۰ یو.آر. ال. کلیک کردیم. این آزمایش‌ها نشان دادند که این وبگاه‌ها خالی بودند (مالکان آنها را خریداری کردند تا مطمئن شوند که آنها استفاده نخواهند شد، اما خودشان هم از آنها استفاده نمی‌کند) یا اینکه آنها برای کشت پیوند استفاده شدند.

این ارقام با ارقام متعلق به خزش فراگیر قبلی بسیار تفاوت داشتند.

تعداد دامنه ها	تعداد یو.آر.ال.ها
38 439	10<
32 258	10-100
41 352	100-1000
15 159	1000-10000
3 928	10000>

شکل ۱۳: تعداد یو.آر.ال ها برای هر دامنه .fr، خزش فراگیر ۲۰۰٦

جدول 12. تعداد یو.آر.ال ها برای هر دامنه .fr ، خزش فراگیر 2007

تقریباً 50 درصـد دامنه های برداشت شده شامل 10 یو.آر.ال یا کم تر هسـتند. این امر می تواند به سب عدم دسترسـی به سرور دائمی در طول خزش باشـد با وجود این ما چنـدین آزمایش «دستی» انجام دادیم یعنی برای کشف آن چه به صورت برخط قابل دسترس بود در یک دو جین نام دامنه با آستانه 10 یو.آر.ال کلیک کردیم این آزمایش ها نشان دادنـد که این وب گاه ها خالی بودند (مالکان آن ها را خریداری کردند تا مطمئن شوند که آن ها استفاده نخواهند شد، اما خودشان هم از آن ها استفاده نمی کند) یا این که آن ها برای کشت پیوند استفاده شدند.

این ارقام با ارقام متعلق به خزش فراگیر قبلی بسیار تفاوت داشتند.

شکل 13: تعداد یو.آر.ال ها برای هر دامنه .fr. خزش فراگیر 2006

ص: 277

سه تفاوت عمده میان سال های 2006 و 2007 درباره عمق دامنه های fr. عبارت اند از:

– دامنه های بسیار زیاد تقریباً خالی در 2007 که می توانست ناشی از افزایش اندازه فهرست هسته باشد.

– برخی دامنه ها در جایی که بیش از 10,000 یو آر ال آرشیو شده بود میان دو خزش به عدد 10 تقسیم شدند این امر پیامد محدود کردن «بودجه» برای 10,000 یو.آر.ال بود؛ همچنین شاید به خاطر رویکرد «به ازای هر دامنه»: دامنه های چندین میزبانه باز نمون نمی شوند.

– از سوی دیگر این رویکرد خزش عمیق تر تعداد بیش تر دامنه های «کوچک» یا «متوسط»، بین 100 و 10,000 یو.آر.ال را اجازه می داد.

ممکن است این ارقام تحت تأثیر تله های روبات ،باشند اما در حال حاضر نمی دانیم با کدام پسوند. سایر کنترل کیفیت ها، بیش از همه کنترل بصری وب گاه های شخصی برای تخمین این ارقام ضرورت دارد. با وجود این ارقام از دید قانون واسپاری فرانسه و اهداف اولیه این ،خزش، رضایت بخش به نظر رسد: با تضمین اینکه همه وب گاه های .fr موجود در مجموعه وظیفه سنگین برداشت وب گاه های کوچک و متوسط شاید به بهای وب گاه های بزرگ تر را داشته باشند

7.4. وب گاه های بزرگ

برای پالایش تحلیل عمق وب گاه ممکن است بر بزرگترین وب گاه ها تمرکز کنیم. هدف، بررسی علت چرایی بیش تر خزش شدن این وب گاه ها نسبت به دیگران است – آیا به این دلیل است که آن ها نسبت به دیگران بیش تر برخط هستند؟

اطلاعات مختلفی از نمایه خزش (سی.دی.ایکس) (1) اخذ شده بود: فهرست 50 دامنه بزرگ (مربوط به سال 2005 تا 2007)؛ و فهرست 1000 دامنه بزرگ در سال های 2006 و 2007.

1.7.4. دامنه ها

از این گذشته میان مجموعه های مختلف تناقص های زیادی وجود دارد نخستین تناقض، اندازه دامنه های بزرگ است تنها سه دامنه دارای بیش از 1,000,000 یو آر ال در مجموعه در 2007 نسبت به 25 یو.آر.ال در مجموعه 2006 – در بخش قبلی ارقام متفاوتی را دیدیم و می توانیم با همان دلایل این ارقام را توصیف کنیم ،هم چنین محتوای 50 سایت بزرگ نخست فهرست بسیار متفاوت است: تنها 20 درصد 50 دامنه بزرگ نخست مربوط به خزش گرهای فراگیر سال 2007 در وضعیتی معادل آن در 2006 ارائه شده اند.

بزرگ ترین وبگاه 2006 (free.fr) که دارای بیش از هفت میلیون یو.آر.ال بود، در سال 2007 فقط 40000 یو.آر.ال «وزن» دارد!

1- crawl index

عکس

نام دامنه ۲۰۰۷	تعداد یو.آر.ال ها	نام دامنه۲۰۰۶	تعداد یو.آر.ال.ها
asso.fr	3 984 821	free.fr	7 405 987
com.fr	1 760 957	amiz.fr	5 194 030
tm.fr	1 270 244	asso.fr	4 036 224
gouv.fr	534 599	1rencontre.fr	3 482 657
cci.fr	495 753	sportblog.fr	2 547 231
co.uk	408 881	gouv.fr	2 360 960
nom.fr	179 256	promovacances.fr	2 113 314
presse.fr	144 207	football.fr	1 895 302
dailymotion.com	108 222	mbpro,fr	1 885 549
notaries.fr	102 081	com.fr	1 856 720

شکل ١٤. ده دامنه بزرگ اول ، از ۲۰۰۵ تا ۲۰۰۷

بـــه علت انجام تنظیمات خاص در مجموعه ۲۰۰۷، بالاترین رتبه دامنه‌ها تقریباً مربوط به دامنه‌های ســـطح دوم است. از سوی دیگر، در خزش‌های ۲۰۰۵ و ۲۰۰٦، دو نوع وبگاه کشف شد: سکوهایی با میزبانی وبنوشـــت‌ها و وبگاه‌های شخصی (free.fr, sportblog.fr) که توسط رویکرد به ازای هر میزبان پشـــتیبانی شـــده‌اند، و وبگاه‌های تجاری با آگهی در صفحه های متعدد (به‌طور مثال .promovacances fr ، آژانس مســـافرتی برخط). خزش فراگیر ۲۰۰۵ نیز چند وبگاه دانشگاهی را نشان می‌دهد (١٦ وبگاه دانشگاهی در ۵۰ دامنه بزرگ اول)، مانند jussien.fr یا cnrs.fr. در سال‌های بعد، این وبگاه‌ها تقریباً در فهرست ۵۰ وبگاه بزرگ اول وارد شدند.

حضور گســـترده وبگاه‌های تجاری توصیف کننده تعداد دامنه‌های سطح بالای عمومی در ۵۰ دامنه بزرگ اول هســـتند: حتی برای مجموعه ۲۰۰۷، تنها ۳۲ درصد در .fr هســـتند. به این ترتیب، توصیف مجموعه منعکس‌کننده وب فرانسه در سال ۲۰۰۷ چنین است: عمدتاً، فضایی برای تجارت، خدمات و روابط اجتماعی.

٤ .۷. ۲ . دامنه‌های سطح دوم

	2006	2007	Evolution
asso.fr	4 036 224	3 984 821	↙
com.fr	1 856 720	1 760 957	↙
tm.fr	1 150 555	1 270 244	↗
gouv.fr	2 360 960	534 599	↙

شکل ۱۵. تحول دامنه‌های سطح دوم از ۲۰۰٦ تا ۲۰۰۷

شکل 14. ده دامنه بزرگ اول، از 2005 تا 2007

به علت انجـام تنظیمات خاص در مجموعه 2007، بالاـترین رتبه دامنه هـا تقریبـاً مربوط به دامنه های سـطح دوم است از سوی دیگر در خزش هـای 2005 و 2006، دو نـوع وبگـاه کشـف شـد سـکو هـایی بـا میزبـانی و بنـوشت هـا و وب گـاه هـای شخصـی (,free.fr sportblog.ir) که توسط رویکرد به ازای هر میزبان پشتیبانی شده اند و وب گاه های تجاری با آگهی در صـفحه های متعدد (به طور مثال

promovacances fr. ، آژانس مسافرتی برخط) خزش فراگیر 2005 نیز چند وب گاه دانشگاهی را نشان می دهد (16 وب گاه دانشگاهی در 50 دامنه بزرگ اول)، مانند jussien.fr یا cnrs.fr. در سال های بعد این وب گاه ها تقریباً در فهرست 50 وبگاه بزرگ اول وارد شدند.

حضور گسترده وب گاه های تجاری توصیف کننده تعداد دامنه های سطح بالای عمومی در 50 دامنه بزرگ اول هستند: حتی برای مجموعه 2007 تنها 32 درصد در fr. هستند به این ترتیب، توصیف مجموعه منعکس کننده وب فرانسه در سال 2007 چنین است: عمدتاً فضایی برای تجارت، خدمات و روابط اجتماعی

2.7.4. دامنه های سطح دوم

شکل 15. تحول دامنه های سطح دوم از 2006 تا 2007

ص: 279

توجه خـاص به دامنه های سـطح دوم در طول برداشت سال 2007 به کتابخـانه ملی فرانسه خزش مقادیر داده انـدکی پایین تر را نسـبت به مشـابه آن ها در برداشت سال 2006 اجازه داد. مهم ترین اسـتثنا دامنه سطح دوم .gou.fr است از 2٫3 میلیون به 500,000 یو.آر.ال سقوط کرده است. یک راه برای توصـیف این رقم کاربرد مشترک دامنه های سطح سوم و حتی چهارم در وب گاه های دولتی (به طور مثال، www rhone.pref.gouv.fr. یا www.auvergne.culture.gouv.fr) است.

تأکید بر فایل های ویدئویی مجموعه بهتری از وب گاه های پخش ویدئویی (1) به بار آورد دیلی موشن (2) 10 تا از پر رتبه ترین ها دامنه ها را وارد کرد (در سال 2006 ، تنها 30,000 یو.آر.ال در این دامنه برداشت شده بود). تعداد فایل های گردآوری شده در یوتیوپ دوبرابر شدند.

5. نتیجه گیری

برای برادشت دامنه ملی فقط یک راه وجود نـدارد. انتخاب های فنی متفاوتی (قبل و حین خزش، و حتی بعد از حذف URL ها) مجموعه را شکل می دهند سیاست مجموعه سازی حتی در مقیاس وسیع اعمال می شود. برای اجرای یک ،خزش دامنه بر طبق قانون و اهداف ،آن شناسایی این انتخاب ها و تخمین پیامد های آن ها ضروری است.

به طور سنتی در رسالت قانون و اسپاری برای تصمیم گیری درباره مرتبط بودن سند به حوزه ،مجموعه سه معیار به کار می رفته است باید با قالبی ،خاص قابل دسترس مردم و داخل مرزهای سـرزمین فرانسه موجود باشد، همه این ها باید ویژگی های جدید وب را پذیرفته باشـند و باید هنگام انجام خزش فراگیر به حساب آمده باشند.

هـدف برداشت وب «فرانسه» تأکید بر fr را توصـیف می کنـد در واقع این امر برآورده نمی شود اگر تصور کنیم 50 تا 60 درصد وب گاه هـای فرانسوی خارج از fir هسـتند اما در حال حاضـر تأکید بر fr. یـک انتخـاب واقع گرایانه و اقتصادی است همه محتوای فرانسه در fr موجود نیست، امـا هر چیزی که در fr است متعلق به فرانسه است. به علاـوه این تأکیـد قابل انعطاف است زیرا روبات اجازه دارد از تغییر مسیرهای fir به دیگر دامنه ها تبعیت کند. سرانجام دامنه .fr به سـرعت با ساده کردن قوانین ویژه این دامنه در حال توسعه است و امید می رود به زودی بخش بزرگ تری از دامنه فرانسه را بنمایاند. انتظار می رود این خواسـته با قوانین CCTLD جدید مربوط به ICANN متوقف نگردد.

همچنین تأکیـد بر تأکید بر fr بسـیار آسـان است زیرا مطمئن هستیم به لطف موافقت نـامه با آفنیک، قادر به برداشت همه جانبه این دامنه سطـح بالا هستیم این راه کاری است که با دومین اصل مهم قانون واسـپاری منطبق است: گردآوری کل تولید فرهنگی کشور هر آن چه به محض قرار گرفتن در دسترس عموم «کیفیت» می یابد یا «ارزش» دارد. از تعداد بسـیار زیاد هسـته آغاز کردن ضـمانتی است برای فراموش نکردن وب گاه های با پیوند کمتر یا نامشهور

1- videobroadcasting websites

2- Dailymotion

این اصول «غیر تبعیض آمیزانه» با ویژگی دیگر قانون واسپاری: میل به گردآوردن تمام قالب های انتشاراتی در حال ظهور ناسازگار نیستند. کتابخانه در گذشته [از قالب های مختلفی] پشتیبانی می کرد: متون، تصویر، صدا یا .ویدئو. با آرشیوسازی وب دیگر رفتار متفاوت با آن ها قابل تصور نیست زیرا این قالب ها به عناصری در شبکه ای یکسان پیوند یافته اند با وجود ،این در صورت ضرورت می توان به راه حل های مناسبی برای انواع مختلف رسانه دست یافت گردآوری نمایه سازی، حفاظت و دستیابی به فایل های متنی و دیداری – شنیداری یا تعاملی همواره به مسائل یکسانی ختم نمی شوند.

هنوز پرسش هایی باقی است: به طور مثال دشوار است که بگوییم، در حال حاضر، برای کسب تصویری جامع از وب فرانسه آیا رویکرد تأکید تنها بر دامنه ها بهتر از نگهداری به ازای هر میزبان به طور جداگانه است لازم است تحلیل بیشتری برای پاسخ به این سؤال انجام شود و کتابخانه ملی فرانسه، به شنیدن گزارش های بین المللی درباره این موضوع بسیار علاقه دارد سرانجام، مسئله، چگونگی تعریف یک وبگاه است زیرا این هویت هوشمند اغلب با میزبانی فنی سازگار نیست اگر وب گاه را به عنوان دامنه تعریف کنیم، رویکرد به ازای هر دامنه باید پذیرفته شود، زیرا به خزش بهتر وب گاه های کوچک یا متوسط منتهی می شود. اما اگر وبگاه را به عنوان هویتی فکری ایجاد شده توسط یک نویسنده یا یک ویراستار (یک یا چند شخص یک مؤسسه عمومی یا خصوصی) تعریف کنیم سایت و دامنه دیگر با هم سازگار نیستد. در واقع، تعداد یا حتی حجم عظیمی از وب گاه ها می توانند تحت یک نام دامنه میزبانی شوند، همچون و بنوشت های میزبانی شده در سکوهای تجاری. این سایت ها– با وجود مرتبط بودن – با راه کاری که در سال 2007 استفاده کردیم بازنمون نشد «به طور مثال، صفحه های شخصی فراوانی، به کل، از free.fr در 2006 برداشت شدند اما یک سال بعد ناپدید شدند برای اجتناب از این مسئله، در خزش های فراگیر آینده به کارگیری رویکردی خاص برای بسیاری از سکوهای عمومی برای میزبانی کردن برخی وب گاه ها یا وب نوشت های شخصی امکان پذیر است زیرا متوجه سکوهای پخش ویدئو در 2007 بودیم.

این تصمیم ها نشان دهنده اصلاحاتی است که می توانست هدف خز شگر هریتریکس باشد. قابلیت های بهتر برای تجزیه و برداشت قالب های پیچیده ،فایل یکی از مسائل ضروری است– از این نقطه نظر هریتریکس پیش از این خود را بسیار پیکربندی شده نشان داد. در چارچوب پروژه «خزشگر هوشمند»، ،تاکنون سه ویژگی ،دیگر توسط IIPC ،IA کتابخانه کنگره کتابخانه ،بریتانیا و کتابخانه ملی فرانسه – که هدف شان توسعه روبات هریتریکس است– ارائه شده است نخستین ،آن ها اجتناب از برداشت محتوایی است که از زمان آخرین خزش تغییری نکرده است: این ویژگی کاهش تکرار به صرفه جویی در تخمین منابع و ذخیره سازی آن ها و بنابراین به خزش عمیق تر وب گاه ها منجر می شود. دومین ویژگی، دادن مجوز به روبات برای اولویت بندی URL های داخل صف است. تلفیق رویکرد تمام گزینشی (در آغاز خزش) با به کارگیری قابلیت های کنترل خزش خودکار بهینه بسیار مفید خواهد بود. سومین توسعه، شناسایی خودکار بسامد تغییر وب گاه ها نیز به روبات اجازه شناسایی سایت هایی که باید به آن ها توجه خاصی مبذول شود، می دهد. از این رو، به طور مثال این امر باعث می شود تاریخ ها و بسامدهای خزش فراگیر را انتخاب کنیم، یا خزش های کانونی را در سایت های پر تغییر اجرا کنیم

در حقیقت اگر قرار است خزش های فراگیر با عمقی متوسط هر وب گاهی را یک یا دوبار در سال برداشت کننـد بایـد هدف از خزش های کانونی را چنین تعریف :کنیم خزش های کانونی بایـد از ابتدا قصد شان آرشـیو وب گاه های بزرگ و عمیق باشد؛ نه وب گاه های فرانسوی fr یا وب گاه های پرتغییر – آرشیو آن ها به بهترین شکل ممکن زیرا حتی خزش های کانونی اغلب برای برداشت کامل وب گاه های عظیم و گردآوری اسـناد موجود در وب پنهان کافی نیسـتند با وجود این باید به خاطر داشته باشیم که تهیه تصاویر تنها راه – و اقتصادی ترین راه– گردآوری حـافظه دیجیتـال فرانسه است و اتخاذ هر تصـمیم در این موضوع در سایر راه های آرشـیوسازی در راهبرد تلفیقی بایـد به حساب آیند.

تشکر و قدردانی

مـایلیم از کریس کارپنتر (1) و تیم آرشـیو ،اینترنت به عنوان همکـار مـا در این چهـار سال پایانی (از ابتـدا!) قـدردانی کنیم به ویژه ایگور رانیتوویک (2) که بر تمـام خزش های BNF مربوط به سـال های 2004 تا 2007 نظـارت داشت همراه با جان لی (3)، برد تافل (4) و میخائیل ماگین (5) که کمک کردند قفسه ها را نصب کنیم و قالب های نوین مجموعه را در پاریس و سان فراسیسکو تحلیل نماییم.

همچنین سـپاس بسیار از مارالیو چوونیک (6) برای دقت در ویرایش داریم سـرانجام، مراتب سـپاس مـان را به گیلـداس ایلین (7)، رئیس واسپاری دیجیتال به خاطر توصیه ها و حمایت هایش تقدیم می کنیم

1- Kris Carpenter

2- Igor Ranitovic

3- John Lee

4- Brad Tofel

5- Michael Magin

6- Mireille Chauveinc

7- Gildas Illien

منابع

Abiteboul, S., Cobena, Masanè s, J. and Sedrati, G. 2002. A First Experience in Archiving the French [1] Web. In Proceedings of the Research and advanced technology for digital libraries: 6th European conference .((Italy, 2002

AFNIC. 2007. French Domain Name Industry report. 2007 Edition. AFNIC, Saint Quentin en Yvelines. [2] http://www.afnic.fr/data/actu/public/2007/afnic-frenchdomain- name-report-2007.pdf

Andersen, B. 2005. The DK-domain: in words and figures. Netarkivet.dk, Aarhus, Copenhagen. [3] http://netarchive.dk/publikationer/DFrevy-english.pdf

Ashenfelder, M. 2006. Web Harvesting and Streaming Media. In Proceedings of the 6th International [4] Web Archiving Workshop (Alicante, Spain). http://www.iwaw.net/06/PDF/ iwaw06-proceedings.pdf

Baeza-Yates, R., Castillo, C. and Lopez, V. 2005. Characteristics of the Web of Spain. In Cybermetrics, [5] 9. http://www.catedratelefonica.upf.es/webes/2005/Characteristics Web-Spain.pdf

Baeza-Yates, R., Castillo, C., Marin, M. and Rodriguez, A. 2005. Crawling a country: Better Strategies [6] than BreadthFirst for Web Page Ordering. In Proceedings of the 14th international conference on World .(Wide Web (Chiba, Japan

Baly, N. and Sauvin, F. 2006. Archiving Streaming Media on the Web, Proof of concept and Firsts [7] Results. In Proceedings of the 6th International Web Archiving Workshop (Alicante, Spain). http://www.iwaw.net/06/PDF/iwaw06-proceedings.pdf

Brin, S. and Page, L. 1998. The Anatomy of a Large-scale Hypertextual Web Search Engine. In [8] Computer Networks and ISDN Systems, 30 (1-7), 107-117. http://www7.scu. edu.au/programme/fullpapers/1921/com1921.htm

Dailymotion. Dailymotion [Accessed: May 10, 2008]. Partagez vos vidé os. [9] http://www.dailymotion.com

Gomes, D and Silva, M. Characterizing a National Community Web. ACM Transactions on Internet [10] Technology (volume 5, issue 3), New York, 508-531. http://xldb.fc.ul.pt/ daniel/gomesCharacterizing.pdf

.[Heritrix. Heritrix Home Page. http://crawler.archive.org [Accessed: May 22, 2008 [11]

.[HTTrack. HTTrack Website Copier. http://www.httrack.com [Accessed: April 17, 2008 [12]

IIPC. International internet preservation consortium-welcome. http://www.netpreserve. org. [13] .[[Accessed: May 24,2008

Illien, G., Aubry, S., Hafri Y. and Lasfargues, F 2006. Sketching and checking quality for web archives: [14] a first stage report from BnF. Bibliothè que nationale de France, Paris. http://bibnum.bnf.fr/conservation/index.html

Illien, G. 2006. Web archiving at BnF. In International Preservation News, Paris, BnF, 27-34. [15] http://www.ifla.org/VI/4/news/ipnn40.pdf

Kimpton, M., Braggs, M. and Ubois, J. 2006. Year by Year: From an Archive of the Internet to an [16] .Archive on the Internet. In Web Archiving, J. Masanè s, Ed, Springer, Berlin, Heidelberg, New York

Koerbin, P. 2005. Report on the crawl and Harvest of the Whole Australian Web Domain Undertaken [17] during June and July 2005. National Library of Australia, Canberra. http:// pandora.nla.gov.au/documents/domain-harvest-report-public.pdf

Koerbin, P. 2008. The Australian Web domain harvests: a preliminary quantitative analysis of the [18] archive data. National Library of Australia, Canberra. http://pandora.nla. gov.au/documents/auscrawls.pdf

Masanè s, J. 2002. Towards continuous Web Archiving: First results and an agenda for the future. In D- [19] Lib Magazine, 8 (12).http://www.dlib.org/dlib/december02/masanes/12masanes.html

Masanè s, J. 2006. Selection for Web Archives. In Web Archiving, J. Masanè s, Ed, Springer, Berlin, [20] .Heidelberg, New York

Mohr, G., Kimpton, M., Stack, M, and Ranitovic, I. 2004.Introduction to Heritrix, an archival quality [21] Web crawler.Paper presented at the 4th International Web Archiving Workshop (Bath, United Kingdom, 2004). http://www.iwaw.net/04/Mohr.pdf

Najork, M. and Wiener, J. L. 2001. Breadth-First Search Crawling Yields High-Quality Pages. In: [22] Proceedings of the 10th international conference on World Wide Web. Elsevier Science, Hong Kong, 114- .118

Sun, Y. Zhuang Z., Councill I. and Giles C L. 2007 Determining Bias to Search Engines from [23] Robots.txt. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. IEEE Computer Society Washington, 149-155. http://www.personal. psu.edu/yus115/docs/sun-robotstxtbias. pdf

:Sun, Y. Zhuang Z. and Giles C. L..2007. A large-scale study of robots.txt. In WWW '07 [24]

ص: 284

Proceedings of the 16th international conference on World Wide Web, ACM Press, New York 1123–1124. http://www2007.org/posters/poster 1034.pdf

YouTube. YouTube – Broadcast Yourself. http://www.youtube.com [Accessed: May 15, 2008 [25]

چکیده

مجموعه های کتابخانه ملی فرانسه (1) بخشی از میراث ملی هستند و به طور تقریبی 31 میلیون سند از همه نوع (کتاب نشریه نسخه های خطی عکس ها، نقشه ها و غیره) را شامل می شوند. چالش های جدید مجموعه با گسترش اینترنت ایجاد شده اند. کتابخانه ملی فرانسه در قالبی بین المللی رهنمون های خط مشی، گردش کارها، و ابزارها را توسعه می دهد تا قسمت های مرتبط و معرف بخش فرانسوی اینترنت را جمع آوری کند و حفاظت و دسترسی به آن ها را سازماندهی کند.

آرشیوهای وب حوزه ملی فرانسه به عنوان خدمتی جدید توسعه یافت به عنوان کاربردی جدید عرضه شد و در آوریل 2008 در دسترس عموم قرار گرفت. از آن ،پس راهبردهایی بوده است و توسعه می یابد تا کتابداران را درگیر آن کند و آن را در دسترس کاربران نهایی قرار دهد

این ،مقاله تجربه کتابخانه ملی فرانسه را به ویژه با تمرکز بر این چهار موضوع بررسی می کند:

- ساختمان مجموعه آرشیو وب به عنوان مجموعه ای جدید و چالش برانگیز،
- کشف منبع: خدمات و ابزارهای دسترسی برای کاربران نهایی
- کاربرد: اطلاعات و ارقام
- مشارکت: راهبردهایی برای ساخت یک انجمن کتابداران و رسیدن به کاربران نهایی

کلید واژه ها: آرشیوهای ملی آرشیو کردن وب گاه ها، ساختمان مجموعه، کشف منبع، کاربران نهایی، کاربرد، فرانسه

1- National Library of France (BnF)

معرفی آرشیوهای وب به عنوان یک خدمت جدید کتابخانه: تجربه کتابخانه ملی فرانسه

اشاره

نوشته: سارا اوبری (1) | ترجمه: زهرا تهوّری (2)

ساختمان مجموعه آرشیوهای وب به عنوان مجموعه ای جدید و چالش برانگیز اینترنت نقش مهمی در زندگی های روزانه ما به عهده گرفته است به عنوان مثال، مدیریت الکترونیکی یادگیری الکترونیکی تجارت الکترونیکی انتشارات آنلاین هنرهای دیجیتال بلاگ ها و فضاهای جدید عمومی بحث و گفتگو بسیاری از فعالیت ها تاحدی یا به طور کامل به سمت وب حرکت کرده و تغییر مکان داده و فعالیت های جدیدی ایجاد کرده.اند با افزایش مستمر و رو به رشد تعداد کاربران اینترنت (امروزه، حدود 35 میلیون در فرانسه) و تعداد فزاینده وب گاه های فرانسوی (فقط 1/7 میلیون دامنه .fr. به نام Top Level Domain or TLD ثبت می شود) (3)، حیاتی است که این نوع انتشارات و رسانه ارتباطی را که هنوز تاحدی برای یک کتابخانه ملی جدید است، بررسی کرد.

یک وبگاه در موارد زیر متفاوت از دیگر انواع انتشارات است:

- محدود به شکل خاصی همچون یک تکه کاغذ یا یک قطعه موسیقی روی دیسک نیست بلکه وابسته به زیرساخت شبکه ای پیچیده تری است یک وبگاه یک فایل پی.دی.اف تنها (شبیه پایان نامه های

1- Sara Aubry, Web Archiving Project Manager, IT Department, National Library of France, Quai François Mauriac, sara.aubry@bnf.fr

2- کارشناس سازمان اسناد و کتابخانه ملی ایران

3- Observatoire du marché des noms de domaine en France, FR Network Information Center, [http://www.afnic. fr/actu/observatoire [last accessed on 2010-06-15].

الکترونیکی) یک JPEG تنها یا یک تصویر TIFF (شبیه تصاویر و عکس ها) نیست بلکه چند وجهی است: یک صفحه وب گردآوری تعدادی عوامل (متون، تصاویر، نوشته ها، شیوه نامه ها (1)، فایل های صوتی و ویدئویی و غیره) است که ممکن است از جاهایی دیگر (سامانه فایل محلی پایگاه داده وبگاه راه دور دیگر و غیره) بیاید و هنگامی که هر کاربر وبگاه را با یک مرورگر مشاهده می کند، جمع آوری شوند. تحلیل گردآوری یک نمونه 2/9 میلیون وبگاه فرانسوی در سال 2007 نشان داد که حدود 1600 نوع رسانه اینترنتی متفاوت وجود دارد. (2)

• یک وبگاه آغازی ندارد و نمی تواند تا آخر خوانده شود موجودیتی عقلانی است که می تواند توسط کاربری متفاوت از کاربر دیگر دیده شود. در ارتباط با شبکه ای از دیگر وب گاه های پیوند یافته توسط پیوندهای ،فرامتنی، شبکه ای از پیوندهایی که حتی قوی تر از شبکه یک انتشارات علمی مبتنی بر استناد ها و اطلاعات کتاب شناختی است وجود دارد.

• وب گاه ها بسیار متعدد هستند و وب هیچ مرزی ندارد. بر اساس آخرین نظر سنجی نتکرافت (3)، بیش از 206 میلیون وبگاه وجود دارد. (4) بیش تر آن ها از هر کشوری در دنیا قابل دسترس هستند و ممکن است بخشی از یک مجموعه ملی بر اساس رهنمودهای خط مشی توسعه مجموعه ملی حقوقی یا سازمانی باشند.

• محتوای وب همیشه درست شبیه یک بخار در حرکت است صفحات وب ممکن است در عرض یک روز چندین مرتبه روزآمد شوند (در صفحات وب روزنامه هایی مانند لوموند (5) و لیبراسیون (6) علائم کوچکی وجود دارد که نشان می دهند محتوی چند دقیقه قبل روزآمد شده است).

• وب گاه ها و محتوای وب نیز گذرا هستند: یک صفحه وب ممکن است در هر زمانی و به چند دلیل ناپدید شود: توقف اختیاری یا غیر عمدی توسط وب مستر عدم تمدید نام دامنه شکستن دیسک یا مشکلات دسترسی شبکه به سرور میزبان و غیره محتوای وب به یک حادثه خاص پیش بینی شده یا غیر مترقبه پیوند دارد و به ویژه در معرض خطر است به مناسبت یک گزینش مشارکتی و طرح گردآوری وب گاه های سیاسی در طول انتخابات ریاست جمهوری 2007 فرانسه کتابخانه لیون (7) دریافت که 52 درصد از 421 وب گاهی که انتخاب شده بودند یا به طور کامل یا به طور تقریبی پنج ماه بعد از رأی گیری بسته شدند. (8)

چارچوب حقوقی

هر زمان که یک نوع جدید ماده نمایش و ایجاد از جمله فناوری های متنوع جدیدی که در فرانسه ظاهر

1- style sheets

2- Legal deposit of the French Web: harvesting strategies for a national domain. France Lasfargues, Clé ment Oury, Bert Wendland, IWAW, 2008: http://iwaw.net/08/IWAW2008-Lasfargues.pdf [last accessed on 15-06-2010].

3- Netcraft

4- Netcraft May 2010 Web Server Survey, http://news.netcraft.com/archives/category/web-server-survey/

.[[last accessed on 2010-06-15

Le Monde -5

Libération -6

Library of Lyon -7

La netcampagne des législatives 2007 en Rhône-Alpes: la course au Net et après, - 8

.[http://www.pointsdactu.org/ article .php3?id_article=863 [last accessed on 2010-06-15

شدنـد، اختراع میشـد کتابخـانه ملی فرانسه نخست آزمـایش می کرد، سـپس سازمـانش را بـا جمع آوری، حفـظ و دسترس پـذیری به این انتشارات دیجیتال متولد شده وفق می داد زمان آن رسـیده است که پس از ،کتاب ها حکاکی ها پارتیسیون های موسیقی عکس ها ،پوستر ها مدارک صوتی تصویری و چندرسانه ای وب گاه ها نیز بایگانی شوند.

قـانون میراث فرانسه (1)، اکنون حق مـالکیت IV (ماده1- L 311 تا پایان (1- L 133) از قـانون حق مؤلف و نگرش هـای حقوقی در جامعه اطلاعاتی 2006-961 را که مطابق دستور عمل 29 / 2001/ EC پارلمان اروپا و شورای 22 می 2001 در مورد سازگاری جنبه های خاص کپی رایت و حقوق مرتبط با جامعه اطلاعاتی (2) است به ثبت می رساند.

این قانون که به طور رسمی سوم آگوست 2006 منتشر شد:

• دامنه واسپاری حقوقی (3) اینترنت را به این موارد گسترش می دهد: مشـمول واسـپاری حقوقی هر علامت، نشانه نوشته، تصویر صدا یا پیغام های هر نوع ارتباط با عموم به وسـیله کانال های الکترونیکی نیز می شود (ماده 39) قانون برای همه نوع انتشارات الکترونیکی آنلاین از جمله مجموعه ای از علائم نشانه ها، تصاویر، صداها یا هر نوع پیغامی که در اینترنت در دسترس عموم باشد، قابل اجر است نه تنها وب گاه ها، بلکه خبرنامه ها و رسانه های جاری نیز شامل این تعریف می شوند؛

• چگونگی تقسـیم مسـئولیت های واسـپاری وب بین سازمان های تحت قیومیت را تعریف می کند: مؤسـسه خبرگزاری ملی که مسـئول حفظ میراث صوتی تصویری فرانسه است وب گاه های مرتبط با ارتباطات صوتی تصویری (به طور عمده رادیو و تلویزیون) و کتابخانه ملی فرانسه تمامی وبگاه ها را جمع آوری خواهند کرد؛ یک حکم در دست اقدام است تا فرایندهای گزینش و دسترسی را به اجرا درآورد؛

• راهبردهای جمع آوری را خاص می کند در کتابخانه ملی فرانسه و اسـپاری حقوقی اینترنت نیاز به اجازه از ناشران ندارد و به جمع آوری خودکار بخش عمده اولویت می دهد: سازمان های تحت قیومیت ممکن است مواد را مطابق با فرایندهای خاص واسـپاری تولید کنندگان از اینترنت جمع آوری کنند. قانون نیز تصریح می کند که هیچ مانعی همچون برقراری ارتباط (4) رمز عبور یا شکل های دیگر محدودیت دسترسی ممکن نیست توسط ناشران برای محدود کردن این فرایند استفاده شود.

دامنه

اگر چه بیشتر وب عمومی توسـط هر کس در فرانسه می توانـد ملاحظه شود، از نظر فنی و حقوقی غیر ممکن است کل وب را بایگانی کرد. کتابخانه ملی فرانسه دستور دارد تا وب گاه های دامنه ملی فرانسه را جمع آوری کند، یعنی:

1- The French Heritage Law, or «Code du patrimoine»

2- Loi n°2006-961 du 1 août 2006 relative au droit d'auteur et aux droits voisins dans la société de l'information , JORF , http://www.legifrance.gouv.fr/affchTexte.do?cidTexte = TEXT000000266350dateTexte= [last accessed on 2010-06-15].

3- legal depository

• به عنوان یک هسته هر وب گاهی با دامنه fr TLD. یا هر TLD مشابه دیگری که به قلمرو مدیریتی فرانسه مربوط می شود (به عنوان مثال re برای جزیره فرانسوی La Reunion)

• هر وب گـاهی (به احتمـال خارج از دامنه fr) که تولیـد کننـده اش از نظر جغرافیـایی تحت قلمرو فرانسه هست (به طور معمول این می تواند در صفحات وب یا با استفاده از سرورهای خاص بررسی شود)؛

• هر وب گـاهی (به احتمـال خارج از دامنه fr) که بتوان ثـابت کرد محتوای تولیـد شـده اش در قلمرو فرانسه به نمایش گذاشـته می شود (بررسی این معیار اخیر چالش برانگیز تر است اما فرصتی برای تفسیر و مذاکره با کتابخانه و تولیدکنندگان اینترنت ایجاد می کند).

ابزارها و روشهای جمع آوری

گر چه ما از یک «واسپاری» حقوقی صحبت می کنیم وب گاه ها در حقیقت توسط ناشران به کتابخانه واسپاری نمی شوند. در عوض توسط تکه هایی از نرم افزارهایی به نام روبات های خزنده آرشیو جمع آوری می شوند یک خزنده آرشیو شبیه خزنده های نمایه سازی موتورهای جسـتجو عمل می کنـد. برنامه ای است که وب را به روشـی خودکار مطابق مجموعه خط مشـی هایی مرور می کنـد. با فهرستی از نشانی های یو.آر.ال (1) شروع و هر صفحه شناسایی شده توسط یو.آر.ال را ذخیره می کند تمامی فرامتن ها را در صفحه می یابد (به عنوان مثال پیونـدهایی به دیگر ،صـفحات ،تصـاویر نوشـته ها یا شـیوه نامه ها، ویـدئوها و غیره) و آن ها را به فهرست یو.آر.ال ها می افزایـد تا به طور مسلسل دیده شوند.

پـارامتر های فنی بر هویت و رفتار خزنـده (دامنه، عمق سـرعت فیلترهای تحریم، و غیره) تأثیر می گـذارد اما از آن جا که فنون وب بسـیار پیچیده هسـتند و خیلی به سرعت توسعه می یابند، خزنده با بسیاری موانع فنی مواجه می شود که مانع آن از جمع آوری تمامی عوامل یک وب گاه یا حتی یک صـفحه وب می شود. بنابراین آرشیوهای وب اغلب ناقص هستند کتابخانه ملی فرانسه از خزنده منبع باز Heritrix که توسط مؤسسات عضو کنسرسیوم حفاظت بین المللی اینترنت توسعه یافته است، (2) استفاده می کند

از آن جا که ممکن نیست جامعیت را هـدف قرار داد یا گزینش دستی وب گاه ها را به عهده گرفت، کتابخانه ملی فرانسه درنظر گرفته اسـت دو روش جمع آوری مکمل را برای رویارویی با چالش های واسپاری حقوقی وب ترکیب کند:

• جمع آوری خودکار بخش عمده ای از وب گاه های فرانسوی خزش های وسـیع در جهت ارائه که تصویری کلی از هزاران فایل از تعداد بسـیار زیـادی از وب گاه ها انجام می شود. به عنوان مثال، برای خزش وسـیع 2010 که هنوز در زمـان نگارش این مقاله هم فعال ،اسـت کتابخانه ملی فرانسه بیشترین حـدّ 10000 یو.آر.ال. را برای هر 1/6 میلیون وب گاه جمع آوری می کنـد خزنـده ها محتوی را بدون هیچ تمایزی بین محتوای دانشگاهی سازمانی تجاری یا مبتذل جمع آوری می کنند. این روش به راستی به

1- URL

2- International Internet Preservation Consortium (IIPC): http://netpreserve.org [last accessed on 2010-06-15].

Heritrix crawler: http://crawler .archive.org [last accessed on 2010-06-15] .

اسلوب واسپاری حقوقی است (به عنوان مثال فرض را بر این قرار نمی دهد آن چه را که مورد علاقه پژوهش گران در 100 سال آینده است، شناسایی کند). با این حال آرشیوهایی که با این روش ایجاد می شوند بسیار سطحی هستند؛ محتوای عمیق یا تحولات وبگاه را حفظ نمی کنند.

• خزش های متمرکز (1): جمع آوری گزینشی خزش های وسیع را کامل می کند. کتابداران موضوعی وب گاه هایی را برای چنین خزش های متمرکز منطبق با طرح های توسعه مجموعه (همکاری با دیگر کتابخانه ها و پژوهشگران نیز ممکن است) انتخاب می کنند خزش های متمرکز می توانند واقعه محور (انتخابات فرانسه در سال های 2002، 2004 ، 2007 ، و 2009) یا موضوعی (خاطرات و بلاگ های شخصی توسعه ،پایدار عمل گرایی وب و غیره) باشند خزش های متمرکز ساخت آرشیوهای کاملتر و بیشتر را از تعداد محدودی از وب گاه ها ممکن می سازند.

امروزه آرشیوهای وب کتابخانه ملی فرانسه یا آرشیوهای وب حوزه ملی فرانسه شامل 12/5 میلیارد .یو.آر.ال. می شود و 145 ترابایت فضای دیسک اشغال می کند. تاریخ قدیمی ترین صفحات وب به 1996 باز می گردد و این صفحات به لطف آرشیو ملی به دست آمدند که سازمانی غیر انتفاعی است و در نظر دارد یک کتابخانه اینترنتی بسازد و پیشگام آرشیو وب است آخرین تاریخ صفحات وب به چند ساعت قبل باز می گردد.

کشف منبع: خدمات و ابزارهای دسترسی برای کاربران نهایی

تعیین جای خدمت و محدودیت های دسترسی

دسترسی به این آرشیوها مانند دسترسی به اسناد فیزیکی واقع در ساختمان های کتابخانه نیست. وب گاه های گردآوری شده در فهرست کتابخانه ثبت نمی شوند زیرا مجموعه بسیار بزرگ و بسیار ناهمگن است؛ غیر ممکن خواهد بود فهرستی جامع از وب گاه های آرشیو شده ایجاد کرد؛ عنوان ها و محتوای مفصل شان را دانست در ،عوض کتابخانه ملی فرانسه فرایندهای نمایه سازی خودکار را ساخته است تا دسترسی سریع به محتوای جمع آوری شده را ممکن سازد. هر فایلی تاریخ می خورد و توصیف می شود تا فقط اطلاعات ضروری (مکان اصلی در وب شکل، اندازه، تعیین جا در آرشیوها، و غیره) را جمع آوری کند. این فرایند نمایه سازی وب گاه های آرشیو شده را قادر می سازد مجدد در محیط انتشاراتی شان نقش داشته باشند و آن ها را با کلیک روی پیوندها درست شبیه وب در حال فعالیت اما در یک محتوای تاریخی و تاریخ خورده مرور کنند.

از آوریل 2008 آرشیوهای وب برای کاربران مجاز در اتاق های مطالعه کتابخانه پژوهشی، در مکان های متفاوت کتابخانه ملی فرانسه (طبقه همکف در کتابخانه فرانسوا میتران و بخش های مجموعه های تخصصی در ریشولیو، لووا، اوپرا آرسنال و ژان – ویلار در آوینیون) (2) قابل دسترس هستند. گر چه آرشیوها اساساً شامل وب گاه های با دسترسی عمومی و رایگان هستند، این محدودیت وضع

1- focused crawls

2- Rez-de-jardin level at François-Mitterrand, and special collections depart-ments at Richelieu, Louvois, Opé ra, Arsenal, and Jean-Vilar in Avignon

شده است تا از فراهم آوری های حقوقی که به تمامی مجموعه های واسپاری و میراث حقوقی مربوط می شوند و در نظر دارند کاملاً به کپی رایت و قوانین شخصی احترام بگذارند، تبعیت کند.

برای دسترسی ،توافقی کاربران نهایی باید بالاتر از 18 سال باشند و مدرکی جهت نیاز دانشگاهی فعالیت های حرفه ای یا شخصی پژوهشی شان برای دسترسی به این آرشیوها ارائه کنند (اکنون کتابخانه ملی فرانسه آخرین و تنها منبع این نوع سند است زیرا تنها کتابخانه ای است که این خدمت را در فرانسه ارائه می کند). کارت های خوانندگان توسط سرویس راهنمای خوانندگان در کتابخانه های فرانسوا – میتران یا ریشولیو پس از مصاحبه شخصی با یک کتابدار صادر می شود بر اساس نیازهای کاربران این مصاحبه تعیین می کند کاربران اجازه ورود به یک یا چند بخش را دارند و دوره اعتبار کارت (3 روز 15 روز سالانه) چقدر است

آرشیوهای وب همراه با تمامی خدمات الکترونیکی کتابخانه (وب گاه های اطلاعاتی، تسهیلات رزرو، فهرست ها، دسترسی به اینترنت) و منابع الکترونیکی (کتاب ها و تصاویر دیجیتالی، نشریات الکترونیکی پایگاه داده های آنلاین، لوح های فشرده، بوک مارک ها) روی 350 کامپیوتر واقع در اتاق های مطالعه کتابخانه پژوهشی در دسترس هستند این کامپیوتر ها در دسترس عموم هستند اما کاربران برای یک جا در کتابخانه و استفاده از کامپیوترها نیاز دارند از قبل جا ذخیره کنند

ابزارهای جستجو و دیدن

آرشیوهای وب از طریق برنامه کاربردی اختصاصی ای که مجموعه «آرشیوهای اینترنت» (1) نامیده می شود، قابل دسترس می شوند واژه «آرشیوها» استفاده می شود تا تأکید کند که این مجموعه ها جامع نیستند. برنامه کاربردی توسط یک نوار نارنجی با حروف www و یک نشانگر ماوس روی آن ارائه می شود تا نشان دهد گر چه صفحات وب در یک جعبه قرار می گیرند هنوز قابل کلیک هستند کتابخانه ملی فرانسه توسعه یک همانندی دیداری و یک علامت خاص (2) را انتخاب کرده است تا آرشیوهای وب را بیش تر قابل دیدن کند.

عکس

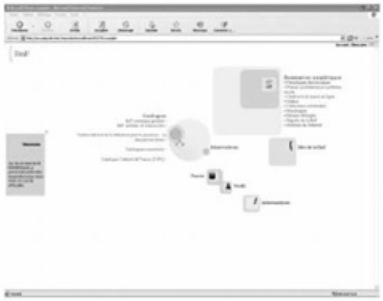
شـــده اســـت تا از فراهم‌آوری‌های حقوقی که به تمامی مجموعه‌های واسپاری و میراث حقوقی مربوط می‌شوند و در نظر دارند کاملاً به کپی‌رایت و قوانین شخصی احترام بگذارند، تبعیت کند.

برای دسترسی توافقی، کاربران نهایی باید بالاتر از ۱۸ سال باشند و مدرکی جهت نیاز دانشگاهی، فعالیت‌های حرفه‌ای، یا شخصی پژوهشی‌شان برای دسترسی به این آرشیوها ارائه کنند (اکنون کتابخانه ملی فرانسه آخرین و تنها منبع این نوع سند است، زیرا تنها کتابخانه‌ای است که این خدمت را در فرانسه ارائه می‌کند). کارت‌های خوانندگان توسط سرویس راهنمای خوانندگان در کتابخانه‌های فرانسوا ـ میتران یا ریشولیو پس از مصاحبه شخصی با یک کتابدار صادر می‌شود. براساس نیازهای کاربران، این مصاحبه تعییـــن می‌کند کاربران اجـــازه ورود به یک یا چند بخش را دارند و دوره اعتبار کارت (۳ روز، ۱۵ روز، سالانه) چقدر است.

آرشیوهای وب همراه با تمامی خدمات الکترونیکی کتابخانه (وبگاه‌های اطلاعاتی، تسهیلات رزرو، فهرست‌ها، دسترسی به اینترنت) و منابع الکترونیکی (کتاب‌ها و تصاویر دیجیتالی، نشریات الکترونیکی، پایگاه‌داده‌هـــای آنلاین، لوح‌های فشـــرده، بوک‌مارک‌ها) روی ۳۵۰ کامپیوتـــر واقع در اتاق‌های مطالعه کتابخانه پژوهشی در دسترس هستند. این کامپیوترها در دسترس عموم هستند، اما کاربران برای یک جا در کتابخانه و استفاده از کامپیوترها نیاز دارند از قبل جا ذخیره کنند.

ابزارهای جستجو و دیدن

آرشیوهای وب از طریق برنامه کاربردی اختصاصی‌ای که مجموعه «آرشیوهای اینترنت»[1] نامیده می‌شود، قابل دسترس می‌شوند. واژه «آرشیوها» استفاده می‌شود تا تأکید کند که این مجموعه‌ها جامع نیستند. برنامه کاربردی توسط یک نوار نارنجی با حروف www و یک نشانگر ماوس روی آن ارائه می‌شود تا نشان دهد گرچه صفحات وب در یک جعبه قرار می‌گیرند، هنوز قابل کلیک هستند. کتابخانه ملی فرانسه توسعه یک همانندی دیداری و یک علامت خاص[2] را انتخاب کرده است تا آرشیوهای وب را بیشتر قابل دیدن کند.



تصویر ۱. صفحه اصلی خدمات و منابع الکترونیکی کتابخانه پژوهشی کتابخانه ملی فرانسه

1. Archives de l›Internet
2. logo

تصویر 1. صفحه اصلی خدمات و منابع الکترونیکی کتابخانه پژوهشی کتابخانه ملی فرانسه

ص: 292

Archives de l›Internet -1

logo -2

برای مرور آرشیوها کتابخانه ملی فرانسه سه ابزار متفاوت ارائه می کند:

- جستجو با یو آر.ال.
- جستجو با کلیدواژه
- مجموعه های مشخصی که در نظر دارند کشف مجموعه های اختصاصی را تسهیل کنند.

هر سه ابزار در برنامه کاربردی واحدی به نام آرشیو اینترنت منسجم شده است (تصویر 2)

عکس

برای مرور آرشیوها، کتابخانه ملی فرانسه سه ابزار متفاوت ارائه می‌کند:

- جستجو با یو.آر.ال.،
- جستجو با کلیدواژه،
- مجموعه‌های مشخصی که در نظر دارند کشف مجموعه‌های اختصاصی را تسهیل کنند.

هر سه ابزار در برنامه کاربردی واحدی به‌نام «آرشیو اینترنت» منسجم شده است (تصویر ۲).

تصویر ۲. صفحه خانگی عملکرد آرشیو اینترنت

- جستجوی یو.آر.ال.[1] کاربران را قادر می‌سازد آرشیو یک وبگاه، یک صفحه وب یا حتی فایل را با وارد کردن جایگاه اصلی اینترنتی‌اش جستجو کنند. به‌عنوان مثال، جستجوی http://www.lemonde.fr ۶۱۷ نتیجه می‌دهد (از ۱۵ ژوئن ۲۰۱۰) که در یک نگاه تقویمی از ۱۹۹۶ تا ۲۰۱۰ نمایش داده می‌شوند (نتایج می‌توانند به تاریخ‌های خاصی محدود شوند). هر تاریخ قابل کلیک است و دسترسی به صفحه‌خانگی روزنامه لوموند می‌دهد که در آن تاریخ قابل دسترس بوده است. این نوع جستجو زمان جستجوی توسعه یک وبگاه و مقایسه ویرایش‌های آن بسیار مفید است، اما درست شبیه جستجوی یک شماره بازیابی در یک فهرست، کاربران باید یو.آر.ال. وبگاه یا صفحه وبی که می‌خواهند جستجو کنند، بدانند (به‌عنوان مثال، از یک کتابشناسی) یا باید قادر باشند آن را پیدا کنند (با استفاده از راهنماها یا پیوندهای بیرونی[2] که به وبگاه‌های همکار ارجاع می‌دهند). کاربران نیز ممکن است وب در حال فعالیت و آرشیوهای وب را هم‌زمان در حال استفاده از دو مرورگر با ظاهری متفاوت، به‌نوبت مرور کنند. در حال حاضر، جستجوی یو.آر.ال. تنها روش جستجوی تمام آرشیوهاست.
- جستجوی کلیدواژه‌ای شبیه یک موتور جستجوی سنتی کار می‌کند: کاربران را قادر می‌سازد اسناد متنی را که شامل یک یا چند واژه است، جستجو نمایند. به‌لطف گزینه‌های جستجوی پیشرفته نیز کاربران ممکن است عین یک عبارت یا جمله را جستجو کنند یا جستجو را به یک وبگاه خاص

۱. URL کوته‌نوشت Uniform Resource Locator است و اشاره به جایگاه اینترنتی صفحه وب دارد.

2. out-links

تصویر 2. صفحه خانگی عملکرد آرشیو اینترنت

جستجوی یو.آر.ال. (1) کاربران را قادر می سازد آرشیو یک وبگاه یک صفحه وب یا حتی فایل را با وارد کردن جایگاه اصلی اینترنتی اش جستجو کنند. به عنوان مثال جستجوی http://www.lemonde fr 617 نتیجه می دهد (از 15 ژوئن (2010) که در یک نگاه تقویمی از 1996 تا 2010 نمایش داده می شوند (نتایج می توانند به تاریخ های خاصی محدود شوند). هر تاریخ قابل کلیک است و دسترسی به صفحه خانگی روزنامه لوموند می دهد که در آن تاریخ قابل دسترس بوده است. این نوع جستجو زمان جستجوی توسعه یک وبگاه و مقایسه ویرایش های آن بسیار مفید است اما درست شبیه جستجوی یک شماره بازیابی در یک فهرست، کاربران باید یو.آر.ال. وب گاه یا

صفحه وبی که می خواهند جستجو کنند، بدانند (به عنوان مثال از یک کتابشناسی) یا باید قادر باشند آن را پیدا کنند (با استفاده از راهنماها یا پیوندهای بیرونی (2) که به وب گـاه هـای همکار ارجاع می دهنـد). کاربران نیز ممکن است وب در حال فعالیت و آرشـیوهای وب را همزمان در حال اسـتفاده از دو مرورگر با ظاهری متفاوت به نوبت مرور کنند. در حال حاضـر، جستجوی یو. آر.ال. تنها روش جستجوی تمام آرشیوهاست.

- جسـتجوی کلید واژه ای شبیه یک موتور جسـتجوی سنتی کار می کند: کاربران را قادر می سازد اسناد متنی را که شامل یک یا چند واژه است جسـتجو نمایند. به لطف گزینه های جسـتجوی پیشرفته نیز کاربران ممکن است عین یک عبارت یا جمله را جستجو کنند یا جستجو را به یک وب گاه خاص

1- URL کوته نوشت Uniform Resource Locator است و اشاره به جایگاه اینترنتی صفحه وب دارد.

2- out-links

محدود کنند. این جستجو هنوز آزمایشی است و فقط پنج درصد آرشیو های وب کتابخانه ملی فرانسه را شامل می شود نمایه سازی تمام متن میلیاردها فایل بینظم با به حساب آوردن نسخه های تکراری و انسجام موقت که آن ها را به یکدیگر پیوند می دهد، یک چالش فنی است که چندین طرح پژوهشی بین المللی در تلاش اند تا آن را برطرف کنند اما هنوز موفق نشده اند.

• علاوه بر گزینه های جستجو مهمترین کارآمدی های این عملکرد عبارت اند از: توانایی نمایش صفحات وب با جمع آوری تعدادی عوامل که ممکن است در تاریخ های متفاوتی یا دست کم در زمان های مختلف آرشیو شده باشند و در نتیجه دوباره یک وبگاه مصنوع ایجاد کنند.

• توانایی کلیک روی پیوندها با در نظر گرفتن انسجام موقت (به عنوان مثال اگر ما به وبگاه الیزه ریاست جمهوری فرانسه (1) در 7 می 2007 نگاه کنیم و روی یک پیوند کلیک کنیم تا وبگاه دولت را ببینیم انتظار داریم ویرایش ماه می 2007 را مشاهده کنیم).

این کارآمدی ها و جستجوی یو.آر.ال توسط ماشین وی یک منبع باز (2) پشتیبانی می شوند و توسط آرشیو اینترنت با مشارکت ICP توسعه می یابند جستجوی کلیدواژه ای بر پایه نرم افزار ناچ واکس منبع باز (3) (ناچ با الحاقات آرشیو وب) بنا شده است. همچنین این نرم افزار که توسط آرشیو اینترنت و گردهمایی کتابخانه های ملی نوردیک (4) توسعه یافته ابزاری است برای نمایه سازی و جستجوی آرشیوهای وب با استفاده از موتور جستجوی ناچ و الحاقات که آرشیوهای وب را جستجو می کند. کتابخانه ملی فرانسه جهت ساخت یک برنامه کاربردی بر اساس این ابزارها آن ها را در ابزارها و فرایندهای توسعه اش یکپارچه و سفارشی کرده و به کارآمدی های کوچک تر گسترش داده است (به عنوان مثال، اجرای طرحی برای سؤالات مداوم یو.آر.ال.).

مجموعه های ویژه

کتابداران موضوعی کتابخانه ملی فرانسه جهت جبران ابهام مجموعه ها و فقدان ابزارهای دسترسی (که هنوز توسعه می یابند) و نیز آگاهی در مورد مجموعه های موضوعی غنی، با همکاری پژوهشگران، مسیرهای هدایت شده (به معنی واقعی کلمه: تورهای هدایت شده یا مجموعه های ویژه در فرانسه) را ساخته اند که مجموعه ای از صفحات مصور و منظم ویرایشی است که شامل پیوندهای مستقیم به آرشیوهای وب می شود مقدمه ای ارائه می کند جستجو را شبیه سازی می نماید و ایده هایی در مورد دیگر جستجوهای ممکن می دهد از سال 2008 سه مجموعه ویژه منتشر شده است:

• کلیک کن و رأی بده: وب انتخاباتی: گزینشی از وب گاه ها از انواع تهیه کننده (سازمان ها، کاندیداها، حامیان، ،تماشاچیان ،افراد و غیره) از مبارزات انتخاباتی ،2002 ، 2004 و 2007 تصویر 3)؛

1- the French Presidency Elysé e website

2- The open source Wayback Machine, http://archive-access. sourceforge.net/projects/ wayback/ [last accessed on 2010-06-15].

3- NutchWAX (Nutch Web Archive eXtensions), http://archive-access. sourceforge.net/ projects/nutch/ [[last accessed on 2010-06-15].

• در مورد خود نوشـتن در وب خـاطرات شخصـی و ادبی: در نظر دارد نتایـج انتقـال از کاغـذ را به وب و روشـی که بلاگ ها نوشـته های شخصی ادبی و انتقادی را تغییر داده اند، بررسی کند؛

• عمل گرایی وب (1): نشان می دهد چطور پس از سال ها وب توسط عمل گراها به عنوان ابزار انتشار و ،ارتباط ابزاری که قبول مسئولیت را تشویق می کند و مکانی برای ،بحث مشارکت سازماندهی و عمل استفاده شده است.

عکس

- در مورد خود نوشتن در وب: خاطرات شخصی و ادبی: در نظر دارد نتایج انتقال از کاغذ را به وب و روشی که بلاگ‌ها نوشته‌های شخصی، ادبی، و انتقادی را تغییر داده‌اند، بررسی کند؛
- عمل‌گرایی وب[1]: نشـــان می‌دهد چطور پس از سال‌ها، وب توسط عمل‌گراها به‌عنوان ابزار انتشار و ارتباط، ابزاری که قبول مســـئولیت را تشویق می‌کند و مکانی برای بحث، مشارکت، سازماندهی، و عمل استفاده شده است.

تصویر ۳. مجموعه ویژه «کلیک کن و رأی بده: وب انتخاباتی»

زمانی که این مقاله نوشته می‌شود، سه مجموعه ویژه‌تر در دست تهیه است: توسعه پایدار، فیلم‌های غیرحرفه‌ای، و سفرنامه‌نویسی.

به‌عنوان یک خدمت مکمّل، کاربران اجازه دارند صفحات وب را چاپ و کپی کنند و نمونه‌های متن را در یک فایل نوت‌پد[2] درج کنند، محتوای نوت‌پد را چاپ کنند یا آن را از طریق ایمیل بفرستند. به‌علت محدودیت‌های حقوقی، در حال حاضر تسهیلاتی وجود ندارد که کاربران را قادر سازد از صفحه عکس بگیرند یا عناصر وب را از آرشیوها بردارند.

طرح‌های توسعه آتی، نمایه‌سازی تمام‌متن و جستجو به‌علاوه دسترسی ناپیوسته وبگاه از کتابخانه‌های منطقه‌ای را همراه با تسهیم مأموریت واسپاری حقوقی با کتابخانه ملی فرانسه در اطراف کشور دربرمی‌گیرد.

1. Web activism
2. note pad

تصویر 3. مجموعه ویژه «کلیک کن و رأی بده: وب انتخاباتی»

زمانی که این مقاله نوشته می شود سه مجموعه ویژه تر در دست تهیه است توسعه پایدار فیلم های غیر حرفه ای، و سفرنامه نویسی.

به عنوان یک خدمت مکّمل، کاربران اجازه دارند صفحات وب را چاپ و کپی کنند و نمونه های متن را در یک فایل نوت پد (2) درج کنند محتوای نوت پد را چاپ کنند یا آن را از طریق ایمیل بفرستند. به علت محدودیت های حقوقی در حال حاضر تسهیلاتی وجود ندارد که کاربران را قادر سازد از صفحه عکس بگیرند یا عناصر وب را از آرشیوها بردارند.

طرح های توسعه آتی نمایه سازی تمام متن و جستجو به علاوه دسترسی ناپیوسته وب گاه از کتابخانه های منطقه ای را همراه با تسهیم مأموریت و اسپاری حقوقی با کتابخانه ملی فرانسه در اطراف کشور در بر می گیرد.

Web activism -1

note pad -2

کاربرد اطلاعات و ارقام

ارزش آرشیوهای وب و علاقه به آن ها فقط با گذشت زمان ثابت می شود پس از اینکه منابع وب از وب ناپدید شدند؛ این بخشی از فلسفه میراث کتابخانه ملی فرانسه و سنتی نیست که انتظار یک بازگشت سرمایه کوتاه مدت را داشته باشد؛ رسالتش فهم اهمیت زمان است.

با این حال نخستین ارزیابی کاربران و کاربرد آرشیوهای وب مرحله ای مهم در اجرای واسپاری حقوقی اینترنت است. اثبات سودمندی عمومی و علمی مجموعه های جمع آوری شده و نیز توسعه و تحلیل نخستین آمارهای کاربرد ما را قادر خواهد ساخت توان بالقوه این مجموعه ها و نیز محدودیت های شان را اندازه گیری کنیم این تحلیل ها در رویارویی با انتظارات پژوهشگران برای توسعه مجموعه و ابزار هر دو مفید خواهند بود

تحلیل کمّی

ابزاری تحلیلی به نام AWStats ایجاد شده است تا ترافیک همزمان عملکرد را تحلیل کند. شبیه ماشین وی بک و ناچ واکس درون سازمانی سفارشی ساخته شده است تا بین استفاده عمومی توسط خوانندگان، استفاده مرجع توسط کتابداران در هنگام ورود یا در میز ،مرجع و استفاده حرفه ای توسط کتابداران موضوعی کتابخانه ملی فرانسه تمایز ایجاد نماید

عکس

کاربرد: اطلاعات و ارقام

ارزش آرشیوهای وب و علاقه به آنها فقط با گذشت زمان ثابت می‌شود، پس از اینکه منابع وب از وب ناپدید شـــدند؛ این بخشی از فلسفه میراث کتابخانه ملی فرانسه و سنتی نیست که انتظار یک بازگشت سرمایه کوتاه‌مدت را داشته باشد؛ رسالتش فهم اهمیت زمان است.

با این حال، نخســـتین ارزیابی کاربران و کاربرد آرشـــیوهای وب مرحله‌ای مهم در اجرای واسپاری حقوقی اینترنت است. اثبات سودمندی عمومی و علمی مجموعه‌های جمع‌آوری‌شده و نیز توسعه و تحلیل نخســـتین آمارهای کاربرد، ما را قادر خواهد ساخت توان بالقوه این مجموعه‌ها و نیز محدودیت‌هایشان را اندازه‌گیری کنیم. این تحلیل‌ها در رویارویی با انتظارات پژوهشگران برای توسعه مجموعه و ابزار هر دو مفید خواهند بود.

تحلیل کمّی

ابزاری تحلیلی به‌نام AWStats ایجاد شـــده است تا ترافیک هم‌زمان عملکرد را تحلیل کند. شبیه ماشین وی‌بک و ناچ‌واکس، درون‌سازمانی سفارشی ساخته شده است تا بین استفاده عمومی توسط خوانندگان، استفاده مرجع توســـط کتابداران در هنگام ورود یا در میز مرجع، و اســـتفاده حرفه‌ای توسط کتابداران موضوعی کتابخانه ملی فرانسه تمایز ایجاد نماید.

جدول ۱. شاخص‌های کاربرد اصلی در سال ۲۰۰۸ و ۲۰۰۹

	۲۰۰۸	۲۰۰۹
میانگین تعداد جلسه‌ها در هر ماه	۳۵	۱۰٦
تعداد کل جلسه‌ها	۳۱٦	۱۲۷۵
تعداد کل صفحات مشاهده‌شده	۳۵۸۹۱	۹۰۰٦۳

جدول ۲. کمیّت‌های جامع در سال ۲۰۰۹

	مشاهده‌ها	مشاهده‌کنندگان	جلسه‌ها (مشاهده‌های بیشتر از ۵ دقیقه)	جلسه‌های طولانی (مشاهده‌های بیشتر از ۱ ساعت)	صفحات مشاهده شده
عمومی	۹۰٤	٤۸۱	۳۳۸	۷۱	۱٤۰٤۵
مرجع	۲۱٦	۱۱۸	۹٤	۲۳	٦۲٤۲
حرفه‌ای	۱۹٤۵	۵۵۲	۸٤۳	۲۳۹	٦۹۷۷٦
کل	۳۰٦۵	۱۲۳۵	۱۲۷۵	۳۳۳	۹۰۰٦۳

جدول 1. شاخص های کاربرد اصلی در سال 2008 و 2009

جدول 2. کمیت های جامع در سال 2009

ص: 296

دو نتیجه ای که توجه به کتابخانه ملی فرانسه را جلب می کند عبارت اند از:

• تعداد کل جلسه ها بین سال 2008 و 2009 سه بار افزایش یافته است (بعداً خواهیم دید که آموزش اصلی و ابتکار عمل های اطلاعاتی برای خوانندگان و کتابداران مرجع توسعه یافته است)؛

• تعداد فزاینده ای از کاربران نهایی وجود دارند که آرشیوها را برای پژوهش عمیق استفاده می کنند. میانگین جلسه در فوریه ،2009، 13 دقیقه طول کشید و در دسامبر 2009 تا 30 دقیقه افزایش یافت بین فوریه و سپتامبر 2009 بیش از دو جلسه وجود نداشت که بیش از یک ساعت یا بیشتر طول بکشد. سپس، تعداد هر جلسه در اکتبر 2009 به 7 جلسه در نوامبر به 32 جلسه، در دسامبر به 22 ،جلسه و در ژانویه 2010 به 23 جلسه افزایش یافت این جلسه ها نشان می دهد که اجرای پژوهش وسیع روی آرشیوهای وب آغاز می شود با نگاهی به فهرست بیش ترین صفحات مشاهده شده به نظر می رسد این طرح ها بیش از همه توسط پژوهشگران علوم اجتماعی و علوم سیاسی اجرا می شود.

تحلیل کیفی

در تکمیل تحلیل آماری و فهرستی جامع از یو.آر.ال.ها و کلید واژه های جست وجو شده، کتابخانه ملی فرانسه یک لاگ داخلی و مشترک مورد استفاده کتابدارانی ایجاد کرد که کاربران را به مناسبت های مختلف از جمله مصاحبه پذیرش میز ،مرجع درخواست ملاقات ،شخصی، نمایش به تهیه کنندگان وب گاه ها و غیره ملاقات می کنند از آوریل 2008 کتابداران 34 کاربر نهایی را با وارد کردن اطلاعاتی همچون تاریخ/ساعت گروه و نام کتابدار شکل در خواست (در محل ،تلفنی، ایمیل، ...) ، نام نوع ،کارت موضوع پژوهش، یادداشت ها /سؤال ها نظرات/ اظهارات، ثبت نام کردند (تصویر 4).

عکس

دو نتیجه‌ای‌که توجه به کتابخانه ملی فرانسه را جلب می‌کند عبارت‌اند از:

- تعداد کل جلسه‌ها بین سال ۲۰۰۸ و ۲۰۰۹ سه بار افزایش یافته است (بعداً خواهیم دید که آموزش اصلی و ابتکار عمل‌های اطلاعاتی برای خوانندگان و کتابداران مرجع توسعه یافته است)؛
- تعداد فزاینده‌ای از کاربران نهایی وجود دارند که آرشیوها را برای پژوهش عمیق استفاده می‌کنند. میانگین جلسه در فوریه ۲۰۰۹، ۱۳ دقیقه طول کشید و در دسامبر ۲۰۰۹ تا ۳۰ دقیقه افزایش یافت. بین فوریه و سپتامبر ۲۰۰۹، بیش از دو جلسه وجود نداشت که بیش از یک ساعت یا بیشتر طول بکشد. سپس، تعداد هر جلسه در اکتبر ۲۰۰۹ به ۷ جلسه، در نوامبر به ۳۲ جلسه، در دسامبر به ۲۲ جلسه، و در ژانویه ۲۰۱۰ به ۲۳ جلسه افزایش یافت. این جلسه‌ها نشان می‌دهد که اجرای پژوهش وسیع روی آرشیوهای وب آغاز می‌شود. با نگاهی به فهرست بیشترین صفحات مشاهده شده، به‌نظر می‌رسد این طرح‌ها بیش از همه توسط پژوهشگران علوم اجتماعی و علوم سیاسی اجرا می‌شود.

تحلیل کیفی

در تکمیل تحلیل آماری و فهرستی جامع از یو.آر.ال.ها و کلیدواژه‌های جست‌وجوشده، کتابخانه ملی فرانسه یک لاگ داخلی و مشترک مورد استفاده کتابدارانی ایجاد کرد که کاربران را به مناسبت‌های مختلف از جمله مصاحبه پذیرش، میز مرجع، درخواست ملاقات شخصی، نمایش به تهیه‌کنندگان وبگاه‌ها، و غیره ملاقات می‌کنند. از آوریل ۲۰۰۸، کتابداران ۳٤ کاربر نهایی را با وارد کردن اطلاعاتی همچون تاریخ/ساعت، گروه و نام کتابدار، شکل درخواست (در محل، تلفنی، ایمیل، ...)، نام، نوع کارت، موضوع پژوهش، یادداشت‌ها/سؤال‌ها/نظرات/اظهارات، ثبت‌نام کردند (تصویر ٤).

سؤالات	پاسخ‌ها
تاریخ و ساعت	می/ ژانویه / اوت ۲۰۰۹
بخش یا خدمت و نام کارگزار	کریستین ژنن
پشتیبانی از درخواست	در محل
نام کوچک و نام خانوادگی خواننده	
عنوان دسترسی	نمایش‌ها
موضوع پژوهش / رشته	- ۲۹ می (۱۷ ساعت و نیم): الیزابت لگرو، بلاگر و عضو گروه یار APA http://2009sediments.wordpress.com/ - ۲۰ ژانویه (۱٦ ساعت): مارتین سونه، پژوهشگر CNRS، نویسنده و بلاگر www.martinesonnet.fr/blogwp - ۱۱ اوت (۱۷ ساعت):سیلور مرسیه که یادداشتی را در بلاگش نوشته است: http://www.bibliobsession.net/2009/09/17/archives-de-liternet-demadez-votre-ticket-pour-la-posterite/ - ۲۱ اوت (۱۷ ساعت): اوریان دزینی که در Paris XIII تدریس می‌کند و نویسنده رساله‌ای در مورد روزنامه‌های خصوصی آنلاین وی به‌ویژه علاقه‌مند به مسیر هدایت‌شده است و از یک همکار بانک سالن در مورد این موضوع سؤال کرده است که آن را به‌سوی من ارجاع داده است.
مشاهدات	

عکس

سؤالات	پاسخ‌ها
تاریخ و ساعت	۲٤ نوامبر ۲۰۰۹
بخش یا خدمت و نام کارگزار	SOL کریستوف تربویی
پشتیبانی از درخواست	در یک محل رسمی
نام کوچک و نام خانوادگی خواننده	فابین گرفه
عنوان دسترسی	کارت سالانه (۲۰۰۹/۱۱/۲٤ تحویل شد)
موضوع پژوهش / رشته	مبارزه انتخاباتی اروپائی سال ۲۰۰۹ در اینترنت
مشاهدات	

سؤالات	پاسخ‌ها
تاریخ و ساعت	۲۰۰۹/۸/۱۹
بخش یا خدمت و نام کارگزار	SOL
پشتیبانی از درخواست	در محل
نام کوچک و نام خانوادگی خواننده	کارول دافینی
عنوان دسترسی	کارت سالانه
موضوع پژوهش / رشته	آرشیوهای وب (کارآموز محافظه‌کار)
مشاهدات	

تصویر ٤. لاگ کاربران نهایی آرشیوهای وب

این ابزار ساده، کتابخانه ملی فرانسه را قادر ساخت مجموعه‌هایی را که مورد درخواست هستند، شناسایی کند؛ این مجموعه‌ها در حال حاضر در حوزه علوم اجتماعی و علوم سیاسی هستند. روزی که کتابخانه ملی فرانسه خدمت را عرضه کرد، یک دانشجوی کارشناسی ارشد که در مورد «اینترنت و انتخابات ریاست جمهوری ۲۰۰۷» کار می‌کرد، درخواست دسترسی کرد. یک دانشجوی دکترای زبان‌شناسی که در مورد تحلیل سخنرانی‌های کاندیداهای زن کار می‌کرد از روم، ایتالیا آمده بود تا در مورد بلاگ رهبر کمونیست، ماری ـ ژرژ بوفه[1] که بعد از انتخابات ریاست جمهوری بسته شده بود، کار کند. فعالیت وب نیکلا سارکوزی، حزب سوسیالیست، انتخابات اروپایی ۲۰۰۹، استفاده از ویدئوها در فعالیت انتخاباتی، کارتون‌ها و کاریکاتورهای سیاسی، نمونه‌هایی از پژوهش در این حوزه‌ها هستند.

پژوهش‌های دیگر شامل موضوعاتی بوده است همچون جستجوی اطلاعات در اتحادیه اروپا، آرشیوهای وزارت بوم‌شناسی و موجودیت‌های نامتمرکزش، وب‌سایت‌های نویسندگان غیرحرفه‌ای، وب‌سایت‌های مدیریت استرس، وب‌سایت‌های شخصی و سؤالاتی درباره اینکه چرا و چگونه

1. Marie-George Buffet

تصویر 4. لاگ کاربران نهایی آرشیوهای وب

این ابزار ساده کتابخانه ملی فرانسه را قادر ساخت مجموعه هایی را که مورد درخواست هستند شناسایی کند؛ این مجموعه ها در حال حاضر در حوزه علوم اجتماعی و علوم سیاسی هستند. روزی که کتابخانه ملی فرانسه خدمت را عرضه کرد یک دانشجوی کارشناسی ارشد که در مورد «اینترنت و انتخابات ریاست جمهوری 2007» کار می کرد درخواست دسترسی .کرد یک دانشجوی دکترای زبان شناسی که

در مورد تحلیل سـخنرانی های کاندیـداهای زن کار می کرد از ،روم ایتالیا آمده بود تا در مورد بلاگ رهبر کمونیست ماری – ژرژ بوفه (1) که بعد از انتخابات ریاست جمهوری بسـته شده بود، کار کند. فعالیت وب نیکلا سارکوزی، حزب سوسیالیست انتخابات اروپایی 2009 استفاده از ویدئوها در فعالیت انتخاباتی، کارتون ها و کاریکاتورهای ،سیاسی نمونه هایی از پژوهش در این حوزه ها هستند.

پژوهش های دیگر شامل موضوعاتی بوده است همچون جسـتجوی اطلاعات در اتحادیه اروپا آرشیوهای وزارت بوم شناسی و موجودیت های نامتمرکزش وب سایت های نویسندگان غیر حرفه ای وب سایت های مدیریت استرس وب سایت های شخصی و سؤالاتی درباره این که چرا و چگونه

Marie-George Buffet -1

وب سایت ها جمع آوری می شوند و قوانین رقابتی تا در یک مورد حقوقی کمک کند.

نظر سنجی ها

نخستین نظر سنجی از اکتبر 2006 تا ژوئن 2007 در محتوای یک دوره درسی در مقطع کارشناسی ارشد به نام «اینترنت در طول فعالیت» انجام شد و توسط کتابخانه ملی فرانسه و یک پژوهشگر دانشگاه در رسانه اجتماعی به طور هماهنگ سازماندهی شد. 17 جلسه مشاهده و 5 مصاحبه انجام شد تا نیازهای کاربران به ابزارها و کارآمدی ها و نگرش شان به یک نوع رسانه جدید شناسایی شوند (برای آزمون نهایی آن ها، دانشجویان مجبور شدند مقاله ای بنویسند در موضوعی که مربوط به منبع وب در حال فعالیت و محتوای آرشیو شده اش بود).

دومین نظر سنجی برای نوامبر 2010 برنامه ریزی می شود این نظر سنجی کاربران فعلی و بالقوه را در نظر دارد و قصد دارد نیازهای شان را به محتوای مجموعه تعریف و ابزار توسعه کاربرد را شناسایی کند (جدول 3).

عکس

وب‌سایت‌ها جمع‌آوری می‌شوند و قوانین رقابتی تا در یک مورد حقوقی کمک کند.

نظرسنجی‌ها

نخستین نظرسنجی از اکتبر ۲۰۰۶ تا ژوئن ۲۰۰۷ در محتوای یک دوره درسی در مقطع کارشناسی ارشد به‌نام «اینترنت در طول فعالیت» انجام شد و توسط کتابخانه ملی فرانسه و یک پژوهشگر دانشگاه در رسانه اجتماعی به‌طور هماهنگ سازماندهی شد. ۱۷ جلسه مشاهده و ۵ مصاحبه انجام شد تا نیازهای کاربران به ابزارها و کارآمدی‌ها و نگرش‌شان به یک نوع رسانه جدید شناسایی شوند (برای آزمون نهایی آنها، دانشجویان مجبور شدند مقاله‌ای بنویسند در موضوعی که مربوط به منبع وب در حال فعالیت و محتوای آرشیوشده‌اش بود).

دومین نظرسنجی برای نوامبر ۲۰۱۰ برنامه‌ریزی می‌شود. این نظرسنجی کاربران فعلی و بالقوه را در نظر دارد و قصد دارد نیازهایشان را به محتوای مجموعه تعریف و ابزار توسعه کاربرد را شناسایی کند (جدول ۳).

جدول ۳. چارچوبی برای نظرسنجی بعدی کاربرد آرشیو وب در نوامبر ۲۰۱۰

کاربران فعلی آرشیو وب	کاربران بالقوه آرشیو وب
۵-۸ مصاحبه	۵-۸ مصاحبه برای هر گروه هدف: • دانشمندان در دیگر حوزه‌های پژوهشی (هنرهای دیجیتال، علوم انسانی، پژوهشگران رسانه‌ای علاقه‌مند به خود وب، و غیره)، • یک گروه وسیع‌تر علاقه‌مند به حافظه وب، • وکلا و متخصصان اطلاعات.
سؤالاتی درباره: • خط‌مشی توسعه مجموعه: گزینش، تناوب و کیفیت جمع‌آوری • کاربران هدف و انواع استفاده برای توسعه • علاقه و موضوعاتی برای مجموعه‌های ویژه • اطلاعات و متادیتاهای ارائه‌شده در نمایش نتایج و نمایش یک صفحه	سؤالاتی درباره: • آگاهی از آرشیوهای وب • علاقه شناسایی‌شده به خود حوزه پژوهشی • انواع اطلاعات جستجوشده اگر مورد نیاز هستند • ضروریات کاربرد واقعی

مشارکت: راهبردهایی برای تأسیس یک انجمن کتابداران و رسیدن به کاربران نهایی

معرفی آرشیوهای وب در کتابخانه به‌معنی ساخت یک مفهوم کلی از مالکیت این نوع جدید مجموعه توسط هر دو کارکنان و کاربران نهایی است.

مجموعه‌ها به‌شیوه‌های مختلفی باعث سرخوردگی می‌شوند (ذخیره‌های وب ناقص هستند و برخی وبگاه‌ها در آرشیوها هم‌زمان ظاهر نمی‌شوند). مانند سرخوردگی عملکرد جستجو: برای کارکنان، به‌ویژه، فقدان یک فهرست و این حقیقت که کتابداران هر کار نظام‌مند توصیفی را در مورد وبگاه‌ها انجام

جدول 3. چارچوبی برای نظر سنجی بعدی کاربرد آرشیو وب در نوامبر 2010

مشارکت: راهبردهایی برای تأسیس یک انجمن کتابداران و رسیدن به کاربران نهایی

معرفی آرشیوهای وب در کتابخانه به معنی ساخت یک مفهوم کلی از مالکیت این نوع جدید مجموعه توسط هر دو کارکنان و کاربران نهایی است.

مجموعه ها به شـیوه های مختلفی باعث سـرخوردگی می شونـد (ذخیره های وب ناقص هسـتند و برخی وب گاه ها در آرشـیوها همزمان ظاهر نمی شوند). مانند سـرخوردگی عملکرد جسـتجو: برای کارکنان، به ویژه، فقدان یک فهرست و این حقیقت که کتابداران هر کار نظام مند توصیفی را در مورد وب گاه ها انجام

ص: 299

نمی دهند برای آنان سخت کرده است که آرشیوهای وب را به عنوان مجموعه های میراث با ارزش مورد توجه قرار دهند (چقدر احتمال دارد یک کتابدار بتواند در مورد یک مجموعه ساخته شده توسط نرم افزار خزش گر و نه توسط خودش احساس مثبتی داشته باشد؟)

بنابراین چالش این است که کیفیت های متمایز آرشیوهای وب – به خصوص این حقیقت که آن ها تنها گواه واحد باقیمانده از تغییرات اساسی جامعه ما هستند که در 15 سال گذشته تجربه شده اند و گذارشان از آنالوگ به دیجیتال –را مشخص کرد، اما همچنین مشخص کردن الگوهای این مجموعه که در واقع آن را کمی شبیه به مجموعه های منظمی می کند که کتابداران یاد گرفته اند در دست بگیرند: مربوط می شود به شعار "دیجیتال متفاوت نیست" – و واژه را منتشر می کند.

گفته می شود که تجارت معمول آرشیو وب ساختن ابزار اولیه استفاده از ارتباط استاندارد، سازماندهی و راهبردهای بازاریابی است.

راهبردهای به حساب آوردن آرشیوهای وب به عنوان بخشی از کار روزانه کتابخانه

سازمان. مدیریت آرشیوهای وب نباید فقط به عنوان یک فعالیت فنی در نظام رسمی کتابخانه دیده شود بیش تر سازمان هایی که به آرشیو کردن وب به عنوان یک مورد فنی نگاه کرده اند، موضوعی تحت آی.تی. که نیاز به مهندس و رهبری توسعه نرم افزار دارد به سختی می توانند کتابداران را درگیر ارتقای آرشیوها کنند برای این که کار آرشیو کردن وب کاری معمول به حساب آید کتابخانه ملی فرانسه تصمیم گرفت فعالیت هایش را در قدیمی ترین واحد تولید ،کتابخانه بخش واسپاری حقوقی به انجام برساند. وب گاه ها اکنون با منابع چاپی واسپاری حقوقی (کتاب ها و نشریات ادواری) در واحدی که سال 2008 به نام واحد «واسپاری حقوقی دیجیتال» (1) ایجاد شد، مدیریت می شوند. این واحد شامل گروهی پنج نفره می شود که فعالیت ها را با مشارکت متخصصان آی.تی. از یک طرف و متخصصان مجموعه از سویی دیگر اجرا می کنند این اجرا در بخشی بزرگ و قدیمی کمک بسیاری کرده است تا متخصصان و فعالیت های آرشیو کردن وب را در رسالت اصلی و تاریخ کتابخانه ملی فرانسه بگنجاند. به عبارت دیگر، این نوع سازماندهی به اتصال قدیم به جدید کمک کرده است و نمایش می دهد که راهبردهای نمونه برداری با مقیاس بزرگ برای حجم های عظیم داده های دیجیتال متولد شده خیلی متفاوت نیست از راهی که تاریخ سنت واسپاری حقوقی منابع چاپی فرانسوی را طی پنج سال گذشته شکل داده است. این نگرش بخشی از یک تلاش وسیع تر توسط کتابخانه ملی فرانسه است تا سازماندهی مهارت های کارکنان را با تغییر دیجیتال انطباق دهد. (2)

شبکه سازی. ساخت شبکه ای از متخصصان مجموعه موضوعی از سراسر کتابخانه مرحله سرنوشت ساز دیگری جهت تشویق مشارکت های فعال و پذیرش آرشیوهای وب به عنوان مجموعه ای

1- digital legal deposit

2- The Human Face of Digital Preservation: Organizational and Staff Challenges, and Initiatives at the‘ Bibliothèque nationale de France’, Emmanuelle Bermès, Louise Faudet, iPres 2009, [http://www.escholarship. org/uc/cdl_ipres09 [last accessed on 2010-06-15].

توسط کتابداران و مدیران بود. در سال 2005 این شبکه به عنوان گروهی متشکل از 20 پیشاهنگ شروع به کار کرد. تا سال 2010 حدود 80 کتابدار به گونه ای درگیر جمع آوری گزینشی بودند، به عنوان مثال، گزینش وبگاه کنترل کیفی و ارتقاء منابع و خدمات برای عموم نخست برای مدیران بخش مجموعه آگاهی ایجاد شد. در نتیجه، بیش تر بخش های موضوعی کتابداران موضوعی را برای طرح آرشیو کردن وب اختصاص دادند اکنون در هر حوزه اصلی (هنر ادبیات فلسفه علوم و غیره) از جمله موضوعاتی که مجموعه های نادر را در بر می گیرد (مانند نقشه ها یا موسیقی) یک کارگزار واسپاری حقوقی وب (1) وجود دارد، فردی که از آموزش خاص بهره مند بوده و دانش اساسی آرشیو کردن وب را کسب کرده است طوری که بتواند در حوزه تخصصی از آن استفاده کند برخی کتابداران موضوعی کتابخانه ملی فرانسه اکنون کتاب ها و وب گاه ها را به دست می آورند از سال 2008 به بعد تعداد کارکنان این شبکه کارگزاران مرحله واسپاری حقوقی وب که به خصوص به عنوان کتابداران مرجع در نظر گرفته شده بودند، افزایش یافته است. این جا هدف توسعه دانش و علاقه به آرشیوهای وب بود در هر کس که با عامه مردم خواه در اتاق های مطالعه یا به صورت آنلاین در ارتباط است.

ابزارهایی برای اشاعه داخلی. کتابخانه ملی فرانسه خوش شانس است که قادر است از مجموعه منابع مهمی برای ارتقای اطلاعات و مهارت ها به طور بین المللی سود برد: آموزش کارکنان، گفتگوهای داخلی ماهانه ،کارکنان و کنفرانس هایی در وقت ،ناهار یک مجله چاپی داخلی ماهانه و البته اینترانت همگی، قبل از شروع برنامه آرشیو کردن وب در دسترس بودند. چالش این بود که از تمامی این منابع ارتباطی استفاده و آن ها را به مسیری هوشمند هدایت کند تا کارکنان در مورد آرشیوهای وب در زمان و با روش مناسب بشنوند به عنوان ،مثال کتابخانه ملی فرانسه درست پایان مبارزات ریاست جمهوری سال 2007 یک نمایش اساسی داخلی از مجموعه وب گاه های منتخب برگزار کرد. همچنین مقالات مجله داخلی یا مقالاتی که در اینترانت کتابخانه ملی فرانسه بود یک طرح ارتباطی ای را دنبال کرد که منافع و علائق کلی کتابخانه را به حساب می آورد به عنوان مثال زمانی که کتابخانه مجموعه فعالیت هایی را که در راستای بالا بردن آگاهی نسبت به توسعه پایدار بود عملی کرد (و کارکنان را تشویق کرد که رفتاری مسئولیت پذیر در این حوزه به عهده بگیرند) گروه آرشیو کردن وب ارتباطات داخلی اش را روی مجموعه های مرتبط با توسعه پایدار نیز افزایش داد در یک کلام طرح ارتباط داخلی جهت ارتقاء آرشیوها هرگز به عنوان یک فرایند مستقل دیده نشد. ما تمامی فرصت های ممکن را به کار گرفتیم تا آرشیوهای وب رل در برنامه کلی کار کتابخانه بگنجانیم بیش از آن که تلاش کنیم این برنامه کار را با تحمیلی دیدن آن بر هم زنیم.

راهبردهایی برای رسیدن به کاربران نهایی

بحث ها: «آیا اصلا آرشیوهای وب استفاده می شوند؟» «چرا منابع را صرف این داده های بیهوده کنیم در حالی که ما هیچ تضمینی نداریم که آیا این نوع ماده مورد علاقه عموم قرار خواهد گرفت؟» «چرا تلاش های مان را روی فراهم آوری یا دیجیتالی کردن موادی که می دانیم بدون شک میراثی برای نسل های

1- Web legal deposit correspondent = correspondant du dé pôt legal du Web

آینده ،هستند متمرکز نمی کنیم؟» – این ها سؤالات معمولی است که خواه مربوط به مدیریت باشند یا مربوط به رسانه ها بیشتر اوقات نیاز به پاسخ دارند. از جهتی با نگاهی به گذشته تاریخ کتابخانه، چنین سؤال هایی جدید نیست در طول ،زمان هر رسانه جدیدی سؤالات مشابهی برانگیخته است. دلیل آن این است که این امر زمان می برد – زمانی برای چیزهایی که از عموم فضای کسب و کار رایج ناپدید می شوند – قبل از این که پژوهشگران یا افراد متوجه شوند آن ها را از دست می دهند آنان به این مستندات نیاز دارند تا تاریخ جامعه را تشریح کنند.

روی هم رفته، موقعیت ویژه ای برای یک سازمان میراثی است تا برنامه آرشیو کردن وب خود را اجرا کند: گرایش به این وجود دارد که بروندادهای ملموس را با مفید و با ارزش نشان دادن آن ها نمایش دهد در حالی که همزمان کاربرد نمی تواند بلافاصله توسعه یابد هنوز برای این کار خیلی زود است زیرا بیش تر کاربران آرشیوهای وب هنوز دارند متولد می شوند راهبردهای کتابخانه ملی فرانسه جهت کنترل این موقعیت دو جانبه است از یک سو با ایجاد ابزارها و آمارهایی که کتابخانه را قادر می سازد توسعه کاربرد را به روشی مشابه همان گونه که برای مجموعه های عادی انتظار می رفت نمایش دهد (دیجیتالی یا غیر دیجیتالی؛ به بخش بالا مراجعه کنید)، از سوی دیگر توجه عموم را به بحث در مورد قابلیت رؤیت از خارج از کتابخانه جلب کند.

ایجاد یک بحث عمومی. گر چه تعداد زیادی از مردم اکنون از آرشیوهای وب استفاده نمی کنند، در واقع بیشتر آنان علاقه به این مسأله را زمانی که از آن ها در این مورد سؤال شد، نشان داده اند. دلیل این امر این است که اینترنت زندگی های شخصی و عمومی افراد را تحت تأثیر قرار داده است بنابراین هر کس چیزی دارد که درباره آن بگوید. نخستین واکنش به طور معمول این است: «من هرگز در مورد آن فکر نکردم اما اکنون که شما به من می گویید تصور می کنم مهم است این حافظه ماست و هر روز از یاد می رود» ارتباطات کتابخانه ملی فرانسه این مردم را برای توسعه حمایت عمومی از برنامه آرشیو کردن وب کتابخانه هدف قرار می دهد – آنان که ممکن است امروز برای دسترسی به مجموعه ها سروکله شان در کتابخانه پیدا نشود، اما این آگاهی را به دست آوردند که کاری است که باید انجام شود.

زنجیره کنفرانس حافظه های وب (1) به عنوان یک گردهمایی عمومی جهت فراهم آوردن این نوع حمایت و قابلیت رؤیت طراحی و در مارچ 2010 اجرا شد هر کنفرانسی یک نصف روز طول می کشد و سه نوع شرکت کننده را که در ترکیب مخاطبان منعکس می شوند نیز گردهم می آورد: پژوهشگران ،وب ناشران و مسئولان همه کسانی که از آن ها درخواست می شود به موضوع مشابه بپردازند به عنوان مثال، عمل گرایی و سیاست وب خاطرات ،وب حمایت از داده های شخصی در مقابل وسعت فضای عمومی، و غیره. کتابخانه ملی فرانسه نزدیک به 100 شرکت کننده را برای هر یک از این رویدادها گردهم آورده است و پوشش رسانه ای مناسبی را (بیشتر بر روی وب) برای دو نشست نخست دریافت کرده است. هدف این است که آرشیو کردن وب یک موضوع بحث عمومی در خارج از کتابخانه اما به همراه کتابخانه شود. این شکل ارتقا (که بازخورد عالی ای را برای ساخت مجموعه و خدمات به پژوهشگران نیز به دنبال دارد)

1- The ‘Mé moires du Web’ (Web memories in French) conference cycle

فعالیت های ارتقاء سازمان یافته مستقیم تر را برای کاربران نهایی امروز کتابخانه ملی فرانسه تکمیل می کند.

بیشتر کاربران آرشیوهای وب هنوز وجود ندارند آنان به احتمال با نسل های آینده ای می آیند که به دنیا آمده اند و با وب به دنیا خواهند آمد و آنان که از آن استفاده می کنند یا آن را به عنوان ابزار اصلی اطلاعاتی و ارتباطی شان استفاده خواهند کرد برای چنین مجموعه ای باید بپذیریم که فقط زمان است که به ما خواهد گفت آیا ما انتخاب درست را انجام می دهیم مدیریت دسترسی و گزینش نیز به معنی مدیریت خطرات است. ما مجبوریم آن چه که امروز غیر منتظره و ناخواسته است اما فردا ممکن است مورد علاقه ،باشد بگیریم و ارتقا دهیم. همچنین مجبوریم کاربران و کاربرد فردا را در نظر بگیریم. این مسأله جدید نیست: چالشی است که مؤسسه میراثی قدیمی با آن آشناست. در این میان مؤسسات هنوز به راهبردهای میان مدت برای ایجاد اطمینان و حس مالکیت جمعی نسبت به این مجموعه های جدید در میان کارکنان و سرمایه داران شان نیاز دارند سازماندهی اشاعه داخلی ارتباطات و تلاش های آموزشی همگی در کل، کلید توسعه جوامع جدید آماده انطباق با مجموعه دیجیتالی هستند.

منابع

Aubry, Sara (2008): 'Les archives de l'Internet: un nouveau service de la BnF', Documentaliste Sciences . 1
.de l'Information, 45(4), pp. 12-13

Bermè s, Emmanuelle and Gildas Illien (2009): 'Metrics and Strategies for Web Heritage Management . 2
.and Preservation', IFLA, http://www.ifa.org/ fles /hq/papers/ifa75/92- bermes-en.pdf

Bermè s, Emmanuelle and Louise Faudet (2009): 'The Human Face of Digital Preservation: . 3
Organizational and Staff Challenges, and Initiatives at the Bibliothè que nationale de France', iPres,
.http://www.escholarship.org/uc/cdl-ipres09

Illien, Gildas (2008a): 'L'archivage d'Internet, un dé f pour les dé cideurs et les bibliothé caires: scé narios . 4
d'organisation et d'é valuation, l'expé rience du consortium IIPC et de la BnF, IFLA, http://archive.ifa.org/IV/
.ifa74/papers/107-Illien-fr.pdf

Illien, Gildas (2008b): 'Re-Inventing Collection Development Policy in the Age of Web Archiving: the . 5
Experience of the BnF, LIBER Annual Conference, http://www.ku.edu. tr/ku/images/LIBER/ LIBER-
.ILLIEN-2008.ppt

Lasfargues, France, Clé ment Oury and Bert Wendland (2008): ʻLegal deposit of the French Web: . 6
.harvesting strategies for a national domain', IWAW, http://iwaw.net/08/ IWAW2008-Lasfargues.pdf

چکیده

WCT نوعی منبع بـاز برای مـدیریت آرشـیو وب گزینشـی است که به عنوان پروژه ای مشترک بین کتابخانه ملی زلانـدنو و کتابخانه بریتانیا توسـعه یافته است. از ژانویه 2007 WCT در کتابخـانه ملی زلانـدنو به طور روزانه اسـتفاده می.شود این مقاله نخستین سال آرشـیو وب گزینشـی ما را با نرم افزار جدید توصیف می کند. کتابخانه ملی زلاندنو، فواید WCT را توسـعه داده و قصد دارد برنامه گزینشی گردآوری با WCT را برای آینده ای قابل پیش بینی ادامه دهد.

ص: 304

یک سال آرشیو وب گزینشی

اشاره

یک سال آرشیو وب گزینشی با (WCT (1 در کتابخانه ملی زلاندنو

نوشته: گوردون پنیتر، (2) سوزانا جو (3)، وانیتا لا لا (4)، گیلیان لی (5)

ترجمه: احترام السادات کیان مهر (6)

مقدمه

WCT نرم افزاری است که منابع برخط گزینش شده، گردآوری شده، و ارزیابی کیفی شده را که توسط کاربران گروهی در محیط کتابخانه ای به کار گرفته می شود پشتیبانی میکند از این نرم افزار، در مواقعی برای گردآوری وب گزینشی استفاده می شود که یک متخصص موضوعی قسمت هایی از یک وبگاه یا کل آن را معمولا در ارتباط با یک ناحیه موضوعی تمرکز یافته یا یک رویداد مهم شناسایی کرده باشد.

نرم افزار، به عنوان پروژه ای گروهی بین کتابخانه ملی زلاندنو و کتابخانه بریتانیا توسعه داده شده است و زیر نظر کنسرسیوم فراهم آوری اینترنت بین المللی مدیریت می شود.

WCT ، نرم افزار منبع باز است و از طریق وبگاه http://webcurator.sf.net برای استفاده جامعه آرشیو وب بین المللی آزادانه قابل دسترس است.

کتابخانه ملی زلاندنو (از این پس کتابخانه) از ژانویه 2007 از WCT به عنوان پایه برنامه آرشیو وب

1- Web Curator Tool

2- Gardon Panynter

3- Susanna Joe

4- Vanita Lala

5- Gillian Lee

6- کارشناس ارشد سازمان اسناد و کتابخانه ملی

گزینشی استفاده می کند.

طی سال اول ویرایش جدید نرم افزار توسعه داده شد و به طور شگرفی، افزایش و بهبود کیفیت فعالیت های گردآوری و نیز گردآوری منابع و بی به طور خودکار را میسر ساخته است.

این مقاله تجربه ما را در استفاده از WCT در محیط کار تعریف می کند

بخش بعدی ،مقاله پیش زمینه فعالیت های بهره برداری وب کتابخانه و WCT را در اختیار می گذارد بخش های زیر تجربه ما را با نرم افزار جمع بندی می کند و سرانجام با توصیف یک رویداد اجرا شده گردآوری با نرم افزار خاتمه می یابد.

2.آرشیو وب گزینشی در کتابخانه ملی زلاندنو

2-1.انگیزه

کتابخانه ملی زلاندنو حکمی قانونی و مسئولیتی اجتماعی برای محافظت تاریخ فرهنگی و اجتماعی زلاندنو به شکل ،کتاب ،روزنامه عکس وبگاه بلاگ ،یوتیوب، و ویدئو بر عهده دارد.

علاوه بر آن میراث مستند کتابخانه ملی زلاندنو فقط به صورت برخط قابل دسترس است از نظر این محتوای دسترس پذیر با ارزش است اما ناپایداری فقدان مالکیت شفاف و صحیح و طبیعت پویای آن چالش های مهمی برای هر مؤسسه ای است که برای اندوختن و نگهداری آن ها تلاش می کند.

وب گاه ها ،WCT برای حل همین مشکلات ارتقا یافت به این ترتیب که به برخی موسسه ها اجازه می داد هر نوع سند برخط شامل صفحه های وب ، وب گاه ها و وب نوشت ها و سایر اشکال جاری، شامل صفحه های HTML ، تصاویر PDF و مدارک word مانند محتوای چند رسانه ای نظیر فایل های دیداری و شنیداری را دریافت کنند این انواع با مراقبت های ،ممکن طوری هماهنگ می شوند که یکپارچگی و اصالت شان حفظ شود

برخط منفعت عمومی از نگهداری و حفاظت دراز مدت میراث برخط زلاندنو غیر قابل محاسبه است. تاریخ اجتماعی برخط ما و اکثر تاریخ دولت و سازمان برای حفظ کردن قابل امکان خواهد بود و برای محققان تاریخ دانان و شهروندان آینده زلاندنو نیز امکان پذیر خواهد بود آن ها خواهند توانست به گذشته مستند مدارک دیجیتال ما در راهی مشابه که زلاندنویی ها تا امروز به واژه های چاپ شده که برای ما از نسل های گذشته باقی مانده است، نگاه کنند.

2-2.تاریخچه درو/ هاروستینگ

کتابخانه ملی ،زلاندنو از سال 1999 تا پایان سال 2006 ، برنامه عملی آرشیو وب گزینشی را اجرا کرده است ،کتابخانه برای نگهداری منابع از نرم افزار کپیر (1) استفاده کرده و منابع را بر اساس مارک (2) هدایت و پایگاه داده ها را فراهم می کند. نرم افزار HTTrack کتابخانه را با یک پس افت منابع گردآوری شده که

1- HTTrack Website Copier Software

2- MARC (Machine readable cataloging)

قابلیت آرشیو شدن برای نگهداری دراز مدت را ندارد رها می کند. در حال حاضر، برای تبدیل این منابع به شکل مناسب قابل آرشیو کردن برنامه انتقال داده در حال اجراست.

2-3. نرم افزار گردآوری وب

WCT، نوعی نمودار کـاری گردآوری را پشتیبـانی می کنـد که شامـل وظـایف تخصصـی است: انتخاب منبع برخط، جست وجو، اجازه نگهـداری ،آن قابـل دسترس بودن ، قابـل توصـیف بودن، تعیین دامنه آن، و دسـته بنـدی هـا، نمودار کردن گردآوری وب یـا یک سـری از گردآوری های وب، اجرا کردن گردآوری ها، انجام دادن مرور کیفیت و تأیید یا تکذیب کردن منابع گردآوری شده و ذخیره منابع تأیید شده در یک مخزن دیجیتال یا آرشیو.

عکس

قابلیت آرشیو شدن، برای نگهداری دراز مدت را ندارد، رها می‌کند. در حال حاضر، برای تبدیل این منابع به شکل مناسب قابل آرشیو کردن برنامه انتقال داده در حال اجراست.

۲ -۳. نرم‌افزار گردآوری وب

WCT، نوعی نمودار کاری گردآوری را پشتیبانی می‌کند که شامل وظایف تخصصی است: انتخاب منبع برخط، جست‌وجو، اجازه نگهداری آن، قابل دسترس بودن ، قابل توصیف بودن، تعیین دامنه آن، و دسته‌بندی‌ها، نمودار کردن گردآوری وب یا یک سری از گردآوری‌های وب، اجرا کردن گردآوری‌ها، انجام دادن مرور کیفیت و تأیید یا تکذیب کردن منابع گردآوری شده و ذخیره منابع تأیید شد ه در یک مخزن دیجیتال یا آرشیو.

شکل ۱. فهرست اصلی نرم‌افزار را نشان می‌دهد.

شکل ۱. فهرست WCT

بیشتر فعالیت‌های آرشیو وب به‌طور عمده‌ای بر تخصص‌های فنی اپراتورهای گردآوری تکیه دارد . از سوی دیگر، WCT گردآوری را مسئولیت کاربران و متخصصان موضوعی (ترجیحاً از مهندسان و مدیران سیستم) از طریق آسان و روان نمودن، به‌طور خودکار، جزئیات فنی گردآوری وب می‌داند.

شکل 1. فهرست اصلی نرم افزار را نشان می دهد.

شکل 1. فهرست WCT

بیش تر فعالیت های آرشیو وب به طور عمده ای بر تخصص های فنی اپراتورهای گردآوری تکیه دارد. از سوی دیگر، WCT گردآوری را مسئولیت کاربران و متخصصان موضوعی (ترجیحاً از مهندسان و مدیران سیستم) از طریق آسان و روان نمودن، به طور خودکار، جزئیات فنی گردآوری وب می داند.

نرم افزار برای کار مطمئن و تأثیر گذار در محیط جایی که کارکنان پشتیبانی فنی می توانند آن را نگهداری کنند طراحی شده است.

WCT نرم افزار منبع بـاز است و به طور آزادانه از طریق وبگاه http://webcurato.sf.net تحت گواهینامه Apache public License قابل دسترس است.

وبگاه دسترسـی کاربران به ،راهنماها دست نامه ها فهرست های ،پستی screenshots ، سؤال ها، مسـتندات فنی و اداری کد منبع ردیابی اشتباهات و صفحه پروژه sourceforg را فراهم می کند.

(پینتر و میسون (1)، نرم افزار و توسعه اش را با جزئیات بیش تری در مقاله خود در کنفرانس لیانسا (2) در سال 2006 توصیف می کنند).

2.4. اعضا و منابع

در کتابخانه ملی زلاندنو ، WCT نرم افزار اولیه و مسئولیت کتابدارن نشر الکترونیکی در کتابخانه «الکساندر تورن بل» (3) است. در سال 2007 معادل دو و نیم برابر انتخابگر ،الکترونیکی به طور مسـتقیم از این نرم افزار اسـتفاده کردنـد و نیز همه گزینش ها را مدیریت نموده و گردآوری و مرور کیفی نمودند .

به هر حال نرم افزار با خط مشی های کتابخانه گردش کار ارتباطات راه دور و پشتیبانی سرویس ها یکپارچه و جامع است و بر گروه گسترده تر و بیش تری از اعضا تأثیر می گذارد.

برای مثال، سـخت افزار و نرم افزار توسط سـرویس های فنی حفظ می شونـد و توسط میز پشتیبانی مـدیریت می شونـد، رده بنـدی توسـط سرویس های محتوا هدایت می شود و آرشیو دیجیتال توسط کتابخانه ملی دیجیتال حفظ و نگهداری می شود.

WCT برای یکپارچه شدن با سیستم های کاری ،موجود به طور محکم و استوار – و تا حد ممکن – طراحی شده است.

برای اسـتقرار وضـعیت از سـخت افزار اسـتاندارد کتابخانه (سرورهای سان اسپارک (4)) سیسـتم های عملیاتی (سولاریس (5))، پایگاه اطلاعاتی (اوراکل (6)) ، وب سرویس ها (Apache HTTP Server and Tomcat)، و خدمات ثبت کاربر (راهنمای الکترونیکی نوول (7)) استفاده می شود.

سیسـتم تولید بین دو سـرور مستقر شده ،است یکی برای مدل کور (8) و دیگری برای نرم افزار گردآوری (هـم آرایـه کردن برای گردآوری هشت منبع و بی همزمـان) و اشتراک پایگـاه اطلاعاتی موجود و سـرورهای فایل با سیسـتم های دیگر کتابخانه سیسـتم آزمون جـداگانه با ترتیب و وضع مشابه حفظ می شود.

1- Paynter and Mason

2- LIANZA Conference

3- Alexander Turnbull Library

Sun SAARC Servers –4

(Operating System (Solaris –5

Oracle –6

Novell e Directory –7

Core module –8

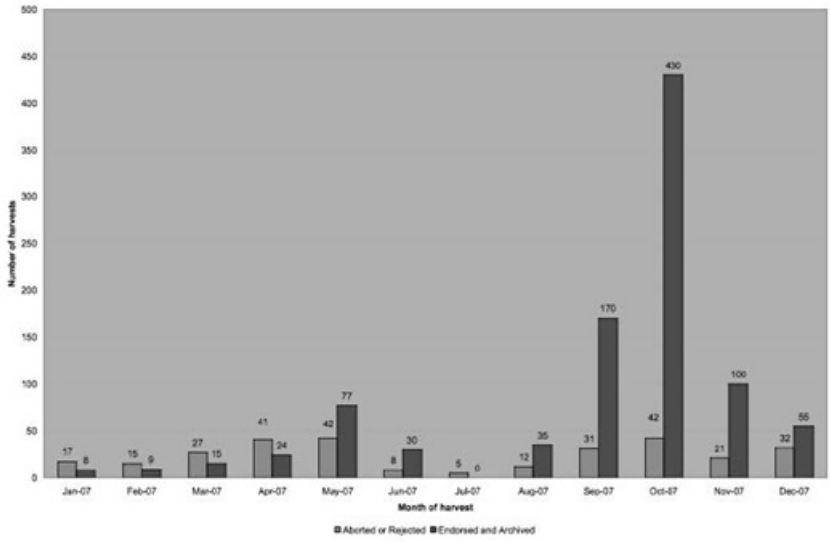
2-5.سطح گردآوری / درو

نمودار 2 تعـداد گردآوری هـای خودکار انجام شـده توسط کتابخانه در سال 2007 را نشـان می دهـد. برای هر ماه ، سـتون ارغوانی رنگ شماره نتایجی را نشان می دهد که در آرشیو تأیید نشده است. این موارد یا متوقف شدهاند (قبل از آن که بتوانند کامل شوند) یا طی مراحل کنترل کیفیت به دلیل آن که برخی جنبه های اساسـی وبگاه را کسب نکرده انـد رها شده اند سـتونهای ،قرمز تعداد نتایج موفقیت آمیزی که در آرشـیو کتابخانه دیجیتال آرشـیو شده اند نشان می دهد. در مجموع 1249 گردآوری منابع خودکار توسط سیسـتم محصول ما در سال 2007 کمک شدهاند. 953 (76 درصد) به طور موفقیت آمیزی گردآوری شده، کنترل کیفی شده و نیز آرشیو شده است.

عکس

۲ -۵. سطح گردآوری/ درو

نمودار۲، تعداد گردآوری‌های خودکار انجام شـــده توسط کتابخانه در سال ۲۰۰۷ را نشان می‌دهد. برای هر ماه ، ستون ارغوانی رنگ شماره نتایجی را نشان می‌دهد که در آرشیو تأیید نشده است. این موارد یا متوقف شده‌اند (قبل از آنکه بتوانند کامل شوند) یا طی مراحل کنترل کیفیت، به‌دلیل آنکه برخی جنبه‌های اساســی وبگاه را کسب نکرده‌اند رها شده‌اند. ستون‌های قرمز، تعداد نتایج موفقیت‌آمیزی که در آرشیو کتابخانه دیجیتال آرشیو شده‌اند نشان می‌دهد. در مجموع ۱۲٤۹ گردآوری منابع خودکار توسط سیستم محصول ما در ســـال ۲۰۰۷ کمک شده‌اند. ۹۵۳ (۷٦درصد) به‌طور موفقیت‌آمیزی گردآوری شده، کنترل کیفی شده و نیز آرشیو شده است.



نمودار ۲. سطح گردآوری گزینش شده در ۲۰۰۷

باقیماندگان ٦۹ (٦درصد) طی گردآوری متوقف شده و ۲۲٤ (۱۸درصد) رد شده‌اند، زیرا از مرحله کنترل کیفی عبور نکرده‌اند.

ســـه فاز گردآوری و برداشـــت منابع وبی به‌طور خودکار در نمودار ۲ قابل رؤیت هستند. نخستین آنها از ژانویه تا می است، یعنی زمانی که انتخابگر الکترونیکی در فعالیت‌های گردآوری از ویرایش ۱,۱ WCT استفاده کرده است. گردآوری در این مرحله کم است و در ماه می ۷۵ گردآوری موفق داشته است، یعنی زمانی که گروه در حال تجربهٔ و گردش کار و خروجی‌های مختلف ـ حتی آنهایی که هنگام کار با نرم‌افزار اولیه کشف نشده بود ـ بودند.

نمودار 2. سطح گردآوری گزینش شده در 2007

باقیماندگان 69 (6 درصد) طی گردآوری متوقف شده و 224 (18 درصد) رد شده اند، زیرا از مرحله کنترل کیفی عبور نکرده اند.

سه فاز گردآوری و برداشت منابع و بی به طور خودکار در نمودار 2 قابل رؤیت هستند. نخستین آن ها از ژانویه تا می است یعنی زمانی که انتخـاب گر الکترونیکی در فعالیت های گردآوری از ویرایش WCT 1.1 اسـتفاده کرده است. گردآوری در این مرحله کـم است و در ماه می 75 گردآوری موفق داشـته است، یعنی زمـانی که گروه در حال تجربه و گردش کار و خروجی های مختلف – حتی آن هایی که هنگام

کار با نرم افزار اولیه کشف نشده بود – بودند.

ص: 309

از ژوئن تـا اواسـط سـپتامبر، انتخـابگران الکترونیکی آزمون ویرایش WCT 1/2 را تقاضـا کردنـد. بنـابراین، مقررات فعالیت های منظم گردآوری حتی بیشتر از قبل قطع می شود.

فـاز نهـایی گردآوری در اوسط سـپتامبر شـروع می شود و افزایش شـگرفی در فعالیت گردآوری را نشان می دهـد. عوامل متعـددی در این افزایش دخیل هسـتند که شامل انتخاب ویرایش WCT 2/1 در سیستم ،تولید افزایش 50 برابری ظرفیت اینترنت کتابخانه (از 2 مگابایت در ثانیه به 100 مگا بایت در ثانیه)، شفافیت بیشتریه خروجی های گردش کار پیرامون گردآوری کردن و فهرست کردن و آغاز دو رویداد بر اساس گردآوری (در زیر توضیح داده شده).

در نتیجه، تقریباً گردآوری در نیمی از سال فقط در ماه (تعداد وب گاه ها) اکتبر انجام می شود.

نکته جـالب در این مرحله این است که اگر چه تعـداد گردآوری هـای موفق به طور چشـمگیری افزایش یـافته است تعـداد «ناتمـام» هـا و «مردود»ها افزایش نیافته است و عددی که تأیید یا رد می شود، محسوب نمی گردد.

گردآوری با WCT ویرایش 1.1.

زمانی که انتخاب گر الکترونیکی از ویرایش WCT .1.1 از ژانویه تا می استفاده کرد، از تجربه با این نرم افزار سود بردند و به فعالیت های گردآوری و نیز مشکلات مختلفی که از طریق ویرایش 2.1 و کامل شدن آن بود پی بردند

این بخش خلاصه ای از این تجربه است و توصیف بیشتر جزئیات از نویسندگان بر اساس در خواست در دسترس می باشد.

1.3.تجربه اولیه

مهم ترین تغییر ایجاد شـده توسط WCT این است که نرم افزار منـابع گردآوری شـده را در فرمت فایل ARC ذخیره می کند که از لحاظ فراهم آوری سودمند است اما برای فعالیت های کنترل کیفیت.

و این موضوع راه اساسی جدیدی را برای انتخابگران الکترونیکی برای رسیدن و نزدیک شدن به تجدید نظر کیفی را به همراه دارد.

با این شـیوه جدید فعل و انفعال، مشـکل می توان گفت که آیا خروجی های کیفی، توسط نرم افزارهای گردآوری کننده ایجاد شده یا توسط کمبودهای نرم افزارهای کنترل کیفی

برای مثال اگر یک بازبین کننـده یک صـفحه وب را در نرم افزار تورق باز کند و نشان داده شود که یک صـفحه وب (تصویر) گم شده اسـت انتخابگر الکترونیکی ابتـدا بایـد تعیین کند که آیا به طور موفق، تصویر از منابع و بی گردآوری شده یا آن را به طور موفقیت آمیزی گردآوری کرده، امـا مرورگر نمی توانـد آن را نشان دهـد خیلی زود مشـخص شـد که برای پشتیبانی اعضا در وجود این تفاوت ها تمرین های اضافی ضروری بوده است

زمـانی که کتابـداران تجربه مهمی از مرور نتایـج گردآوری نرم افزار HTTrack را کسب کردنـد، نمی دانسـتند که چگونه از عـدم موفقیت گردآوری منابع و بی در WCT گره بگشایند.

تمام مشـکلات جدید گردآوری در نسخه جدید ضبط و مشاهده شد که همان نوع اشتباه ها دوباره رخ داده است. بنابراین بخشی از تمرین حفظ و یک روند بازشناختی نیز توسعه و به WCT به طور دستی اضافه شد.

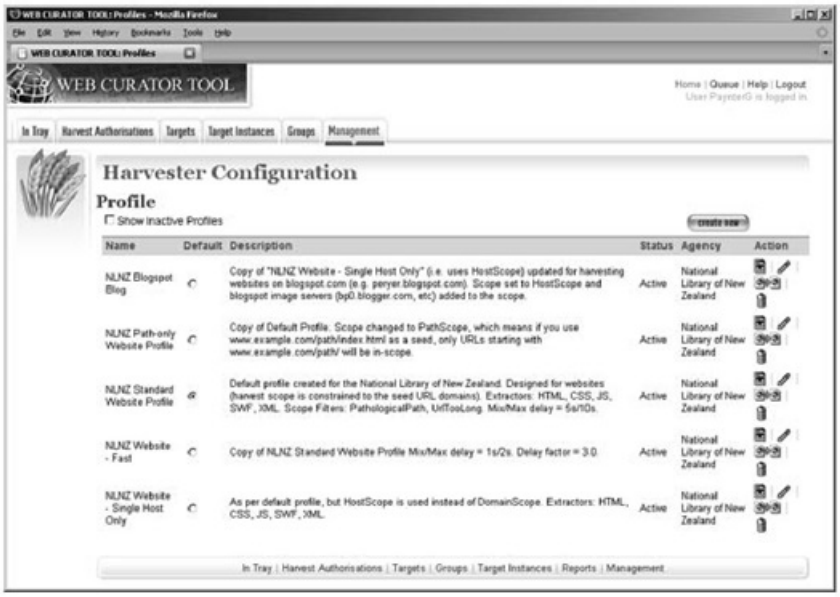
مشکل دیگر که در ماه های نخست نمایان شد این بود که توسعه سودمند پروفایل های گردآوری شده، به خصوص در مورد بلاگ ها زمان بر است توسعه شکل 3. مجموعه ای از پروفایل های گردآوری شده را نشان می دهد که سرانجام به وجود آمده اند.

عکس

زمانـــی که کتابداران تجربه مهمـــی از مرور نتایج گردآوری نرم‌افزار HTTrack را کســـب کردند، نمی‌دانستند که چگونه از عدم موفقیت گردآوری منابع وبی در WCT گره بگشایند.

تمام مشـــکلات جدید گردآوری در نسخه جدید ضبط و مشاهده شد که همان نوع اشتباه‌ها دوباره رخ داده اســـت . بنابراین، بخشی از تمرین حفظ و یک روند بازشناختی نیز توسعه و به WCT به‌طور دستی اضافه شد.

مشـــکل دیگر که در ماه‌های نخست نمایان شد این بود که توسعه سودمند پروفایل‌های گردآوری شده، به‌خصوص در مورد بلاگ‌ها، زمان بر است. توسعه شکل ۳. مجموعه‌ای از پروفایل‌های گردآوری شده را نشان می‌دهد که سرانجام به‌وجود آمده‌اند.



شکل ۳. پروفایل‌های گردآوری کنندهٔ در حال استفاده در کتابخانه در دسامبر ۲۰۰۷

۳.۲. مشکلات گردآوری

مشکلات زیر در ویرایش ۱.۱ وجود داشته و در ویرایش ۲.۱ بر طرف شده است .

کمبودهای تورق نرم‌افزار

برخی اوقات محتوای ظاهر شـــده از دست می‌رود به دلیل اینکه نرم‌افزار تورق (مرور) نمی‌تواند آن را به‌طور صحیحی منتقل کند.

شکل 3. پروفایل های گردآوری کننده در حال استفاده در کتابخانه در دسامبر 2007

3.2.مشکلات گردآوری

مشکلات زیر در ویرایش 1.1 وجود داشته و در ویرایش 2.1 بر طرف شده است

کمبودهای تورق نرم افزار

برخی اوقات محتوای ظاهر شده از دست میرود به دلیل اینکه نرم افزار تورق (مرور) نمی تواند آن را به طور صحیحی منتقل کند.

ص: 311

نسخه های میانجی کاربر

برخی نسخه های کاربر برای انتخاب گران الکترونیکی، سیر کردن سریع در پیرامون نرم افزار را مشکل می ساخت.

اشتباه «در مرحله ایست درگیر شدن»

مشکل دوباره به وجود آمده در ویرایش جدید WCT این بود که گردآوری باید همان گونه که انتظار می رفت با موفقیت به مرحله پایانی می رسید ولی سیستم به وضعیت «گردآوری شده» انتقال نمی یافت و در عوض در وضعیت توقف باقی می ماند مشکلاتی که به این شرایط هدایت می شود نتایجی از نسخه های مورد نظر که در بالا آمده و کامل شده است.

وب گاه های بزرگ

کتابخانه بریتانیا به این موضوع پی برد که نرم افزار از عهده گردآوری های بزرگ و گسترده بر نمی آید و مشکلاتی با نرم افزار گردآوری همزمان وجود دارد.

3.3. خلاصه تجربه با ویرایش 1.1

با وجود برخی ای محدودیت ها ویرایش اول WCT بهبودی وسیع و بزرگی بر نرم افزار HTTrack قبلی بود و آن مرحله مهم و با اهمیتی بود وقتی که ما اولین وب گاه های گردآوری شده را به آرشیو دیجیتال ارائه دادیم

طی سال اول WCT 1250 منابع وبی را به طور خودکار گردآوری کرد و تقریباً 950 مورد دوباره ،بررسی تأیید و به آرشیو اضافه شد.

گردآوری با WCT با ویرایش 1.2

ویرایش WCT 1.2.6، در اواسط سپتامبر 2007 به محصول گسترش داده شد و از آن در ماه های بعدی در دو رویداد گردآوری استفاده شد و سپس در پایان اکتبر با ویرایش 1.2.7 جابه جا شد.

4.1. نرم افزار گردآوری تاریخ

یکی از نواحی ای که در آن پیشرفت زیادی به وجود آمد کنترل کیفیت است. افزایش در این ناحیه به طور شگرفی تأثیر کنترل کیفیت را ثابت کرد و باعث پیشرفت بیشتر ، سریع تر، ماهرانه تر شد.

نرم افزار گردآوری تاریخ در کنترل کیفی ثابت کرده است که بسیار مفیدتر از تقویت کنترل کیفی است. این نرم افزار همه منابع گردآوری شده را که دارای یک هدف ویژه است با چکیده اطلاعات، نظیر تاریخ شروع تاریخ انتقال URL گردآوری شده و افتاده، زمان صرف شده و وضعیت جاری را فهرست می کند.

این امر بسیار مفید است به طوری که از اطلاعاتی که در روند کنترل کیفیت نیاز بود محکم تر است.

2.4.نرم افزار تورق (مرورگر)

هنوز یک تغییر ساده و تأثیر گذار برای مرور میانجی ،بود افزایش پیشنهاد ها برای دریافت دیداری یک سایت گردآوری شده از سه راه مختلف صورت می گیرد: دریافت دیداری مرحله جاری گردآوری شده در نرم افزار مرورگر دریافت دیداری سایت ،جاری یا دریافت دیداری ویرایش آرشیو شده قبلی (در آرشیو اینترنت یا آرشیو مکانی).

این مراحل در Tab دیگر با تورق ویندوز باز می شود و اجازه می دهد که مرورگر یک کپی گردآوری شده را با ویرایش دیگر سایت مقایسه کند.

این پیشرفت بزرگ باعث شد که نرم افزار تورق به ما اجازه دهد که به طور موفقیت آمیزی مرور و تورق سایت های دیگر را – که قبلاً قابل دیدن نبودند – مشاهده کنیم به هر حال، چند خروجی حل نشده باقی می ماند.

وب گاه های تحویل داده شده که از جاوا اسکریپ استفاده می کند باعث مشکلاتی می شود، به ویژه دوباره عناصر وظیفه ای مانند منوی پایین – بالا که می تواند سایت های گردآوری شده را برای سیر کردن مشکل سازد.

گردآوری سبک برگه ها هنوز می تواند مشکل باشد، گر چه نسخه های جمع آوری شده بیش از نسخه های تکراری است نسخه های برجسته ،دیگر شامل مشکلات تکرار شده URL ها با فضاها و جمع آوری تصاویر پیش زمینه جاسازی شده مطمئن.

3.4.نرم افزار هرس درخت تصمیم

نرم افزار هرس درخت تصمیم سیستم روز آمد شد اما کتابخانه نرم افزار هرس را در وب گاه های قدیمی تر استفاده نمی کند. بنابراین از این تغییرات استفاده زیادی به عمل نمی آید. به هر حال، ما با نرم افزار تجربه کرده ایم و مدتی می تواند خیلی آهسته برای سایت های بزرگ استفاده شود وظیفه جدید نرم افزار هرس ثابت کرد که می تواند برای انتخاب کردن و نگاه کردن فایل های دل خواه از منابع گردآوری شده بسیار مفید باشد

4.4.وب گاه های بزرگتر جمع آوری می شوند

در ماه های آخر سال ما متوجه شدیم که فعالیت های گردآوری ما شامل سایت های بزرگتر و بزرگ تر هستند. بزرگ ترین گردآوری ،کتابخانه به عنوان قسمتی از برنامه گردآوری در اندازه 21 گیگا بایت کامل و کنترل شد گر چه پس از کنترل کیفیت رد شده بود.

وب گاه های متعدد 10 گیگا بایتی به طور موفقیت آمیزی کنترل آرشیو و گردآوری شدند زمانی که اندازه وب گاه ها رشد کرد، انتخاب گران الکترونیکی تکیه بیش تری بر پروفایل های عادی و پروفایل های

برتر کردنـد به ویژه فیلتر هـا که به انتخاب گران الکترونیکی اجازه داد که گردآوری کننـده را از جمع آوری قسـمت های ویژه وبگاه متوقف کنند.

5.4.گسترش ذخیره دیجیتال با ارزش

نتیجه افزایش تعـداد و انـدازه وب گاه های گردآوری شده این بود که WCT به ذخیره دیجیتـال با ارزش بیش تری نیاز داشت (فضای موقت برای ذخیره گردآوری ها زمانی که کنترل کیفی شده اند و قبل از این که آرشیو شوند).

در اکتبر و نوامبر این مشـکلی ویژه و خاص بود زیرا تعداد وب گاه های انتخاب شده گردآوری شده و تأیید شده ناگهان افزایش یافت و موقتاً به طور طولانی تری فهرست شد و نوعی انبار منبع به وجود آمد. در نتیجه فضای اختصاص داده شده برای ذخیره منابع دیجیتال با ارزش که زودتر ظاهر شـده از دست رفت و ناگهان به نظر رسـید که بسـیار کم ذخیره شده است در نتیجه یک دیسک بزرگ تر و جدید برای ذخیره با ارزش منابع دیجیتال بود و ابزارهای گزارش گیری به کتابداران اجازه می دهد که از دیسک به طور موقت استفاده کنند و نمایش دهند.

این وضـعیت ،همچنین نسـخه ای از آرشـیو موقتی دیجیتال را فراهم کرد که در آن فضای دیسک با ذخیره ارزشـمند دیجیتال مشترک شـده مطمئنا تفصیل کنترل های نهایی فایده ای نداشته است.

6.4.ارتباط

ارتباط خوب بین اعضای کتابخانه های مختلف از طریق نرم افزار تحت تأثیر واقع می شود، اما شکاف های مختلفی در ارتباط از زمانی که WCT از ابتدا استفاده می شد در محصول شناسایی شد.

گر چه در ویرایش 1.2 پیشرفت هایی حاصل شد خیلی از قلم افتادگی ها، خارج از نرم افزار از طریق گزارش های دوره ای از طریق پایگاه اطلاعاتی WCT نشانی شده اند (مدرک نرم افزار، شامل یک فرهنگ لغت داده کامل است).

برای مثال، گزارش هفتگی حاصل شده از گردآوری ،منابع به فهرست نویس ها برای اعلام آن ها به وبگاه هایی که نیاز است فرسـتاده شد و تعـداد و انـدازه وب های گردآوری شـده جاداده شـده توسط انتشارات الکترونیکی کتابداران برای سـرویس های فنی اعضا پیش بینی شد. زمانی که این ها در یک شکل کامل گردآوری شدند ما امیدواریم که به آن ها یک سری گزارش های ساختار WCT را اضافه کنیم.

7.4.فهرست گردش کار و دسترسی

زمانی ممکن است از WCT برای توصیف وب گاه ها استفاده شود و آن خط مشی کتابخانه است برای توصیف مجموعه کامل کتابخانه در فهرستش و فراهم کردن یک پیوند رکورد فهرست کتابخانه به آیتم های دیجیتال نگهداری شده در محزن دیجیتال

(به هر حال انبارش های دیجیتال ،پیچیده نظیر سریال ها و وب گاه ها به طور عام قابل دسترس نخواهند

بود تا آرشیو میراث دیجیتال ملی با مخزن دیجیتال موقت ما در سال 2008 جایگزین شود).

یـک مـانع و اشـکال در جست وجو کردن فهرست کتابخـانه برای وب گـاه هاست و ناتوانی جمع آوری کامل رویـدادها از هر وبگاه که به تنهایی فهرست شده است

نشانی این نسخه کتابخانه بریتانیا توسعه یک نرم افزار میانجی وب را طراحی می کند که دسترسی به منابع گردآوری شده را بر اساس موضوع و رویداد به عنوان یک هدف اضافی برای کاربرانی که می خواهند به طور ویژه برای وب گاه ها جست و جو کنند فراهم می کند ما پیشـرفت کتابخانه بریتانیا را با علاقه بسیار زیادی دنبال خواهیم کرد.

5. گردآوری رویداد انتخابات هیئت محلی

کتابخـانه، بـا به کار بردن نرم افزار گردآوری HTTrack در سال 2007 مسـئولیت گردآوری برخی رویدادهای عظیم ورزشـی را بر عهده داشته است (جام آمریکا) (1) و انتخابات (انتخابات پارلمان ملی).

کتابخـانه دو رویـداد گردآوری را طراحی کرد که به طور اتفـاقی با اسـتقرار ویرایش WCT 1/2 همزمـان .بود گزینش هـای هیئت دولتی محلی ، رویـدادی سه سـاله است و هر قـدرت محلی در زلانـدنو به نگهـداری و هـدایت گزینش اعضـایش احتیـاج دارد کتابخـانه برنامه گردآوری رویـداد دوازده هفته ای را بـا تمرکز بر گزینش هـای هیئت محلی در سـال 2007 و شـروع شدن در سـپتامبر را به عهده گرفت. گردآوری رویداد هیئت محلی اولین گردآوری رویداد عظیمی با استفاده از WCT بود و همچنین هنوز بزرگ ترین تلاش در کتابخانه بود با 238 وب گاه انتخاب شده

این سـایت هـا شامـل وب گـاه هـا، بلاـگ ها، وب گاه های شوراهای منطقه ای ، شـهری و ناحیه ای، اخبارهای سایت ها و سایت های عمومی یا بلاگ ها با محتوای مربوط و مناسب هم گزینشـی یا دولت محلی با این حال طیف وسـیعی از وب گاه ها انتخاب شـده انـد که تفسـیرهای رسـمی و غیر رسـمی را با پراکنـدگی جغرافیایی وسـیعی نمایش می دهند کلیه سایت های انتخاب شده در محدوده وضـعیت قانونی زلاندنو بودند، بنابراین نیازی به جست وجوی دستور صریح و واضح و قانون گذاری برای به دست آوردن آن ها نداریم

سایت های انتخاب شـده برای تعیین گردآوری، اولویت بنـدی شدنـد و در آن موقع یک جـدول گردآوری برای پوشاندن کلیه تاریخ ها در مدت 16 هفته طراحی شـده بود. جـدول با تعـداد زیادی سایت انتخاب شده و با قابلیت اداره خاتمه داده شد درگیری مقتضـی اولیه برای پایان دادن منابع اعضا نیز وجود داشت. تعداد کمی از گردآوری ها یک فاصـله زمان کافی بین هر گردآوری برای 2 عضو برای کامل کردن خزش گرها و روند تکرار کیفیت را مطمئن ساخت.

عامل تأثیر گذار دیگر بر برنامه گردآوری از گردش کاری هایی ریشه می گرفت که در حال حاضـر از WCT استفاده می کنند در گردآوری های قبلی ایجاد نمودار گردآوری تکراری عظیمی از کنترل کیفیت امکان پذیر بود که می توان بعد از تکمیل همه گردآوری ها انجام داد.

America's Cup Yachting -1

به هر حال با تعداد زیادی از سایت های انتخاب شده در این دوره این رویکرد بر WCT هم غیر عملی و نشدنی است و هم سنگین و طاقت فرساست.

با گردآوری و سری زمانی، اهداف گروه ها با استفاده از سیستم های گروهی و ظیفه ای به وجود آمد. سایت های برگزیده مرتب شدند و 36 گروه با دو مقوله وبگاه ، یا – در مورد سایت های شورا– منطقه جغرافیایی تعیین شدند.

این گروه ها غیر رسمی بودند، زیرا استفاده اولیه گروه ها در این رویداد (دوره) – برای آسان تر شدن برنامه گردآوری صورت می گرفت و نه برای گروه بندی رسمی استفاده از گروه در مدیریت گردش کاری در تیم مفید بود و متناوب شدن گردآوری توسط گروه ها نیز مفید بود. همچنین نوعی تعادل فشار در بین سخت افزارهای گردآوری به وجود آورد.

اگر چه هدف، شروع گردآوری در بعد از ظهر بود، حتی با وجود هشت گردآورنده همزمان این کار ممکن نشد برخی گردآوری های برنامه ریزی شده چندین ساعت در نوبت می ماندند تا یک گردآورنده آزاد شود، اما خوشبختانه صف های طولانی برداشت موردی به وجود نیاورد و با کوچک شدن بسیاری از سایت ها صف های برداشت به سرعت ناپدید شد این سیستم با استفاده از هشت گردآورنده همزمان به خوبی با دوره های گردآوری طولانی و فشرده ، مقابله کرد

مشکلات وقتی به وجود آمد که خزش های وبگاه شوراها در اندازه های بسیار بزرگتر با آرشیو دیجیتال وب گاه هایی که منتظر فهرست نویسی پیش آرشیوی بودند همزمان شدند. نتیجه یک تفسیر مشکل بحرانی در ظرفیت فضای دیسک بود. از آن جا که دیگر مناطق کتابخانه به طور مشترک از همین امکان ذخیره سازی استفاده می کردند عواقب بسیار جدی داشته است. در یک مقطع مجبور شدیم گردآوری را تازمانی که فضای دیسک کافی در دسترس باشد به تعویق بیندازیم.

با وجود پیچیدگی های اخیر ، بیش تر وب گاه هایی جمع آوری شده بررسی کیفی و تأیید شده و با موفقیت آرشیو شده اند. در پایان برنامه برداشت ، بیش از 600 وبگاه برداشته شده تأیید و یا برای این رویداد خاص آرشیو شده اند که موفقیت (دستاورد) قابل توجهی برای برنامه آرشیو وب کتابخانه است.

به طور کلی انتخابات هیئت ،محلی آزمایش مهمی درباره گردآوری رویداد با استفاده از WCT بود. ما به سهولت قادر به برنامه ریزی و مدیریت این رویداد بزرگ آن ها برروی سیستم بودیم ، و تجربه های مفیدی در استفاده از ویژگی های سیستم به دست آوردیم که تاکنون موقعیت استفاده از آن ها را نداشتیم موضوع با اهمیت تر این که گردآوری این رویداد نشان داد که WCT با موفقیت می تواند با تقاضاهای فعالیت گردآوری وب و اندازه افزایش یافته تطبیق دهد مناطقی را که به بررسی و نظارت دقیق تر نیاز داشتند نیز مشخص کند و درس های ارزشمندی آموخته و تجربه ای به دست دهد.

متعاقب انتخابات هیئت محلی سال 2007 ، رویداد دیگری که بر پایه گردآوری قرار داشت جام جهانی راگبی 2007 بود این رویداد در مقیاس بسیار کوچک تر از گردآوری رویداد اول بود، اما شامل وب گاه های خارج از محدوده مقررات واسپاری قانونی زلاندنو می شد که امکان قانونی جمع آوری با استفاده از WCT را برای اولین بار نیاز داشت.

6. نتیجه گیری

انتخـاب گران الکترونیکی کتابخـانه به طور روز مره برای انتخاب نمودار گردآوری، و مرور وب گاه ها از WCT استفاده می کنند سپس آن ها را برای آرشیو دیجیتال پیشنهاد می دهند.

انتخاب گران الکترونیکی کار را با ویرایش 1.1 شروع کردند و پس از آشنایی بیشتر با این نرم افزار، از نوعی محدودیت روزافزون آگاه شدند که Specification ویرایش جدیدی را اطلاع می داد. نرم افزارهای کنترل کیفی به روز شده تفاوت عظیمی را به وجود آوردند و بسیاری از سایت ها که برای مرور آن ها با ویرایش های اولیه مشکل بودند ، جمع آوری و با ویرایش 1.2 مرور شدند.

هنوز برخی مسائل پر دردسـر و آزار دهنده وجود دارد و یا در حال ضبط آن ها در bugtraker روی وب گاه های منبع باز هم ضبط کرده ایم.

از آخرین هفته ها ویرایش جدید نرم افزار (ویرایش 3.) قابل بهره برداری شـد و انتظار داریم که بتواند در گردآوری گردش کار، نتایج بهتری ایجاد کند.

پس از دو دوره گردآوری کتابخانه ، ظرفیت مدل ها برای ذخیره آتی را توسـعه داده و نیاز به پهنای باند دارد. ما همچنین دسترسی نرم افزارها و سطح دامنه گردآوری وب گاه های زلاندنو را طراحی می کنیم.

منابع

.HTTrack Website Copier, (Accessed 2008-05-15) .1

Paynter, G. W. and Mason I. B. (2006) Building a Web Curator Tool for The National Library of New .2 Zealand. New Zealand Library Association (LIANZA) Conference, . October 2006. (Accessed 2008-05-15).

Web Information Management

:Vol. 1

Basics and Worldwide Initiatives

Edited by

Gholam Ali Montazer (Ph.D)

and

Tarbiat Modares University

Farzaneh Shadanpour

National Library and Archives of Islamic Republic of Iran

ص: 318

جلد 2

مشخصات کتاب

مدیریت منابع اطلاعاتی وب جلد دوم

دیدگاههای فناورانه اخلاقی و مدیریتی

به کوشش:

دکتر غلامعلی منتظر و

فرزانه شادان پور

زمستان 1391

فهرستنویسی پیش از انتشار کتابخانه ملی جمهوری اسلامی ایران

سرشناسه:منتظر، غلامعلی 1348 - ، گردآورنده

عنوان و نام پدیدآور: مدیریت منابع اطلاعاتی وب / به کوشش غلامعلی منتظر و فرزانه شادان پور.

مشخصات نشر:تهران سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران 1391

مشخصات ظاهری:2 ج.

شابک:دوره: 0 - 344-446 - 1964 - 978؛ ج. 2: 72 - 345-446-964-978؛

وضعیت فهرستنویسی:فینا

مندرجات:ج.1 مبانی و تجربه های جهانی - ج..2 دیدگاههای فناورانه، اخلاقی و مدیریتی.

موضوع:وب -- سایت ها -- مدیریت

موضوع: منابع اطلاعاتی -- مدیریت

موضوع:وب -- آرشیوسازی

شناسه افزوده:شادان پور، فرزانه 1344 - ، گردآورنده

شناسه افزوده:سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

رده بندی کنگره:1391 4م TK5105/8888

رده بندی دیویی:005/72

شماره کتابشناسی ملی:3077380

خیراندیش دیجیتالی : انجمن مددکاری امام زمان (عج) اصفهان

ویراستار کتاب : خانم مرضیه محمدی سرپیری

ص: 1

اشاره

سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

عنوان مدیریت منابع اطلاعاتی ،وب جلد اول مبانی و تجربه های جهانی

به کوشش دکتر غلامعلی منتظر (دانشیار دانشگاه تربیت مدرس) و فرزانه شادان پور(مربی سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران)

ویراستار ادبی آرزو تجلی کار شناس ارشد جامعه شناسی، سازمان اسناد کتابخانه ملی جمهوری اسلامی ایران)

تنظیم و تصحیح مهشید برجیان کار شناس ارشد کتابداری و اطلاع رسانی سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران)

ویراستار استنادی مقالات تألیفی فروزان رضایی نیا کارشناس کتابداری و اطلاع رسانی سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران)

نمونه خوانی و اصلاحات مهشید ،برجیان فاطمه ،رمضانپور آمنه ،هزارخانی زهرا ،زاهدی محمد رضا ،میقانی ملیحه حاجی زاده مقدم

طراحی جلد و صفحه آرایی شهره خوری

ناظر فنی چاپ نصرت الله امیرآبادی

ناشر سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

شمارگان: 500 نسخه

بها: 20000 تومان

:نشانی تهران بزرگراه شهید حقانی (غرب به شرق)

بعد از ایستگاه مترو، بلوار کتابخانه ملی

تلفن فروشگاه 81623318 - 81623315- 88941946

دورنگار: 88947496

وب سایت: www.nlai.ir

پست الکترونیک انتشارات Publication @nlai.ir

سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

مدیریت منابع اطلاعاتی وب

جلد دوم

دیدگاه های فناورانه اخلاقی و مدیریتی

ص: 3

فهرست مطالب

ص: 5

سخن نخست

از ویژگیهای قرون گذشته بی خبری بود و تمایز جدی عصر جدید نسبت به گذشته دسترسی آسان به اطلاعات .است بشر با از سر گذراندن سه موج و پارادایم ،کشاورزی صنعت و اطلاعات امروز در قرن بیست و یکم پا در عصر انفجار اطلاعات نهاده است این امر فی نفسه نه مطلوب است نه مذموم، بلکه به نحوه مدیریت ما نسبت به اطلاعات باز میگردد.

بشر امروزی به دلیل رشد روزافزون علم و فناوری در شرایط هشدارآمیز عدم قطعیت بسر میبرد و همین مدیریت و تصمیم گیری را با چالش جدی روبرو ساخته است. اگر اطلاعات درست مدیریت شود و در تصمیم گیریها به موقع به کار آید و از دو ویژگی صحت و سرعت برخوردار باشد میتواند منشأ تصمیمهای تحول آفرین شود. ویژگی دیگر این عصر ظهور و حضور همه جانبه اطلاعات دیجیتالی .است دورانی فرا رسیده است که در آن بناست دانش مدون و تفکر مضبوط بشر علاوه بر کاغذ و حتی بیش از آن بر محمل «بیت» ها مسیر ،تولید نشرو اشاعه و مصرف را بپیماید. هم اطلاعات تولید شده تحت وب و هم میزان استفاده از این اطلاعات با سرعت فزاینده ای رو به رشد است. کشور ما بنابر اطلاعات وثیق از حیث تعداد کاربران و میزان حضور و فعالیت آنها دروب جایگاه نخست را در منطقهٔ خاور میانه داراست. این روند رو به رشد با نصب العین قرار دادن آرمانهای بلند انقلاب اسلامی در ترویج تفکر رهایی بخش اسلام ولایت مدار وظیفه خطیری بر دوش نهادها و دستگاههای مسئول تولید، سیاستگذاری و نشر محتوا در محیط وب قرار می دهد و آن انجام بررسیهای علمی و مستند به منظور ابتنای سیاستگذاریها و عملکردها بر مبانی صحیح و کارآمد و متناسب با نیازهای گوناگون کاربران در این محیط است. اما وجه دیگر، صیانت از این محتوا و انتقال آن به نسلهای آینده است که با توجه به ناپایداری محتوای قرار گرفته بر اینترنت و فناوری پیشرفته ای که برای چنین امر خطیری لازم است

از اهمیت مضاعفی برخوردار می شود.

سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران بنابر مأموریت خویش دایر بر صیانت از میراث فکری کشور و اشاعه آن عزم راسخ داشته است که برای مدیریت منابع اطلاعاتی مهم و رو به رشد وبی نیز چاره اندیشی نماید؛ بنابر این در سال 1389 نخستین بار در کشور به تهیهٔ ساز و کار لازم برای ایجاد

آرشیو ملی وب همت گماشته است.

از دیگر سو، سازمان با علم به این که مـدیریت در این حوزه مشارکت همـهٔ صاحبان اندیشه در حوزه تولید سازماندهی و اشاعه اطلاعات تحت وب را می طلبد مصـمم» شد نخستین کنفرانس ملی مدیریت منابع اطلاعاتی وب را برگزار نماید تا اهل علم و فناوری در این مجمع با هم اندیشی و تضارب آراء همچون گذشته این سازمان را یار و یاور باشند.

این اثر مجموعه ای است فراهم آمـده از تلاش پژوهشـگرانی که با وجود نو بودن مباحث مطرح شـده در محورهای موضوعی ،کنفرانس به ارائه ثمرهٔ پژوهشـهای خود همت نمودند؛ که با برگزیده ای از مقالات ترجمه ای در این عرصهٔ پژوهشی ادغام و به طبع رسیده است. رجاء واثق دارم که با الطاف الهی از این پس مدیریت منابع اطلاعاتی وب و آرشـیو وب به طور خاص موضوع پژوهش و ابتکار عمل اهل دانش و فناوری در کشورمان قرار خواهد گرفت و در این عرصه نیز فرزندان این مرز و بوم تجسم گفتار نغز رسول اعظم صلی الله علیه و آله خواهند بود که علم اگر تا ثریا ،برود مردانی از فارس بدان دست خواهند یافت».

اسحق صلاحی

رئیس کنفرانس و

رئیس سازمان اسناد و کتابخانه ملی

جمهوری اسلامی ایران

به جای مقدمه

گسترش روزافزون اطلاعات در شبکه اینترنت و سادگی بارگذاری انواع داده ها بر وب جهان را با شکل جدیدی از تولید انتشار و مصرف اطلاعات مواجه کرده است تغییر جایگاه شهروندان جامعه از مصرف کنندهٔ صرف اطلاعات به مولّد و ناشر اطلاعات و فارغ از سازوکارهای موسوم سبب ساز روابطی جدید در عرصهٔ ارتباطات اجتماعی و فرهنگ شده است. از سویی حجم رو به تزاید داده ها و چرخه عمر کوتاه اطلاعات موجود در وب موجب شده که گردآوری»، «پالایش»، «سازماندهی»، ذخیره سازی» و «اشاعه» آنها در زمرهٔ مسائل پژوهشی در نهادهای علمی و نیز بخشهای پژوهش و نوآوری شرکتها قرار گیرد؛ ضمن اینکه حفظ و دسترسی پایدار به اطلاعات موجود در وب که خود جزئی از میراث فکری ملتها محسوب میشود به دغدغه ای جدی برای سازمانها متولی حفظ و اشاعه میراث ،فکری به ویژه کتابخانههای ،ملی بدل شده است.

این حوزه در جهان موضوعی نسبتا جدید است و پیشینهٔ آن به کمتر از پانزده سال میرسد لیکن با سرعتی شتابان در حال رشد است و محققان مختلفی را از زوایای مختلف فنی حقوقی اقتصادی و حتی اخلاقی به سوی خود جذب کرده که گواه آن نیز طیف وسیعی از مقاله ها کتاب ها و گزارشهای سازمانی است که در طی چند سال اخیر در سطح جهانی منتشر شده است به رغم این نکات در ایران همچنان این ،زمینه حوزه ای بکر و کمتر مورد توجه محسوب میشود و در طی سالهای اخیر کمتر تحقیق بدان پرداخته لیکن تزاید اطلاعات فارسی بر روی وب و برنامه های ملی کشور مبنی بر توسعه کاربریهای مختلف بر شبکههای اطلاعاتی (از جمله توسعهٔ دولت الکترونیکی، یادگیری الکترونیکی و کتابخانههای دیجیتالی) لزوم توجه به این موضوع را بیش از پیش نمایان میسازد به همین دلیل سازمان اسناد و کتابخانه ملی جمهوری اسلامی همزمان با برگزاری نخستین کنفرانس ملی مدیریت منابع اطلاعاتی «وب» درصدد برآمد تا این حوزه را هرچه بیشتر به متخصصان و پژوهشگران بازشناساند کتاب پیش رو حاصل همین نیت

متولیان این موضوع مهم است.

این کتاب مجموعهای قریب به 30 مقالهٔ برگزیده از مهمترین منابع علمی منتشر شده در جهان و نیز قریب به 15 مقالهٔ برگزیده از صاحب نظران ایران است که در قالب دو جلد تقدیم

حضور خوانندگان ارجمند می.شود این مقالات در چهار موضوع اصلی به شرح زیر تقسیم شده اند:

- مبانی مدیریت و آرشیو وب
- تجارب جهانی و مسائل بومی در مدیریت و آرشیو وب
- مسائل فناورانه
- مسائل اخلاقی و مدیریتی

بی گمـان این مجموعه میتوانست به افزودنیهای دیگر)هم از منابع خارجی و هم از دیـدگاه سایر متخصـصان ایرانی) به اثری پربارتر بـدل گردد لیک نخستین گامی است که در این حوزه برداشـته شده و مطمئنا در مراحل بعدی با همت سایر ،اندیشـمندان ویراست هایی غنی تر از آن حاصل خواهد آمد نگارنده امیدوار است این مجموعه به مثابه بذری باشد که در کشتزار ذهن پژوهشـگران کاشـته شده و ان شاء الله در آینده ای نه چندان دور به نهالی پرطراوت در عرصه علم و عمل در جامعه اسلامیمان مبدل گردد.

در پیدایی این اثر کسان بسیاری همراهی و همکاری داشته اند که مقدم بر همه اندیشمندانی است که متن هر مقاله به خامهٔ دانش افزای آنان امکان وجود یافته است از این رو نگارنـده سـپاس فروتنانه خود را نثار نگارنـدگان و مترجمان ارجمنـد این اثر مینمایـد گردآوری، تنظیم و آماده سازی مطالب کتاب به همت خانمها فرزانه شادان پور و مهشـید برجیان بوده و ویراسـتاری آن را خانم آرزو تجلی بر عهده داشته.اند ویراسـتار اسـتنادی مقالات تألیفی را سـرکار خانم فروزان رضایی نیا به انجام رساندهاند و سـرکار خانم دکتر میترا صمیعی زحمت چکیده نویسـی شـماری از مقالات را که فاقد چکیده بودند متقبل شدند نمونه خوانی و اصلاحات اثر حاصل تلاش خانمها مهشید برجیان فاطمه رمضانپور ،آهنگری آمنه هزار خوانی زهرا ،زاهـدی ملیحه حاجی زاده مقدم و آقای محمد رضا میقانی بوده است ضـمن اینکه زیبایی متن و صفحه آرایی آن مدیون حسن سـلیقه سرکار خانم شهرهٔ خوری است زحمات لیتوگرافی چاپ و صحافی کتاب نیز برعهده جناب آقای امیر آبادی بوده که برخود فرض میدانـد از همهٔ این بزرگواران صـمیمانه تشـکر کند. گمان پدیدآمدن این اثر به همت مسؤولان گرانمایه سازمان اسـناد و کتابخـانه ملی جمهوری اسـلامی بوده است و نگارنـده امیـدوار است خداونـد آنان را در مسـیر خـدمت به فرهنگ و دانش ایران اسلامی مورد تأیید قرار دهاد.

اللهم وفقنا لما تحب و ترضی

غلامعلی منتظر

تهران- بهمن ماه یکهزار و سیصد و نود و یک خورشیدی

فصل اول : مبانی مدیریت و آرشیو وب

اشاره

ص: 3

چکیـده فیـدهای وب با فرمت آر.اس.اس یا مبتنی بر اتم(1) .ایکس. ام. ال، اسـناد توصـیفی در حال توسـعه ای هستند که کانون هاب) پویای وبگاهها مشـخص میکننـد و به مشترکان کمک میکنند تا با تازه ترین محتوای وبگاه مورد علاقه و روزآمد خود در ارتباط باشـند در این مقاله نشان میـدهیم چگونه فیـدهای وب میتوانند ابزاری مفید برای اسـتخراج اطلاعات و تشـخیص تغییر صـفحه وب باشـند. معمولاً صفحه هـای وب که با آیتمهای فیـد ارجاع می شونـد عبارت انـد از پسـتهای وب نوشت و یا مقاله های خبری، و داده هایی با ماهیتی پویا (سـپس زودگذر) که به صورت موضـعی در یک کانال فید خوشه بندی میشود ما کانالهای وب را پایش و از صفحه های وب مرتبط، متن و منابع مربوط به مقاله های وب را استخراج میکنیم نتیجه کار با بر چسب زمان و فراداده اضافی استخراج شده از فید غنی شده، در یک شیء داده ای محصور می شود شـیء داده ای به شکل اطلاعات خاصی خواهد بود که فاقد هرگونه عناصر تمپلیت یا تبلیغات میباشد این عناصر بی ربط، که معمولاً boileplate نامیده می شوند نه تنها از دید برنامه خزشـگر وقت گیر و جاگیر هستند، بلکه مانع فرآیند تجزیه و تحلیل دادههـا میشونـد مـا نخست بـا خزش فیـدهای وب برای یـک دوره زمانی و مشاهـده جنبه های زمانی ،آنها روی مجموعه ای از آنها نوعی بررسـی الگوریتم مورد اسـتفاده برای استخراج مقاله را ارائه میکنیم؛الگوریتمی آماری کرده، سپس که از معانی فید(2) به طور اختصاصی تر شـرح و عنوان آیتمهای فید به منظور شناسایی گره DOM در صـفحه اچ تی ام ال که حاوی مقاله است، اسـتفاده می.کند از اشیای داده ای ساخته شـده با این شـیوه میتوان برای مجموعه همپوشانی معنایی برای آرشـیو و یا در زمینه یک خزش تدریجی اسـتفاده کرد که آن را از طریق تشـخیص تغییر در سطح شیء داده ای کارآمدتر میکند آزمایشهای انجام شده بر روی روش استخراج به منظور روایی رویکرد مورد نظر، بـا نتایـج خوبی – حتی در مواردی که تکنیکهـای دیگر شـکست خورده بودنـد – انجام میشونـد. در نهایت در مورد برنامه های مفید براساس اسـتخراج و تغییر تشـخیص شـیء وبگاه بحث میکنیم. کلید واژه ها آرشیو کردن ،وب شیء داده ای فید ،وب پویایی صفحه های وب

Atom -1

Semantic -2

آرشیو اشیای دادهای با استفاده از فیدهای وب

نوشته ماریلنا اویتا(1) و پیرسنلارت(2) | :ترجمه لیلی سیفی(3)

1. مقدمه

آرشیو وب بایگانی کردن (تارنما) [15] به فرآیند جمع آوری مکرر محتوای بخشهایی از شبکه جهانی ،وب به منظور حصول اطمینان از حفظ آن و اجازه دسترسی به اطلاعات – حتی پس از بین رفتن وب اطلاق می شود برنامه خزشگر آرشیو وب به دنبال همان مراحل اولیه به محض اینکه خزشگر موتور جست وجو شاخصهایی را برای صفحه های وب ایجاد میکند اجرا می.شود با این حال، خزشگر آرشیو زمانی که نسخه های جدید کشف میشوند، شاخصها را به نسخه های قدیمی انتقال نمی دهد، بلکه آنها را ذخیره و نسخه ها را به موقع ارجاع میدهد نتیجه نهایی مجموعه ای از صفحه های وب است که می توان به صورت ناپیوسته مرور کرد، و در شکل ایده آل، می توان به صورت موقتی و معنایی جست وجو کرد. خزش منسجم سایتهایی که در به روزرسانی عرضه اطلاعات خود بسیار سریع هستند، کار آسانی .نیست گزینه مرسوم خزش تصویرهای لحظه ای است؛ اما خزش کل مجموعه در بازه زمانی دوردست و) به اندازه کافی مکرر از لحاظ استفاده از پهنای باند شبکه بسیار گران قیمت است و در پایان نیز هم برای برخی مناطق سایت کاهنده و برای بخشهای پویاتر سایت ناقص خواهد بود علاوه بر این چون

1- Marilena Oita

2- Pierre Senellart

3- استادیار دانشگاه بیرجند Leili.seifi@gmail.com

اجرای یک خزش تلفیقی زمان بر است و در عین حال ممکن است منبع تغییر کند مشکلاتی ناشی از عدم انسجام زمانی تصویر لحظه ای معین به وجود خواهد آمد.

هنگـام تصـمیم گیری برای خزش تـدریجی آنچه که انـدازه کافی مکرر است، بایـد مشـخص شود که با سایتهایی انطباق بهتری دارد که در ساختارشان دارای بخشـهای موقت ناهمگن هسـتند هر خزشـگر تـدریجی نسـخه کامل سایت را در یک مرحله خزش میکنـد و این روند نوعیراه اندازی (1) است که به خزشـگر اطلاع میدهـد که کدام محتوای سایت اضافه یا به روزرسانی شده است، و خزشـگر فقط محتوای اصـلاح شـده را خزش کرده و آن را به صورت یک ساختار داده ای دلتا(2) ذخیره می کند. مشـکل انجام خزش تدریجی الزاماً تعیین این پویایی است چنـد وقت به چنـد وقت صـفحه های وب جدیـد اضافه یا اینکه صـفحه های موجود اصـلاح میشونـد و کارآمدی تشـخیص تغییرات با علم به اینکه عوامل زیادی وجود دارند که میتوانند روی فرآیند تشـخیص اثر منفی داشته باشند موضوع مورد توجه ما تکنیکهایی است که می توانـد برای بهبود فرآینـد تشـخیص مورد اسـتفاده قرار گیرد ، در مورد خاص که در آن صـفحه های که باید خزش شوند دارای .آر.اس.اس یا فیدهای اتم الصاقی هستند.

روش سـنتی تشـخیص تغییر بین دو نسـخه پی در پی یک صفحه ،وب مقایسه در محتوا با استفاده از معیارهای شباهت ،است که با در هم سازی و امضاهای محتوا به منظور ویرایش فاصـله ها برای انعطاف پذیری متفاوت است. تغییر به هر شیوه ای که شناسایی شود اهمیت آن با توجه به نوع محتوای یک صـفحه وب مورد ارزیابی قرار نمی گیرد. با این حال درک اینکه آیا تغییرات مربوط به محتوای اصـلی صـفحه از نقطه نظر معنایی باشـند بسـیار مهم است؛ زیرا در برخی برنامه های کاربردی، تغییراتی که تنها روی قسمتهای [boilerplate [12 صفحه از قبیل ،منوها تمپلیتهای ارائه یا تبلیغات اثر میگذارند ممکن است به راحتی نادیده گرفته شده است.

مطـالعه انـدکی در مورد فیـدهای وب صورت گرفته است در حـالی که پدیـده ای به سـرعت در حـال تحول است. ما توجه خود را به این واقعیت معطوف می داریم که میتوان آنهـا را به عنوان ابزاری در تجزیه و تحلیـل یک وبگاه قبل و در حین خزش اسـتفاده کرد فیـدهای وب علاـوه بر اینکه راهی برای تبلیغ محتوا هسـتند جهت طبقه بنـدی منـابع اطلاعاتی و نوع محتوا توسط موتورهای جست و جو اسـتفاده می شونـد. به طور خلاصه از طریق فیدهای ،وب جنبههای مهم وبگاه پویا را میتوان در چوب یک خزش وب – به منظور آگاه تر ساختن آن از اطلاعاتی که تهیه میکند استخراج و بهره برداری کرد.

ماهیت فید

آموزنده است اطلاعات جدید و زمان ورود آنها، و انتشارشان توسط کانال را ترکیب می کند؛ توصـیفی است توضـیح میدهد چه نوع منابع جدیدی همراه با عنوان توضیح، و سایر عناصر تگ ممکن افزوده شود.

1- Trigger

2- Delta data Structure

هدف ما استخراج دادههای ساختاری از صفحه های وب، با کمک فیدهاست. اساس و پایه رویکرد ما این است که یک آیتم از یک کانال [چگونه با یک شیء داده ای در یک صفحه وب انطباق مییابد.

بنابراین، فراداده ای که در مورد آیتم مورد نظر در فید به دست می آید را میتوان برای تشخیص و استخراج شیء داده در صفحه وب استفاده کرد.

مفهوم شیء داده ای دارای تفسیرهای مختلفی در علوم رایانه است به منظور روشن ساختن اهمیت آن در زمینه مورد بررسی باید گفت که یک شیء داده ای نمونه ای از منابع ارجاعی فید است و فراداده هایی دارد که با عبارات «و(1)».. [5] و «به(2)» [13] با هم مرتبط میشوند خود مفهوم شیء داده یک ترکیب است و از این نظر میتوان آن را به عنوان سند منطقی ما میگوییم معنایی در تقابل با «سند فرامتن»(صفحه اچ.تی.ام.ال) تلقی کرد.

حتی اگر شییء اغلب یک مقاله وب ،باشد میتواند مدخل فرهنگ لغت ،نظر، پیامی در یک نشست ویدئو، یک وضعیت، و هر نوع منابع دیگری باشد که به طور منحصر به فرد با آیتمهای فید وب مرتبط شده است رویکردهای مستقیم برای شناسایی محتوای اصلی یک صفحه ،وب از جمله در نظر گرفتن عناصری که پس از تمیز کردن پایه مطالعه تراکم متن در مناطق خاصی از صفحه [12]، و یا حتی شناسایی برجسته ترین مناطق بصری صفحه ،[29] با پیچیده تر شدن شیوه رمزگذاری و تکامل یافتن خود درک مستقیم مقاله ،وب در حال منسوخ شدن هستند اشیای داده ای میتوانند مطابق با بخشهای کوچکی از یک صفحه ،وب ساده یا مرکب شامل چندرسانه ایها و حتی جاوا اسکریپتهای خوبی باشند که باید مورد بهره برداری قرار گیرند. بسته به زمینه میتوانیم یک یا چند شیء داده چندگانه در هر صفحه داشته و تفاوت بین آنها را با استفاده از معناشناسی با قابلیت خوانده شدن توسط ماشین و یا با استفاده از فناوری هوشمند بشر انجام داد ما اهرمهای معنایی هستیم که توسط فیدهای وب به صفحه های وب پویا آورده میشوند فیدها را – که در فرم ایکس ام ال مانند با عناصر استاندارد نوشته شده اند – می توان برای به تصرف در آوردن برخی جنبه های مهم اطلاعاتی استفاده کرد که می خواهیم استخراج کنیم با استفاده از تکنیکهای کلاسیک بازیابی اطلاعات که از ساختار یک آیتم فید به دست می آوریم، توصیفگر معنایی شییء که به عنوان ورودی برای الگوریتم استخراج می شود، استفاده خواهد شد. در مرحله بعد، منطقه ای در یک صفحه وب را شناسایی میکنیم که شامل محتوای شیء داده بوده و آن را با استفاده از تجزیه و تحلیل تراکم معنایی استخراج می.کند میتوان با داشتن محتوای استخراج شده در زمان و بخش مهمی از خواص آن پرس و جوهای پیچیده را از نقطه نظر زمانی و معنایی اجرا کرد.

در بخش بعدی برخی کارهای مرتبط را ارائه و در بخش 3 نتایج حاصل از مطالعاتمان بر روی فیدهای وب توصیف میکنیم که به منظور تعیین ارزش این فیدها در فرآیند آرشیو وب انجام شد. ما در بخش 4 شرح میدهیم که چگونه معناییهای مشخص شده توسط عناصر خاص موجود در یک فید را میتوان برای استخراج اشیای داده ای استفاده کرد. در بخش ، آزمایشهای مربوط به استخراج

1- and

2- to

اشیای داده ای نشان داده میشوند نتیجه گیری ما با ترسیم اهمیت اشیای داده ای استخراج شده در زمینه برنامه های کاربردی است.

2.کارهای مرتبط

آرشیو وب چه به عنوان یک ضرورت یا وظیفه دیده شود به تازگی از اهمیت زیادی برخوردار شده است که علت آن ماهیت فرّار ،وب و به ویژه ارزشی است که اطلاعات از دست رفته میتواند برای نسلهای آینده داشته باشد. آرشیو اینترنتی [11] یکی از مبتکران جنبش آرشیو وب است. بسیاری از بازیگران دیگر وجود دارد که به طور فعال خزش را به عنوان بخشی از رسالت حفظ میراث خود اجرا کرده و برای اینکه مجموعه آرشیو وب جهانی به صورت واحد تقارب ،یابد تلاشهایی در شرف انجام است [28]. اگرچه فیدهای وب به طور معمول به عنوان انواع دیگری از اسناد وب توسط خزشگرهای آرشیو نمایه میشوند تلاش محدودی در بهره برداری از این ویژگیهای در روند آرشیو کردن صورت گرفته است آرشیو پرس که پروژه ای برای آرشیو کردن وب نوشت است [20] در حال توسعه نوعی پلاگین وردپرس است که پستها را با استفاده از فیدهای وب آرشیو می.کند اشکال اصلی این است که تنها می توان محتوایی را گرفت که توسط فید آر.اس.اس قابلیت تحویل را داشته باشد. هر فید آر.اس. اس. در واقع میتواند پوشش کامل مقاله و فایلهای رسانه ای را داشته باشد. اما این مورد بسیار نادر است زیرا هر فید اغلب فقط یک شیوه محتوای تبلیغاتی است در مقابل سرنخ های فید را با هدف بهره برداری از اطلاعات واقعی مربوط تحت کنترل در می آوریم. علاوه بر این، خود را به یک چارچوب ،وبلاگ نویسی یا وبلاگ نویسی به طور خاص محدود نمی کنیم.

هنگام مطالعه پویایی صفحه های وب دو دیدگاه وجود دارد این تغییر میتواند ناشی از انتشار محتوای جدید و یا ناشی از تغییراتی باشد که در صفحه های موجود رخ میدهد. برای جلوگیری از تجمع بیش از حد بر روی سرور وبگاههایی که محتوای به موقع را تهیه میکنند ، [22]، رویکرد رایانه انطباقی پیشنهاد می شود. مدل ناهمگن پواسون(1) در فیدهای وب، به منظور یادگیری الگوهای ارسالی در بلاگها و پیش بینی بررسی مجدد بهینه برای محتوای جدید مورد استفاده قرار می گیرد در زمینه خزش افزایشی در حال حاضر مشکل مشابه این است که آمار خود را در جنبه های زمانی فیدها با هدف بیان ارزش خود در فرآیند یادگیری راهبرد انتشارات ارائه میدهد در زمینه خزش تدریجی نیز همان مشکل را داریم آمارهای ما در مورد جنبه های موقتی با هدف بیان ارزشهایشان در فرآیند یادگیری راهبرد انتشار انجام می شوند. به منظور مدل سازی رفتار وبگاهها در طول زمان و خزش مؤثرتر ،آنها مطالعات دیگر دربارهٔ درک درست پویایی صفحه های وب متمرکز شدهاند ،[8 1 ، 16] میزان سرعت تغییرات محتوای وب و نوع ماهیت این تغییرات معانی ضمنی در مورد ساختار و همبستگی با موضوع صفحه ها برای انطباق خودکار خزش با آهنگ پیشبینی شده تغییرات مقدماتی وجود دارد [10] از مدل فضای برداری متنی

1- Poisson

برای شناسایی الگوهای صفحه و برای آموزش فیلترهای کالمن(1) استفاده میشود در پایان تغییر رویدادی است که با پیش بینی مطابقت ندارد با این ،حال فرضیه خطی سیستم و عدم تمامیت مدل فضای برداری اشکالهای ممکن را نشان میدهد. برای شناسایی و ذخیره سازی محتوای اصلاح شده یک صفحه وب خزشگر آرشیو کد منبع باز 23] Heritrix] با استفاده از عبارتهای غیر مستدل و منظم برای فیلتر کردن تغییرات بی ربط استفاده میشود در تلاش برای برآورد منصفانه آهنگ تغییر صفحه وب مدل رسمی تر پیش بینی توسط چو(2) و گارسیا- مولینا(3) [3] مورد مطالعه قرار گرفت که در آن ،نویسنده بررسی می که آیا تغییرات یک صفحه وب از فرآیند پواسون همگن تبعیت میکند یا نه با این وصف، شناسایی آهنگ تغییر توسط هر دو رویکرد به عنوان چالشی مطرح میشود ما استخراج اشیای داده ای در داخل این موضوع را توصیه میکنیم هنگامی که دستیابی اشیای داده ای به موقع انجام می شود، یک معیار شباهت را میتوان در محتویات و یا بر روی خواص اشیای داده ای استفاده کرد. با این فرض که آنها دو نسخه پی در پی یک صفحه وب را با استفاده از چند روش برای تشخیص تغییر به دست آورده اند پهلوان(4) بن سعد(5)،و گانکارسکی (6) [19] روی تشخیص تغییراتی تمرکز میکنند که براساس یک نسخه قدیمی در نسخه جدید رخ داده است برای این منظور، الگوریتم وپس(7) [29] برای شناسایی بخشهای معنایی مرتبط یک صفحه وب استفاده میشود که به منظور تشخیص تغییرات ساختاری و محتوایی مقایسه می شوند. ابتکارهایی روی ظاهر بصری یک صفحه وب ایجاد میشوند تا محتوایی را با هم گروه بندی کنند که به نظر میرسد در صفحه از اهمیت یکسانی برخوردار باشند این ابتکارها مشکل را به طور جامع پوشش نمیدهند و الگوریتم محاسباتی گران قیمت است. رویکرد ما محدودتر است، چون نیاز داریم که از طریق فید عبور کنیم؛ در عین حال، می تواند مؤثرتر باشد: با شناسایی مناطق معنایی «مهم» در صفحه وب میتوانیم بر روی تغییراتی تمرکز کنیم که به تلاش کمتری وابسته هستند.

یک مقاله وب شناسایی شده با استفاده از عنوان و شرح آیتم فید باید از کد اچ.تی.ام.ال. صفحه وب مرتبط استخراج شود کار زیادی در استخراج بدون نظارت دادههای ساختاری از صفحه های وب انجام شده است؛ بسیاری از اینها مبتنی بر ام. دی. آر(8). [14]، اکس آلژ(9)[2] یا دونده جاده [4] هستند. در واقع این روشها تلاش میکنند تا با استنتاج گرامر برای کد اچ تی ام ال به طور خودکار لفافهای تولید کند که حاوی اطلاعات مورد علاقه به طور کلی به ،آن سوابق داده گفته میشود ،باشد به شیوه ای که به دانش قیاسی در مورد صفحه های هدف و محتویاتشان وابسته نباشد معمولاً صفحه های مختلف که همان قالب را دارند به منظور مقایسه زوجی آنها و کشف الگوهای مشترک و قواعد کد گذاری مورد نیاز هستند این

1- Kalman

2- Cho

3- Garcia-Molina

4- Pehlivan

5- Ben Saad

6- Gancarski

7- VIPS

8- MDR

9- Ex Alg

قواعد یا از طریق بررسی شباهتها و تفاوتهای بین صفحه ها [4] یا با ساخت کلاسهای هم ارز [2] تأکید میشوند. برخلاف کارهای پیشین ام.دی.آر. [14] ساختار درختی DOM صفحه اچ.تی.ام.ال. را در نظر گرفته و منطقه دادهها را با پیدا کردن گرهی کلی شناسایی میکند که شامل بیشترین تعداد فرزندانی است که الگوهای مشابه را با توجه به اندازه گیری شباهت ارایه میدهند. حتی اگرچه ما توجه خود را روی سایتهای حاوی مقاله های وبی متمرکز میکنیم – الزاماً به شیوه ای که ما در یک صفحه وب میبینیم – متوالی نیستند اما ما با یک مشکل مشابه استخراج دادهها احتمالا ساختاری با یک ماهیت پیچیده تر و منزوی تر از این رویکردها مواجه هستیم و همانطور که در بخش 4 توضیح داده خواهد شد، به نحوی از تکنیکهای مربوط استفاده می کنیم.

جدول .1 انواع فیدهای مجموعه ای

عکس

قواعد، یا از طریق بررسی شباهت‌ها و تفاوت‌های بین صفحه‌ها [٤]، یا با ساخت کلاس‌های هم ارز [۲] تأکید می‌شوند. برخلاف کارهای پیشین، ام.دی.آر. [١٤] ساختار درختی DOM صفحه اچ.تی.ام.ال. را در نظر گرفته، و منطقه داده‌ها را با پیدا کردن گرهی کلی شناسایی می‌کند که شامل بیشترین تعداد فرزندانی است که الگوهای مشابه را با توجه به اندازه‌گیری شباهت ارایه می‌دهند. حتی اگرچه ما توجه خود را روی سایت‌های حاوی مقاله‌های وبی متمرکز می‌کنیم ـ الزاماً به شیوه‌ای که ما در یک صفحه وب می‌بینیم– متوالی نیستند، اما ما با یک مشکل مشابه استخراج داده ها (احتمالاً ساختاری با یک ماهیت پیچیده‌تر و منزوی‌تر از این رویکردها) مواجه هستیم، و همانطور که در بخش ٤ توضیح داده خواهد شد، به نحوی از تکنیک‌های مربوط استفاده می‌کنیم.

جدول ۱. انواع فیدهای مجموعه‌ای

Type	Number	Proportion
Atom	21	6.1%
RDF	30	8.8%
RSS 0.91	1	0.2%
RSS 2.0	288	84.7%
Total	340	100.0%

دو رویکرد دیگر برای مشکل استخراج مقاله اصلی از یک صفحه وب به‌تازگی توسط خولچوتر [۱۰]، فن خسار [۱۱] و نجدی[۱۲] (۱۲) و پاسترناک [۱۳] و راث[۱۴] [۱۸] پیشنهاد شده است. در حالی که [۱۲] از تراکم متن بر روی صفحه وب برای شناسایی مقاله استفاده می‌کند، [۱۸] از روش تقسیم‌بندی متن توالی برای رسیدن به نتیجه مشابه استفاده می‌کند. الگوریتم استخراج اشیای داده ما، مشابه به ام.دی.آر.، از برخی ابتکارات در کد اچ.تی.ام.ال. به‌منظور شناسایی منطقه‌ای که در آن می‌توان مقاله را یافت استفاده می‌کند، اما برخلاف تمام روش‌های برشمرده شده ما با استفاده از معنایی برگرفته شده از فید برای استخراج اشیای داده‌ای استفاده می‌کنیم. این امر به وضوح تنها برای صفحه‌هایی ممکن است که به یک فید وب پیوند شوند. در بخش ۵، ما نتایج را با نتایج الگوریتم Boilerpipe [۱۲]، که به‌عنوان پایه از آنچه که می‌توان بدون بهره‌گیری از اطلاعات فید وب انجام داد، مقایسه می‌کنیم. طبق بررسی‌های ما، هیچ کار قبلی، که اطلاعات معنایی را (که ممکن است از فید وب یا از هر منبع دیگر آمده باشد) جهت استخراج بخش مربوط به صفحه وب ـ به‌شیوه‌ای کلی و بدون نظارت ـ اهرم نمی‌شود.

10. Khoschutter
11. Fankhauser
12. Nejdi
13. Pasternack
14. Roth

دو رویکرد دیگر برای مشکل استخراج مقاله اصلی از یک صفحه وب به تازگی توسط خولچوتر(1)، فن خسار(2)،و نجدی(3) (12) و پاسترناک(4) و راث(5) [18] پیشنهاد شده است. در حالی که [12] از تراکم متن بر روی صفحه وب برای شناسایی مقاله استفاده میکند [18] از روش تقسیم بندی متن توالی برای رسیدن به نتیجه مشابه استفاده میکند. الگوریتم استخراج اشیای داده ،ما مشابه به ام.دی.آر.، از برخی ابتکارات در کد اچ تی ام ال به منظور شناسایی منطقه ای که در آن میتوان مقاله را یافت استفاده می کند، اما برخلاف تمام روشهای بر شمرده شده ما با استفاده از معنایی برگرفته شده از فید برای استخراج اشیای داده ای استفاده میکنیم این امر به وضوح تنها برای صفحه هایی ممکن است که به یک فید وب پیوند شوند در بخش 5 ما نتایج را با نتایج الگوریتم Boilerie [12] که به عنوان پایه از آنچه که میتوان

بدون بهره گیری از اطلاعات فید وب انجام داد مقایسه میکنیم طبق بررسیهای ما، هیچ کار قبلی که اطلاعات معنایی را (که ممکن است از فید وب یا از هر منبع دیگر آمده باشد) جهت استخراج بخش مربوط به صفحه وب – به شیوه ای کلی و بدون نظارت – اهرم نمی شود.

ص: 10

1- Khoschutter

2- Fankhauser

3- Nejdi

4- Pasternack

5- Roth

3.فیدهای وب آمارها و روندها

به منظور اثبات ارزش فیـدها به عنوان ابزارهای تجزیه و تحلیل در فرآینـد تشـخیص تغییر وب طی یک دوره کمی بیش از یک ماه دو بار در روز تعداد 400 فیـد وب را همراه بـا تمام صـفحه های وب مرتبط با آن خزش کرده ایم. نخست، چگونگی انتخاب این فیلها را توصـیف و سپس برخی آمار جالب مجموعه داده هایمان را گزارش میکنیم.

3-1. فراهم آوری

مجموعه ای از فیـدهای وب بـا عبور از بخش بزرگی از طریق یـک موتور جست وجوی فیـد به نـام جست وجوی آر.اس.اس.4[21](1) جمع آوری شد. این موتور جست و جو تعداد فیدها و صفحه های کانال مرتبط برای یک کلید واژه را بررسی میکند ما شیوه ای را که در آن نتایـج به شـکل سوابق برگردانده می،شد کنار گذاشتیم و تمام یو. آر. ال.(2) های تجزیه شـده را جهت تجزیه و تحلیل بیشتر در فهرسـت فایل قرار دادیم.

کلید واژه های انتخابی برای ردیابی واسط جست و جو نام دامنه های زیر بود: هنر(3)، زیست شناسی(4)، محیط(5)، دارو(6)، علم(7)، و جهان(8). به منظور دریـافت مترادفهای این کلمات برای (مثال مترادف هنر عکاسـی(9) است) از شـبکه ورد(10) استفاده کرده ایم و کیسه هایی از واژه های معرف زیر دامنه ساختیم.

این کیسه هـای واژه برای ردیابی خودکار واسط نام جست وجوی آر.اس.اس.4 و برای سـاخت برای هر ،دامنه فهرستی از یو آرال های فید مورد استفاده بود [بار] معنایی واژه ها به ما امکان تمرکز جست وجو برای فیدها و شناسایی صفحه های وبی را داد که به عنوان یک موضوع خاص تلقی می شدند. بازتاب در رسانه دامنه های مورد علاقه را میتوان از طریق چشمهای فید گرفتار شده به این شیوه – که معمولاً برخی از الگوها و ویژگیهای خاص را ارائه میدهد – مشاهده کرد.

هـدف از این انتخـاب نیمه خودکـار به دست آوردن درک و بینش کلی نسـبت به تنوع ،فیـدها، برحسب فرمتهـا الگوهـای به روز رسانی و ساختارهـای متنوع صـفحه هـای وب متناظر بود. علاوه بر این، به منظور حصول اطمینان از پوشـش سیسـتم عاملهای وب نوشت رایـج و همچنین نتایـج بیشتر خبرگرای بازگردانده شده توسط نام جست و جوی آر.اس.اس 4 تعـدادی از سایتهای وب نوشت به صورت دستی از فهرست بهترین وب نوشت [24] انتخاب شدند. به این ترتیب، فهرستی از حدود 400 سایت به دست آمد که

1- Search4RSS

2- URL

3- Art

4- Biology

5- Environment

6- Medicine

7- Science

به طور سیستماتیک خزش میشدند (این عدد پوشش انواع فیدهای وب است که بدون هیچ گونه زیر ساخت درگیر آرشیو مدیریت پذیر هستند). در پایان دوره ،خزش متوجه شدیم که برخی از انواع فیدها هرگز به روزرسانی نشده برخی دیگر ناپدید شده و برخی را نتوانستیم تجزیه کنیم. با فیلتر کردن ،آنها آرشیوی از 340فید فعال و صفحه های مرتبط با آنها را به دست آوردیم.

فید وب به یک صفحه وب اولیه (کانال) است که معمولا یا صفحه اصلی سایت یا هابی است که ما میتوانیم از آن پیوندهای اطلاعات ارائه شده به شکل مقاله های وب را پیدا کنیم. بقیه فید، آیتمهای فید فردی مربوط به مقاله های جدید و یا به روزرسانی شده را توصیف میکند برای هر دامنه برای هر سایت پیگیری شده فید و منابع مرتبط – به طور عمده صفحه کانال و صفحه های وبی – را ذخیره کرده ایم که با هر آیتمی اشاره شده بودند به منظور از دست ندادن آیتمهای جدیدی که بتوان در فید اضافه کرد، اجرای این خزشگرها دو بار در روز (علاوه بر اجرای خزشگرهای همان فیدها که روزانه توسط آرشیو وب اروپا اجرا میشود) انجام گرفت.

2-3. ویژگیهای فید وب برای تجزیه و تحلیل ،فید از ادی(1) [7] – کتابخانه تجزیه فید برای جاوا(2) – استفاده کرده ایم، که بر اساس تجزیه مبتنی بر سکس(3) قادر به تجزیه حتی دنیای واقعی به شکل ایکس ام ال است. ادی، از فرمتهای آر.اس.اس استاندارد ،اتم و آر.دی.اف(4) برای ،فیدها پشتیبانی می.کند ساختار داده فید بازگردانده شده توسط این تجزیه را میتوان برای استخراج همه نوع اطلاعات مفید در مورد کانال و آیتمهای تشکیل دهنده مورد تفحص قرار داد به طور خاص جهت ،کانال زبان شبیه ،فراداده شعار شرح و عنوان و همچنین برای هر مورد دیگر به علاوه نویسنده و مقوله هایی که در آن مقاله طبقه بندی شده است.

نوع فید: اجازه دهید نخست نگاهی به انواع فرمتهای فید که در مجموعه داده هایمان برشمردیم، داشته باشیم همانطور که در جدول 1 نشان داده شده است بیشتر فیدها از گویش آر.اس.اس. 2/0 استفاده می میکنند در حالی که اقلیتی هم وجود دارد که از اتم یا آر.دی.اف استفاده میکنند از آر.اس. اس. 0/91، تنها یک بار در میان 340 فید استفاده شد و از آر.اس.اس 1/0 هرگز استفاده نشد، که ممکن است منسوخ شدن این دو فرمت فید را نشان دهد با این حال اینکه این اعداد نیز به دلیل استفاده از جست وجوی آر.اس.اس.4، به عنوان منبع اصلی ما برای فیدها دچار تورش شوند، کاملاً امکان پذیر است.

تعداد آیتمها: تعداد تعداد آیتمهای ارائه شده در یک فید مفروض را بررسی کردیم. در واقع هر چند به طور نظری این امکان وجود دارد برای فید به همه آیتمهای منتشر شده قبلی اشاره کند به ندرت برای محدود کردن اندازه فید وب حاصل استفاده میشود در واقع بسیاری از فیدها کوتاه شده، تنها جدیدترین آیتم k برای یک مفروض را ارائه میدهد در شکل 1 هیستوگرام تعداد آیتمها به ازای هر فید در مجموعه داده

1- Eddie

2- Java

3- SAX

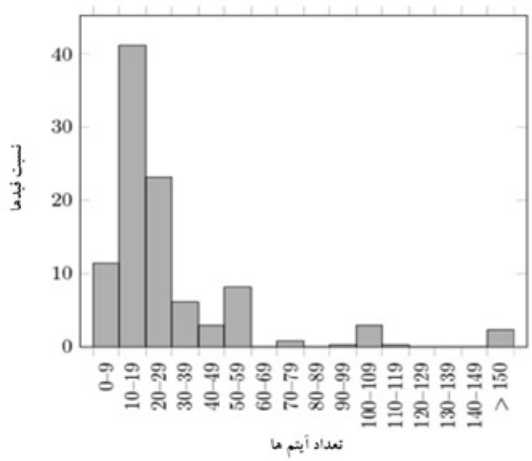
4- RDF

را نشان میدهیم. حدود 75 درصد فیدها، اطلاعاتی حدود کمتر از 30 آیتم را در یک زمان ارائه می دهند. سایر پیکهای مشاهده شده در شکل 1 بـا مقادیر جادویی» K = 50 و K = 100 توضـیح داده می شوند. k اگر فیـدی تنها شامل 10 آیتم مرسوم ترین عـدد باشد، بدان معنی است که با دو بار در روز خزش آن، ما توانسته ایم روزانه حداکثر 20 مقاله جدید را ذخیره کنیم همانطور که خواهیم دید تعداد قلیلی از فیـدها با یک فرکانس به روزرسانی بالاتر از آن وجود دارد، که به همین دلیل برخی از به روزرسانیها در خزش ما در واقع فیدها باید بیشتر از دو بار در روز خزش شوند از دست میروند.

اطلاعات موقتی: در خصوصیات آر.اس.اس به همین ترتیب در سایر فرمتهای فید) اطلاعات

عکس

را نشان می‌دهیم. حدود ۷۵ درصد فیدها، اطلاعاتی حدود کمتر از ۳۰ آیتم را در یک زمان ارائه می‌دهند. سایر پیک‌های مشاهده شده در شکل ۱ با مقادیر «جادویی» k = ۵۰ و k = ۱۰۰ توضیح داده می‌شوند. اگر فیدی تنها شامل ۱۰ آیتم (مرسوم‌ترین عدد) باشد، بدان معنی است که با دو بار در روز خزش آن، ما توانسته‌ایم روزانه حداکثر ۲۰ مقاله جدید را ذخیره کنیم. همانطور که خواهیم دید، تعداد قلیلی از فیدها با یک فرکانس به‌روزرسانی بالاتر از آن وجود دارد، که به همین دلیل، برخی از به‌روزرسانی‌ها در خزش ما (در واقع، فیدها باید بیشتر از دو بار در روز خزش شوند) از دست می‌روند.



شکل ۱. تعداد آیتم‌ها به‌ازای هر فید در مجموعه داده‌ها

اطلاعات موقتی: در خصوصیات آر.اس.اس (به همین ترتیب در سایر فرمت‌های فید)، اطلاعات زمانی را می‌توان از طریق عناصر lastBuildDate ، TTL، و updateFrequency برای کانال، و pubDate و lastModified برای آیتم‌ها به‌دست آورد. از طریق آزمایش‌ها، ما مشاهده کرده‌ایم که اگرچه pubDate مؤلفه‌ای اختیاری است، اما در اکثریت قریب به اتفاق فیدها ارائه می‌شود. این در مورد انواع دیگر عناصر مرتبط به زمان (timerelated) مذکور صدق نمی‌کند، هر چند lastBuildDate را می‌توان به‌نحوی به‌عنوان تاریخ انتشار جدیدترین آیتم استنباط کرد. اهمیت این مشاهدات از آن روست که نشان می‌دهد که فیدها را می‌توان برای تعیین زمانی که داده جدیدی به یک کانال اضافه می‌شود، مورد استفاده قرار داد و در تشخیص تغییر کمک‌رسان است. با تجزیه و تحلیل یک فید برای یک دوره زمانی، می‌توانیم الگوها را در انتشار راهبرد، شناسایی و به‌طور خودکار خزش را با آن منطبق کنیم.

بازه به‌روزرسانی: تمام تاریخ انتشارات مربوط به آیتم‌های ظاهر شده که در طول دوره آزمایش را جمع‌آوری کرده‌ایم؛ از آنجا که هر آیتم یک تاریخ انتشار دارد، تعداد تاریخ‌های انتشار برابر است با آیتم‌ها. ما به طیف وسیعی از بازه‌های به‌روزرسانی بین دو انتشار، و همچنین نشانه‌های وجود راهبرد انتشار منظم

زمانی را میتوان از طریق عناصر last BuildDate ، TTL و updateFrequency برای کانال و pubDate و lastModified برای آیتم ها به دست آورد از طریق آزمایشها ما مشاهده کرده ایم که اگر چه pubDate مؤلفه ای اختیاری است اما در اکثریت قریب به اتفاق فیدها ارائه میشود این در مورد انواع دیگر عناصر مرتبط به زمان (timerelated) مذکور صدق نمی کند هر چند lastBuildDate را میتوان به نحوی به عنوان تاریخ انتشار جدیدترین آیتم استنباط کرد اهمیت این مشاهدات از آن روست که نشان میدهد که فیدها را میتوان برای تعیین زمانی که داده جدیدی به یک کانال اضافه میشود مورد استفاده قرار داد و در تشخیص تغییر کمک رسان است. با تجزیه و تحلیل یک فید برای یک دوره زمانی، می توانیم الگوها را در انتشار راهبرد شناسایی و به طور خودکار خزش را با آن منطبق کنیم.

بـازه به روزرسانی تمام تاریـخ انتشارات مربوط به آیتمهای ظاهر شـده که در طول دوره آزمایش را جمع آوری کرده ایم؛ از آنجا که هر آیتم یـک تاریـخ انتشار دارد تعـداد تاریخهای انتشار برابر است با آیتمها. ما به طیف وسـیعی از بازه های به روزرسانی بین دو انتشار و همچنین نشانه های وجود راهبرد انتشار منظم

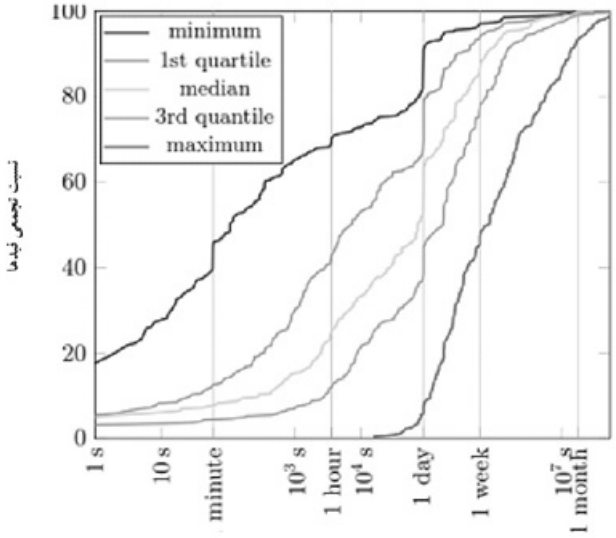
ص: 13

برای فید توجه زیادی میکنیم در شکل ،2 متوسط بازه به روزرسانی بین دو انتشار هر فید را به صورت نمودار تجمعی (به رنگ سبز در وسط نشان داده ایم. توجه داشته باشید که محور x مقیاس لگاریتمی .است شکل 2 نشان میدهد که به عنوان مثال 20 درصد فیدها یک بازه به روزرسانی متوسط کمتر از یک ساعت و حدود 10 درصد فیدها یک بازه به روزرسانی متوسط دقیقاً یک روزه دارند که مربوط به فیدهایی است که به طور منظم و خودکار هر روز به روز رسانی میشوند. در سطح جهانی، مهم که توجه داشته باشید که هیچ فاصله به روز رسانی نمونه واری وجود ندارد و حتی بدون در نظر گرفتن موارد شدید میتواند به محدوده کمتر از یک ساعت تا به بیش از یک هفته برسد. همچنین شکل 2 مقادیر دیگر فاصله به روز رسانی هر فید را نشان میدهد که کمک میکند تنوع الگوهای به روزرسانی برای یک فید مفروض را بتوان دید به این ترتیب، حتی اگر چه 60 درصد فیدها دارای یک بازه به روزرسانی متوسط یک روز یا کمتر بودند، کمتر از حدود 10 درصد آنها همیشه حداقل یک به روزرسانی در هر روز، و بیش از 90 درصد آنها حداقل به روزرسانی روزانه در دوره مشاهده داشتهاند. در واقع، شکاف بزرگی بین مقادیر بازه به روزرسانی متوسط و مقادیر حداقل و حداکثر وجود دارد که باعث بروز مشکل پیش بینی به روزرسانی بعدی یک فید مفروض میشود.

نسبت تجمعی فیدها

عکس

برای فید توجه زیادی می‌کنیم. در شکل ۲، متوسط بازه به‌روزرسانی بین دو انتشار هر فید را به‌صورت نمودار تجمعی (به رنگ سبز، در وسط) نشان داده‌ایم. توجه داشته باشید که محور x مقیاس لگاریتمی است. شکل ۲ نشان می‌دهد که به‌عنوان مثال، ۲۰درصد فیدها یک بازه به روزرسانی متوسط کمتر از یک ساعت، و حدود ۱۰درصد فیدها یک بازه به‌روزرسانی متوسط دقیقاً یک روزه دارند، که مربوط به فیدهایی است که به‌طور منظم و خودکار، هر روز، به روز رسانی می شوند. در سطح جهانی، مهم است که توجه داشته باشید که هیچ فاصله به روز رسانی نمونه‌واری وجود ندارد، و حتی بدون در نظر گرفتن موارد شدید می‌تواند به محدوده کمتر از یک ساعت تا به بیش از یک هفته برسد. همچنین شکل ۲، مقادیر دیگر فاصله به‌روزرسانی هر فید را نشان می‌دهد، که کمک می کند تنوع الگوهای به‌روزرسانی برای یک فید مفروض را بتوان دید: به این ترتیب، حتی اگرچه ٦۰ درصد فیدها دارای یک بازه به‌روزرسانی متوسط یک روز یا کمتر بودند، کمتر از حدود ۱۰ درصد آنها همیشه حداقل یک به‌روزرسانی در هر روز، و بیش از ۹۰ درصد آنها حداقل به‌روزرسانی روزانه در دوره مشاهده داشته‌اند. در واقع، شکاف بزرگی بین مقادیر بازه به‌روزرسانی متوسط، و مقادیر حداقل و حداکثر وجود دارد، که باعث بروز مشکل پیش‌بینی به‌روزرسانی بعدی یک فید مفروض می‌شود.



شکل ۲. نسبت تجمعی فیدها با مقادیر یک چهارم مفروض فاصله به‌روزرسانی‌ها

در جدول ۲ برخی آمارهای دیگر در مورد فواصل به‌روزرسانی در سطح دامنه را نشان می‌دهیم. برای هر دامنه، میانگین فاصله زمانی به‌روزرسانی متوسط به‌عنوان انحراف استاندارد ادغام شده فواصل به‌روزرسانی داده‌ایم، که روشی است مبتنی بر آمار جهت خلاصه کردن انحراف از اتحاد مجموعه‌ای از

در جدول 2 برخی آمارهای دیگر در مورد فواصل به روزرسانی در سطح دامنه را نشان میدهیم برای هر ،دامنه میانگین فاصله زمانی به روزرسانی متوسط به عنوان انحراف استاندارد ادغام شده فواصل به روزرسانی داده ایم که روشی است مبتنی بر آمار جهت خلاصه کردن انحراف از اتحاد مجموعه ای از

ص: 14

اعـداد. همان طور که دیـده میشود تغییرات زیادی در میان دامنه ها وجود دارد که نشانه دیگری از عدم فاصـله زمانی به روزرسانی معمولی است همچنین در اینجـا انحراف اسـتاندارد بسـیار بالاـ را در برخی دامنه هـا خـاطر نشـان میکنیم به خصـوص یـک دامنه مفروض میتواند وبگاههای با ماهیت بسـیار متفاوت ،اخبار، وب نوشـتها نوشـته های و یکی و ماننـد آن را نشان دهنـد در دامنه ،هنر به عنوان مثال مشاهده کرده ایم که سایتهای مختلفی هسـتند که مقاله های کوچک در مورد نقاشی یا عکس از رده 100 ورودی در هر روز چاپ میکنند بنابراین مفهوم آیتم متفـاوت از یک مقاله تخصصـی است که ممکن است حاوی متنهای(مثل آیتم خبری نسـبت به یک مقاله ای باشـد که به طور انحصاری از تصاویر و یا فیلم ها تشـکیل شده است. هر چه رده وب نوشـتهای محبوب دسته بندی همگن ساختاری داشته باشد، انحراف معقول تری از فواصل به روزرسانی را دارا خواهد بود.

در اینجـا، مطـالعه مجموعه دادههـا را نتیجه گیری میکنیم که در یک اقـدام خاص منعکس کننـده وضـعیت موجود و تنوع فیـدهای وب به منظور گردش به سمت بحث و گفت و گو راجع به روش ما برای استخراج اشیای داده ای وب است.

4 – استخراج اشیای داده ای

ما در این بخش الگوریتم را برای پیـدا کردن شـیء داده ای در یک صـفحه وب مفروض بحث میکنیم شـیئ که مربوط به یک آیتم فیـد وب است.

4-1 جمع آوری اطلاعات معنایی

نخست از آیتمهای فیدی نوعی بافت معنایی ایجاد کردیم که جهت اسـتخراج شئ متناظر عمل میکند آیتمهای فید سه مؤلفه اجباری دارند پیوند عنوان و شرح.

پیوند. پیوند به ما یو آر.ال. صفحه وبی را میدهد که در آن شیء داده ساکن است.

عنوان: آیتم فید باید دارای عنوان باشد، که معمولاً یک متن کوتاه به منظور توصیف محتوای یک

مقاله است.

شرح بیشتر اوقات شـرح شامل کل محتوای مقاله نمی شود بلکه تنها چند خط اول آن که دراچ .تی.ام.ال برای اهداف ارائه کدگذاری شده همراه بـا پیونـدی در انتهای ماننـد ادامه مطلب ...» صـفحه اصـلی وب در موارد دیگر شـرح تنها شامل یک جمله می شود که به جای در برگرفتن چنـد خط اول آن ، مقاله را خلاصه می.کند دقت هر چه باشد میتوانیم برخی معناهای قابل اعتماد در مورد مقاله را با بهره گیری از این شرح استخراج کنیم.

با بازیابی تمامی مطالب ،متنی از عنوان و شرح شروع میکنیم کد اچ تی ام ال شرح علامت گذاری و فقط متن نگه داشته میشود توالی های منتج لغات به دو نوع از نهادهای معنایی تبدیل میشوند مفاهیم و ان–گرمها(1).

برای به دست آوردن مفاهیم کلمات را لب خوانی (tokenize) و ریشه آن را قطع کرده، و براساس فراوانی lexemes مرتب کرده (به عنوان اندازه اهمیت در نظر گرفته شدند) و فقط آنها برجسته ها را نگه داشتیم تا ،حدی مفهوم شبیه یک برچسب است چون اصطلاحی است که یک شیء داده ای را توصیف کند. در واقع کلماتی را که از مفاهیم می آیند میتوان به عنوان کلید واژه های جست وجو در یک صفحه ،وب به منظور شناسایی منطقه شیء داده ای به کار برد اما میتوان توجه را به مناطق دیگری نیز معطوف داشت که سرشار از مفاهیم هستند مانند مناطقی که حاوی آرا یا دسته بندیهایی برای یک مقاله وب مفروض هستند.

به همین دلیل تمرکزمان را به آن – گرم بر می گردانیم آن گرم در زمینه ما معرف دنباله عناوین n است که از عنوان و ،شرح همانگونه که ظاهر میشوند گرفته میشوند گزینه برای n، مصالحه ای بین مثبت های کاذب و منفیهای کاذب در فرآیند استخراج بوده و بیشتر مورد بحث قرار خواهد گرفت.

2-4. استخراج

عکس

برای به‌دســـت آوردن مفاهیم، کلمات را لب‌خوانی (tokenize) و ریشه آن را قطع کرده، و براساس فراوانی lexemes مرتب کرده (به‌عنوان اندازه اهمیت در نظر گرفته شدند) و فقط آنها برجسته‌ها را نگه داشتیم. تا حدی، مفهوم شبیه یک برچسب است، چون اصطلاحی است که یک شیء داده‌ای را توصیف می‌کند. در واقع، کلماتی را که از مفاهیم می‌آیند می‌توان به‌عنوان کلید واژه‌های جست‌وجو در یک صفحه وب، به‌منظور شناسایی منطقه شیء داده‌ای به‌کار برد، اما می‌توان توجه را به مناطق دیگری نیز معطوف داشت که سرشـــار از مفاهیم هستند، مانند مناطقی که حاوی آرا یا دسته‌بندی‌هایی برای یک مقاله وب مفروض هستند.

به‌همیـــن دلیل، تمرکزمان را به ان ـ گرم، برمی‌گردانیم. ان ـ گرم در زمینه ما معرف دنباله عناوین n است، که از عنوان و شرح، همانگونه که ظاهر می‌شوند، گرفته می‌شوند. گزینه برای n، مصالحه‌ای بین مثبت‌های کاذب و منفی‌های کاذب در فرآیند استخراج بوده و بیشتر مورد بحث قرار خواهد گرفت.

جدول ۲ . آمارهای فید بهازای هر دامنه

Domain	Number of feeds	Average mean update interval	Pooled standard derivation of update interval
Art	87	12 days, 14 hours, 12 min	82 days, 6 hours, 32 min
Biology	80	7 days, 13 min	8 days, 17 hours, 43 min
Blogs	29	15 hours, 35 min	8 hours, 39 min
Environment	7	19 hours, 49 min	4 days, 15 hours, 18 min
Medicine	8	3 days, 19 hours, 16 min	1 day , 22 hours, 43 min
Other	13	4 days, 16 hours, 48 min	4 days, 19 hours, 46 min
Science	112	22 days, 12 hours, 45 min	14 days, 21 hours, 35 min
Universe	4	4 hours, 44 min	7 hours, 5 min
Total	340	12 days, 15 hours, 17 min	37 days, 16 hours, 49 min

٤-۲. استخراج

در اینجا نوعی الگوریتم از پایین به بالا را نشـــان می‌دهیم، که با توجه به فید، آیتم‌ها را شناسایی کرده و برای هر آیتم فید مؤلفه لفاف بســـته‌بندی شیء داده‌ای را با تطبیق دادن ان ـ گرم در برابر محتوای متنی گره‌های برگی استخراج شـــده از صفحه وب اچ.تی.ام.ال. می‌یابد. این الگوریتم در الگوریتم ۱ خلاصه شده است.

نخست مفهوم گره مفهومی را معرفی می‌کنیم:

تعریف ۱.گره مفهومی یک گره برگی (گره بدون فرزند) است که در مفهوم متنی‌اش شامل حداقل یک مفهوم (یا ان ـ گرم) از عنوان و شرح آیتم باشد.

ما تمام گره‌های برگی صفحه را استخراج و برای هر یک، تراکم معنایی ایجاد می‌کنیم.

تعریف ۲. تراکم معنایی یک گره مفهومی به‌عنوان تعداد مفاهیم همسان (یا ان ـ گرم) تقسیم بر طول محتوای متنی گره‌های مربوط تعریف می‌شود.

ما گره‌های مفهومی را طبق نزدیک‌ترین جد(نیا) طبقه‌بندی می‌کنیم که یک مؤلفه در ســـطح بلوک است. یک غیرمستدل مؤثر در واقع گرفتن نزدیک‌ترین جدهاست (نیاهاست)که یک مؤلفه div می‌باشد،

در اینجا نوعی الگوریتم از پایین به بالا را نشان میدهیم که با توجه به ،فید آیتمها را شـناسایی کرده و برای هر آیتم فید مؤلفه لفاف بسته بندی شیء داده ای را بـا تطـبیق دادن آن – گرم در برابر محتـوای متنی گره هـای برگی اســتخراج شـده از صـفحه وب اچ.تی.ام.ال. میبابـد. این الگوریتم در الگوریتم 1 خلاصه شده است.

نخست مفهوم گره مفهومی را معرفی میکنیم:

تعریف 1.گره مفهومی یک گره برگی (گره بدون فرزند) است که در مفهوم متنیاش شامل حداقل

یک مفهوم (یا آن – گرم) از عنوان و شرح آیتم باشد.

ما تمام گره های برگی صفحه را استخراج و برای هر یک تراکم معنایی ایجاد میکنیم.

تعریف2. تراکم معنایی یک گره مفهومی به عنوان تعداد مفاهیم همسان یا ان – گرم تقسیم بر طول

محتوای متنی گرههای مربوط تعریف میشود.

ما گره های مفهومی را طبق نزدیکترین جد (نیا) طبقه بندی میکنیم که یک مؤلفه در سطح بلوک است. یک غیر مستدل مؤثر در واقع گرفتن نزدیکترین جدهاست نیاهاست که یک مؤلفه divمیباشد،

چون در آزمایشـهای مان مشاهـده کرده ایم که یک شـیء داده تقریباً همیشه در یک مؤلفه div محـدود است. پس از این تجزیه و تحلیل میتوانیم بگوییم که کدام یک گره های مفهومی هسـتند که همان جد را به اشتراک می.گذارند فهرست اجداد به ما مناطق معنایی صـفحه را میدهد که توسط گره های معنایی در سطح کد مدل شده اند.

تعریف .3 گره معنایی پایینترین جـد مشترک سـطح بلوک مجموعه ای از گره های مفهومی است. به منظور روشن شـدن اینکه کدام یک از گرههای معنایی نشان دهنده لفاف بسته بندی مقاله است، اندازه تراکم معنایی زیر را برای هر کدام محاسبه کرده و گرهی را در نظر میگیریم که بزرگترین مقدار را برای آن دارد.

تعریف .4 گره لفاف بسته بندی شیء داده ای گرهی معنایی است که حاوی بیشترین تعداد گره های متراکم مفهومی است.

در شـرح قبلی اجازه اسـتفاده از مفاهیم یا آن – گرمها برای پیدا کردن گره های مفهومی و محاسـبه تراکم معنایی را داده ایم هنگام همسان کردن با شـروطی که متناظر با مفاهیم است، تعـداد مناطق معنایی افزایش خواهد یافت در حالی که همسان کردن با آن – گرمها به وضوح این تعـداد را کاهش خواهد داد. دلیل وقوع این اتفاق این است که آن گرمها نسـبت به مفاهیم نسـبت به محتوای مقاله معنی دارتر هسـتند. اضافه بر این در بعضـی موارد انتخاب ان – گرم بیش از حـد محـدود کننده است. این امر کاملاً به ندرت محدودکننده اتفاق میافتد، بیشتر زمانی که شـرح بیشتر مقاله را با کلمات مختلف خلاصه می کند، به جای اینکه چند خط اول آن را ارائه دهد. در نتیجه به منظور تشخیص این نوع موارد که در آن ان–گرمها یک گزینه نیستند مفهوم ثبات گره معنایی را معرفی میکنیم.

تعریف 5. یک گره از لحاظ معنایی سازگار است اگر متن آن حاوی یک نسبت بزرگی از مفاهیم به

دست آمده از عنوان و شرح آیتم باشد.

ما میگوییم یک نسـبت بزرگی (در عمل 0/5)از مفاهیم زیرا لازم نیست حضور همه را به منظور اثبات یک گره لفاف بسـته بنـدی بررسـی کرد. از سوی دیگر، زمـانی که نـامزد گره جـد شامل نیمی از مفاهیم موجود نباشـد ممکن است گمان بریم که آن لفاف بسـته بنـدی مقاله نیست. اگر این اتفاق بیفتـد میتوان نتیجه گرفت که آن گرمها به علت نقص مفروضات کسب معنایی مؤثر نبودنـد و در نتیجه مقدار n در آن گرمها را کاهش داده و روش گره لفاف بسـته بندی را تکرار میکنیم در آزمایشهای انجام شده با n – 3 (هدف به دست آوردن تعداد مناسب ان – گرمهـای قابـل توجه است) شـروع کردیم که بهـترین نتیجه را داد و در صورت شـکست، به طور مسـتقیم سـعی در همسـان کردن با مفـاهیم .داشتیم به طور کلی اشـیای داده ای مقاله های خبری، پسـتهای وب نوشت دارای نظرات مرتبط هسـتند. ما میخواهیم بین نظرات یک مقاله و خود مقاله به دلایل زیر تمایز روشنی قائل شویم:

1–از نظر مفهومی اطلاعات مورد نظر در مقاله همان مواردی نیست که در موردش نظر داده شده است.

2–خزش مقاله باید از خزش آرا از هم تفکیک شود هر زمان که یک نظر اضافه شد، مقاله را باید

تغییر یافته در نظر گرفت یک خزش جدید لازم است و شیء حاصله میتواند در مقایسه با نسخه های قبلی آن خیلی اضافی باشد؛ در عوض این مقاله باید به آرای خود اشاره داشته، جداگانه ردیابی شده، و در به روزرسانی هماهنگ باشد.

معمولاً وب مستر مراقبت افزودن یو آر.ال. در فیدی است که می تواند برای پیگیری آرا مورد استفاده قرار گیرد، اما وقتی که این مورد نیست میتوانیم منطقه آرا را با استفاده از غیر مستدلهای در الگوریتم شناسایی کنیم.

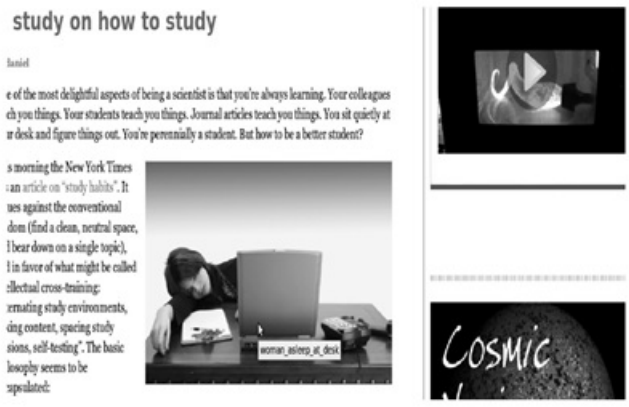
عکس

۱۸ مدیریت منابع اطلاعاتی وب

تغییریافته در نظر گرفت، یک خزش جدید لازم است و شیء حاصله می‌تواند در مقایسه با نسخه‌های قبلی آن خیلی اضافی باشد؛ در عوض، این مقاله باید به آرای خود اشاره داشته، جداگانه ردیابی شده، و در به‌روزرسانی هماهنگ باشد.

۳- معمولاً، وب مستر مراقبت افزودن یو.آر.ال. در فیدی است که می‌تواند برای پیگیری آرا مورد استفاده قرار گیرد، اما وقتی که این مورد نیست، می‌توانیم منطقه آرا را با استفاده از غیرمستدل‌های در الگوریتم شناسایی کنیم.

الگوریتم ۱. استخراج اشیا داده‌ای



شکل ۳. یک نمونه مقاله وبی و آیتم داده‌ای مرتبط با فید

-/http://feedproxy.google.com/r/Cosmic Variance Blog/3

/uatEVOIO0g

/http://blogs.discovermagazine.com

cosmicvariance/2010/09/07/a-study-on-how-to-study/comments

Wed, 08 Sep 2010 03:16:54 +0000

daniel

?/http://blogs.discovermagazine.com/cosmicvariance

p=5353

scientist is that you8217;re always learning. Your colleagues teach you

.things. Your students teach you things. Journal articles teach you things

You sit quietly at your desk and figure things out. You8217;re perennially

a student. But how to be a better student? This morning the New

<[[[…] York

One of the most delightful aspects

<[[<….of being a scientist is that you8217;re always learning

/http://blogs.discovermagazine.com/cosmicvariance

/a-study-on-how-to-study/feed/2010/09/07

6

/http://blogs.discovermagazine.com/cosmicvariance

/a-study-on-ow-to-study/2010/09/07

>>

3-4. یک مثال

به منظور شرح و توضیح بیشتر ما مثالی از یک صفحه وب را که قسمتی از آن در شکل 3 ارائه شده است در نظر خواهیم گرفت صفحه وب با استفاده از تمایز دهنده اچ تی ام ال پاک و خوب فرمت خواهد شد. این مرحله به منظور انتخاب معقول گرههای برگی درخت dom در صفحه لازم است (و برای

ص: 19

تجزیه و تحلیل تنها آنهایی نگهداری میشود که حداقل یک واژه معنایی را در بردارند).

مشاهده میکنیم که عنوان این مقاله در آیتم ،فید و در دو خط اول متن وجود دارد که در شرح فید کد گذاری شده است. برچسب زمانی نیز وجود دارد که مربوط به تاریـخ انتشار است. تگ مؤلفه ای نیست که به طور متـداول ظاهر شود بنابراین، ما در الگوریتم خود در مورد مفید بودنش فرض ایجاد نمیکنیم.

عکس

تجزیه و تحلیل تنها آنهایی نگهداری می‌شود که حداقل یک واژه معنایی را در بردارند).

مشـــاهده می‌کنیم که عنوان این مقاله در آیتم فید، و در دو خط اول متن وجود دارد، که در شـــرح فید کد گذاری شـــده اســـت. برچســـب زمانی نیز وجود دارد که مربوط به تاریخ انتشـــار است. تگ <content:encoded> مؤلفه‌ای نیست که به‌طور متداول ظاهر شود، بنابراین، ما در الگوریتم خود در مورد مفید بودنش فرض ایجاد نمی‌کنیم.

در شکل ٤، مشاهده می‌کنیم که عنوان مقاله همچنین در سمت راست منطقه آرا وجود دارد، یعنی جایی‌که برخی از تازه‌ترین مقاله‌های وبگاه نیز ارائه شده است. عنوان تنها برای یک اشاره کافی نیست: بلکه به شرح آیتم هم نیاز داریم. برخی نمونه‌های تصادفی ان ـ گرم‌ها از شرح عبارت‌اند از «جنبه‌های توضیحی[1]» (n=۲)، «دانشـــمند بودن[2]» (n=۳) ، «این صبح، جدیـــد[3]» (n=٤). به‌طور کلی ان ـ گرم‌ها فرصت/ احتمال کمتری برای خوشـــه‌بندی زیاد در ســـایر مناطق صفحه‌هـــای وب دارند. هرچه توالی بزرگ‌تر باشد، به همان اندازه سرعت شناسایی مقاله بیشتر خواهد بود.

شکل ٤. توضیحی برای این واقعیت که عنوان به تنهایی برای شناسایی منطقه مقاله وب کافی نیست.

1. Delightful Aspects
2. Being a Scientist
3. This morning the New

در شکل ،4 مشاهده میکنیم که عنوان مقاله همچنین در سـمت راست منطقه آرا وجود دارد، یعنی جایی که برخی از تازه ترین مقاله های وبگاه نیز ارائه شـده است عنوان تنها برای یک اشاره کافی نیست بلکه به شـرح آیتم هم نیاز داریم برخی نمونه های تصادفی آن – گرمها از شرح عبارت اند از «جنبه های توضیحی(1)» (n–2)،«دانشمند بودن(2)»(n–3) ،«این صبح جدید(3)»(n–4). به طور کلی آن – گرمها /فرصت/ احتمال کمتری برای خوشه بنـدی زیاد در سایر مناطق صـفحه های وب دارنـد هر چه توالی بزرگتر باشـد به همان اندازه سـرعت شناسایی مقاله بیشتر خواهد بود.

شکل 4.توضیحی برای این واقعیت که عنوان به تنهایی برای شناسایی منطقه مقاله وب کافی نیست.

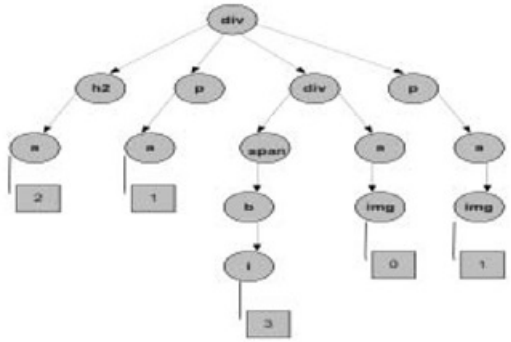
1- Delightful Aspects

2- Being a Scientist

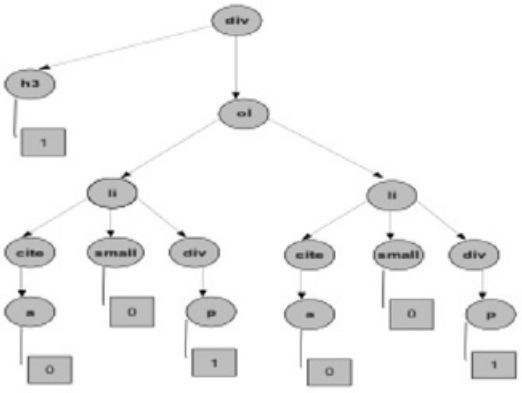
3- This morning the New

شکل .5 زیر شاخه DOM ساختگی (به منظور درک و توضیح بیشتر) متناظر با مقاله وب. مقدار الصاق شده به گره برگی معرف تعداد مفاهیم در برگرفته در متن اصلی است.

عکس



شکل ۵. زیر شاخه DOM ساختگی (به‌منظور درک و توضیح بیشتر) متناظر با مقاله وب. مقدارالصاق شده به گره برگی معرف تعداد مفاهیم در برگرفته در متن اصلی است.



شکل ٦: زیر شاخه ساده شده DOM متناظر با منطقه نظری‌ها. متوجه شدیم که این منطقه به‌عنوان یک فهرست (مرتب شده با آن) ساماندهی شده و همچنین مفاهیم (یا ان ـ گرم‌ها) مقاله را، در نسبت کوچک‌تر نشان می‌دهد.

شکل :6 زیر شاخه ساده شـده DOM متنـاظر بـا منطقه نظریه هـا متوجه شـدیم که این منطقه به عنوان یـک فهرست (مرتب شـده با آن) ساماندهی شده و همچنین مفاهیم (یا آن – گرمها) مقاله را، در نسبت کوچک تر نشان میدهد.

برعکس وقتی شـرح مقاله خیلی کوتاه باشد و یا فرمول بندی ایده وجود داشته باشد، استفاده از مفاهیم گزینه بهتری خواهد بود در این مثال ،خاص الگوریتم زوج مقاله را فقط با این صـبح جدید شـناسایی می.کند این اتفاق به دلیل اینکه این صـبح جدید یک آن – گرم منحصر به فرد در صفحه وب است، رخ می دهد.

استفاده تنها از جنبههای توضیحی یا دانشمند بودن مقاله را شناسایی نخواهد کرد، چون آن گرمها

نیز در اولین نظر ظاهر خواهند شد (شکل 6 ، div/ol/li/div/p).

بنابراین نمیتوانیم فرض کنیم که توالی ها منحصر به فرد هستند و یا اینکه میتوانیم آنها رامنحصر

به فرد کنیم. پس ما نمونه های مختلفی از عنوان و شرح را در نظر میگیریم.

برای توضـیح اصل الگوریتم روند تطبیق آن – گرمهای دو گره معنایی توضـیح داده شده در شکله و 6 را در نظر میگیریم در عمل برای این صفحه وب الگوریتم سه گره معنـایی ممکن را که در واقع ساختار پیچیـده تری دارنـد بر می گردانـد. همانطور که میبینیم گره های برگی در اولین زیر شاخه dom شـکل) (ه به لحاظ مفهومی غنی تر از دومی هستند (شکل6) ،بنـابراین اولین گره را انتخاب میکنیم چون با توجه به اندازه تراکم معنایییمان معنی دارترین است.

5.آزمایشها

به منظور اثبـات روایی رویکرد مـان برای اسـتخراج اشـیای داده ای به طور کامـل سیسـتم را پیـاده سازی کرده به منظور ارزیابی دقت ،آن آزمایشهایی را انجام داده ایم.

آزمایشـهایی با اسـتفاده از فیدهای جمع آوری شده بر اساس پاسخهای موتور جست وجوی آر.اس. اس 4 همان مجموعه دادههای ذکر شده در بخش 3 انجام شده است به یاد بسـپارید که مجموعه دادهها ، از لحاظ ساختار و نوع اشـیای دادهای خیلی متنوع .بود برای هر فید ما کانال ساختار داده را بازیابی کرده و روش استخراج را برای تمام اجزای آیتم آن اعمال کرده ایم.

به عنوان اولین ،آزمون تلاـش کرده ایم تـا منطقه صـفحه وب مرتبـط بـا عنوان آیتم را برگردانیم. بـا این حـال، این روش به چنـد دلیل نتایج ضعیفی ارائه میکند تطبیق عنوان ممکن است به خاطر ویژگیهای کدگذاری به طور کامل امکان پذیر نباشد یا اینکه ممکن است در چندین محل مختلف در صـفحه ظاهر شود (شـکل4) علاـوه بر این بـا توجه به محـل مؤلفه عنوان و سـطوح انبـاشتی آن در بلوکهای کـد اچ.تی .ام.ال.، شناسایی محدوده کل شیء داده ای کار آسانی نیست.

حال ما عملکرد الگوریتم را مقایسه میکنیم یعنی اینکه دقت اشـیای داده ای اسـتخراج شـده با [12] Boilerpipe که پیشرفته ترین روش برای شـناسایی محتوای اصلی یک صفحه در غیاب اطلاعات معنایی اضافی است. ما تأکید میکنیم که هر چه ما اطلاعات بیشتری نسبت به آنچه Boilerpipe بدان دسترسی دارد استفاده کنیم به دست آوردن دقت بیشتر از علاقه به این روش که کلی تر است – نمی کاهد.

مشاهده کردیم که نتیجه کارمان اغلب دقیقتر است چون تراکم متن در صـفحه وب را در نظر نمی گیریم بلکه انسـجام معنایی آن مطابق با آیتم فید را در نظر گرفتیم. مواردی وجود دارد که در آن گره

ممکن است حاوی مقدار زیادی متن باشد در حالی که ممکن است با توجه به اندازه گیری تراکم معنایی ما فاقد ارزش قضاوت شود.

علاوه بر این توجه داشته باشید که زمانی که صفحه وب حاوی مقاله های مختلف متوالی باشد، روش ما بین آنها تمایز قایل خواهد شد و مقاله خاص متناظر با یک آیتم را شناسایی خواهد کرد. در مقابل، Boilerpipe محتوای متنی تمام مقاله ها و یا تنها متراکم ترین نوع را بسته به مورد در بر خواهد گرفت.

جدول3. نتایج آزمایشها

عکس

ممکن است حاوی مقدار زیادی متن باشد در حالی که ممکن است با توجه به اندازه‌گیری تراکم معنایی ما فاقد ارزش قضاوت شود.

علاوه بر این، توجه داشته باشید که زمانی که صفحه وب حاوی مقاله‌های مختلف متوالی باشد، روش ما بین آنها تمایز قایل خواهد شـــد و مقاله خاص متناظر با یک آیتم را شناسایی خواهد کرد. در مقابل، Boilerpipe محتوای متنی تمام مقاله‌ها و یا تنها متراکم ترین نوع را، بسته به مورد در بر خواهد گرفت.

روش	استخراج های درست	دقت (درصد)
Our technique	1038/1314	79.0%
Boilerpipe	821/1314	62.5%

جدول۳. نتایج آزمایش‌ها

آزمایش‌ها برای ٦٠ سایت انتخابی تصادفی از مجموعه داده در دامنه هنر (اولین مورد خزش شده)، متناظر با تمام ١٣١٤ آیتم های فید انجام شد. ما به‌طور دستی مقاله وب را، نتیجه الگوریتم استخراج ما و نوع Boilerpipe [١٢] بررسی کردیم. ما فقط نتایج متنی استخراجی را مقایسه کردیم چون خروجی رایگان قابل استفاده پیاده سازی Boilerpipe بود. روش ما در واقع استخراج محتوای کل منطقه شناسایی شده شامل پیوند ها و تصاویر است.

نتایج عددی در جدول ٣ داده شـــده است. ما یک شـــیء را همانطور که به‌درستی استخراج شده زمانی که متن دقیقاً اســـتاندارد طلایی است، در نظر می گیریم که با توضیح دستی صفحه وب به دست آمده است. انطباق‌های جزئی نادیده گرفته شدند. دقت این الگوریتم (در حدود ٧٩ درصد) در مقایسه با Boilerpipe (در حدود ٦٢ درصد) رضایت‌بخش است. در نهایت، توجه داشته باشید که زمانی که روش ما با شکست مواجه شد، الگوریتم در منطقه غنی‌تر مفهومی را شناسایی می‌کند که یکی از اشیای داده‌ای است، که هنوز هم مربوط به مقاله است، اگرچه توسط آن شناسایی نشده است.

٦. بحث در مورد کاربردها

ما این مقاله را با بحث در مورد تعداد کاربردها به‌پایان می‌رسانیم که می‌توان در فرآیندآرشیو وب تلفیق کرد، و نیز آنی که روش استخراج اشیاء داده‌ای را استفاده می‌کند که ما پیشنهاد می‌کنیم.

ماندگاری آرشیوهای وب. زمان عامل خیلی تأثیرگذاری برای تقسیر محتوای خزش شده است.

درحالی‌که ممکن اســـت داده‌ها دست نخورده باقی بمانند، روشی که ما آن را درک و ارائه می‌کنیم متفاوت است، که دلیل عمده آن واقعیت خود زبان، فرهنگ، و وسایل فناوری تکامل بیان است. یکی از جدی‌ترین مشکلات برشمرده شده درآرشیو وب زمانی است که فرمت داده‌های خزش شده منسوخ و یا به‌طور کلی استفاده نشود. راه‌حل‌های ارائه شده توسط نویسندگان [٩]، [٢٧] و [٢٦] شبیه‌ساز نرم‌افزاری یا سخت‌افزاری، انتقال محتوا، یا شامل یک پروکسی است که قابلیت‌های ترجمه فرمت را ترکیب خواهد

آزمایشها برای 60 سایت انتخابی تصادفی از مجموعه داده در دامنه هنر (اولین مورد خزش شده)، متناظر با تمام 1314 آیتمهای فید انجام شد ما به طور دستی مقاله وب را نتیجه الگوریتم استخراج ما و نوع 12] Boilerie] بررسی کردیم ما فقط نتایج متنی استخراجی را مقایسه کردیم چون خروجی رایگان قابل استفاده پیاده سازی Boilerpipe بود روش ما در واقع استخراج محتوای کل منطقه شناسایی شده شامل پیوندها و تصاویر است.

نتایج عددی در جدول 3 داده شده است ما یک شـیء را همانطور که به درستی اسـتخراج شده زمانی که متن دقیقاً استاندارد طلایی است در نظر میگیریم که با توضـیح دستی صفحه وب به دست آمده است. انطباقهای جزئی نادیده گرفته شدند. دقت این الگوریتم در حدود 79

درصد در مقایسه با Boilerpipe در حدود 62 درصد رضایت بخش است در نهایت توجه داشته باشید که زمانی که روش ما با شکست مواجه شد الگوریتم در منطقه غنی تر مفهومی را شناسایی میکند که یکی از اشیای داده ای است که هنوز هم مربوط به مقاله است اگرچه توسط آن شناسایی نشده است.

6. بحث در مورد کاربردها

ما این مقاله را با بحث در مورد تعداد کاربردها به پایان میرسانیم که میتوان در فرآیند آرشیو وب تلفیق کرد، و نیز آنی که روش استخراج اشیاء داده ای را استفاده میکند که ما پیشنهاد میکنیم. ماندگاری آرشیوهای وب: زمان عامل خیلی تأثیرگذاری برای تقسیر محتوای خزش شده است. در حالی که ممکن است دادهها دست نخورده باقی بمانند روشی که ما آن را درک و ارائه میکنیم متفاوت است که دلیل عمده آن واقعیت خود ،زبان فرهنگ و وسایل فناوری تکامل بیان .است یکی از جدی ترین مشکلات بر شمرده شده در آرشیو وب زمانی است که فرمت دادههای خزش شده منسوخ و یا به طور کلی استفاده نشود راه حلهای ارائه شده توسط نویسندگان [9] [27] و [26] شبیه ساز نرم افزاری یا سخت افزاری ،انتقال محتوا یا شامل یک پروکسی است که قابلیتهای ترجمه فرمت را ترکیب خواهد

کرد و این کـار را به صورت پویا براساس درخواست کاربر و یا زمانی که نیاز تشـخیص داده شود انجام خواهـد داد. در حالی که این کارها در حـال تلاـش برای مبـارزه بـا تکامـل فنـاوری هسـتند این امکـان را به وجود می آورد که روش مخـالف را در بر بگیرد انطبـاق دادهها با فناوریهـای موجود به منظور انجام این کار میتوان محفظه سازی اطلاعات مرتبط صـفحه وب خزش شـده را و ذخیره سازی شـیء حاصـل مستقل از فرمت کدگذاری اصلی تصور کرد. در این صورت این واقعیت که فناوری تکامل می یابد دیگر چیزی برای مقاومت در مقابل آن نخواهد بود به جای آن امکان ارائه دادههای موجود با روشهای جدید را با تطبیق محتوای واقعی با پیش تنظیمات ارتقا خواهد داد استخراج اشیای داده ای که ما در این مقاله نشان میدهیم اولین گام در جهت ذخیره سازی اطلاعات آزاد از روش خاص کدگذاری است.

در حالی که میتوان اسـتدلال کرد که برای آرشـیو وب فرم اصلی صفحه وب اهمیت دارد این امر بیشتر با نیاز قرار دادن اطلاعات واقعی در مفهوم درست آن مرتبـط است توجه داشـته باشـید که هـدف مـا تغییر روشـهای موجود ذخیره سـازی محتوای آرشـیوها نیست بلکه برای مواردی که کاربرد دارد (داده های پویا یا فیدهای مرتبط به آن) میتوان برخی کاربردهای همپوشانی جالب را برای کسانی که از این مجموعه اسـتفاده میکنند یا برای خود آرشیویست ها ایجاد کرد استخراج اشیای داده ای در مفهوم آرشیو وب خیلی با تجزیه و تحلیل معنایی محتوا در زمـان و امکـان ارائه خـدمات ارزش افزوده مرتبط است. در واقع آنچه که میتوانـد بسـیار مفیـد باشـد این است که قادر به اجرای پرس وجوهای پیچیده در محتوا باشد و تعاملات و امکانات را به مصـرف کنندگان هدف اطلاعاتی که به عنوان آرشـیو وب ارائه می شوند اضافه نماید.

بازسازی صفحه های وب از شکل 2 بیاد میآید که صـفحه های وب مرتبط با فیدهای وب میتوانند بسـیار پویا باشند و در مواردی با فواصل دقیقه ای روزآمد میشوند در این شـرایط، تلاش برای جذب نسـخه های پی در پی صـفحه وب کانال متناظر غیر معقول به نظر میرسد با این حال از آنجا که فیـدهای وب مرجعی از تعـداد آیتمها را نگه میدارند و خوشـبختانه تعداد نسـبتاً زیادی برای چنین کانال پویایی)، هنوز این امکـان وجود دارد که در فیـد وب منظم خزش و آرشـیو کرد کـاربرد روش اسـتخراج دادههـا می توانـد بازسازی صـفحه وب در یک نقطه مفروض زمانی با استفاده از آیتمهای خزش شده فیدهای وب و ارجاع به مؤلفه های تمپلت ،باشد بنابراین، به نسخه ای از صفحه وب اشاره می کند که در واقع به روش کلاسیک خزش نشده است.

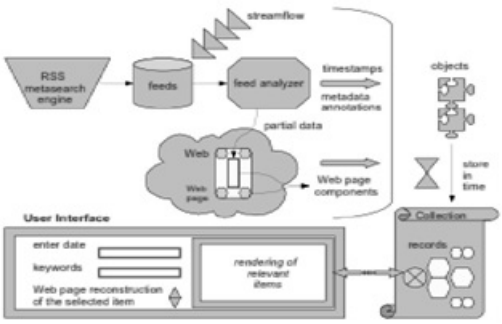
علاـوه بر این، بـا اسـتفاده از الگوریتم (احتمالاًـ برخی از فناوریهای هوشـمند)، میتوانیم نه تنها گره DOM را – که شامـل مقـاله است – شناسـایی کنیم بلکه سـایر گره هـای معنایی صـفحه را که شامل نظرها مقوله ها یا تگ هاست نیز شـناسایی کنیم این مناطق میتواننـد برای محتوای متنی و منابع به طور مسـتقل از تمپلت صـفحه وب (که شامل باقی مانده های پس از اسـتخراج است). اسـتخراج و ذخیره شوند. رابطه بین اجزای به دست آمـده را میتوان با اسـتفاده از معناها و تجزیه و تحلیل اسـتفاده شده برای این اجزا دوباره اسـتنباط و ایجاد کرد ما میتوانستیم روش اطلاعات ارائه شده را با ترکیبی از اشیای (در ترکیب) دوباره اختراع کنیم تا با تنظیمهای کاربر انطباق داده شود.

شکل .7 کاربردهایی برای بازسازی صفحه های وب از اشیای داده ای

شکل 7 جریان بازسازی صفحه های وب از اشیای داده ای را همانطور که در [17] توصیف شده است نشان میدهد.

خزش و بهره برداری معنـایی از آرشـیوها در نهـایت مـا میخواهیم به طور خلاصه دو برنامه کاربردی دیگر را یادآوری کنیم که قبلاً در این مقاله ذکر شـده بودنـد. در مرحله نخست، این امکان وجود دارد که از روش اسـتخراج شـیء داده ای برای تشـخیص تغییر اسـتفاده کرد با اسـتفاده از روش تجزیه و تحلیـل زمانی روی فیـدهای مشابه در همان موردی که در این مقاله ذکر شـد میتوانیم راهبرد انتشار کانالی را که قسمت پویای وبگاه را ارئه میکند تعیین کنیم با این ،آگاهی میتوان آهنگ خزش را با فرکانس تقریبی انطباق داد. علاوه بر این در مورد مقاله های خزش شده ،قبلی می توانیم تشـخیص بدهیم که نسـخه جدیدی جهت خزش ظاهر شده یا نه این کار را میتوان با استفاده از الگوریتم اسـتخراج اشـیای داده ای روی مقاله های خزش شده و نسخه جدید محتمل آن صفحه وب (فعلی انجام داد با مقایسه نتایج اشیای داده ای اشاره کننده به همان یو آر.ال، میتوانیم ببینیم که آیا در متن یا منابع در دوره مداخله زمانی تغییر رخ داده است.

عکس



شکل۷. کاربردهایی برای بازسازی صفحه‌های وب از اشیای داده‌ای

شکل ۷ جریان بازسازی صفحه‌های وب از اشیای داده‌ای را همانطورکه در [۱۷] توصیف شده است نشان می‌دهد.

خزش و بهره‌برداری معنایی از آرشیوها. در نهایت، ما می‌خواهیم به‌طور خلاصه دو برنامه کاربردی دیگر را یادآوری کنیم که قبلاً در این مقاله ذکر شده بودند. در مرحله نخست، این امکان وجود دارد که از روش استخراج شیء داده‌ای برای تشخیص تغییر استفاده کرد. با استفاده از روش تجزیه و تحلیل زمانی روی فیدهای مشابه، در همان موردی که در این مقاله ذکر شد، می‌توانیم راهبرد انتشار کانالی را که قسمت پویای وبگاه را ارئه می‌کند، تعیین کنیم. با این آگاهی، می‌توان آهنگ خزش را با فرکانس تقریبی انطباق داد. علاوه بر این، در مورد مقاله‌های خزش شده قبلی، می‌توانیم تشخیص بدهیم که نسخه جدیدی جهت خزش ظاهر شده یا نه. این کار را می‌توان با استفاده از الگوریتم استخراج اشیای داده‌ای روی مقاله‌های خزش شده و نسخه جدید محتمل آن (صفحه وب فعلی) انجام داد. با مقایسه نتایج اشیای داده‌ای اشاره‌کننده به همان یو.آر.ال، می‌توانیم ببینیم که آیا در متن یا منابع در دوره مداخله زمانی تغییر رخ داده است.

دوم، آرشیو وب حاوی اشیای داده‌ای (شاید علاوه بر صفحه‌های وب) را می‌توان توسط تحلیلگران به‌طور مؤثرتری نسبت به استفاده صرف آرشیو صفحه‌های وب استفاده کرد. به‌عنوان مثال، یک زبان‌شناس می‌تواند روی اصطلاحات جدیدی که در مقاله‌های روزنامه‌های ظاهر می‌شوند، بدون در نظر گرفتن اصطلاحاتی که در نظر ظاهر می‌شود، تمرکز کند. به‌طور کلی، هدف این است که معنای قابل بهره‌برداری و موقتی (در سطوح بهتر) به مجموعه‌ای از وب آرشیوهای رساتر و قابل انطباق با نیازهای کاربران اضافه شود.

تشکر و قدردانی. ما از وب آرشیو اروپا (مخصوصاً جولین ماسان[1]، گابریل واسل[2] و رادو پاپ[3]) به‌خاطر بحث و تبادل نظر در مورد عنوان این مقاله و کمک‌هایشان در به‌دست آوردن مجموعه داده‌های آزمایش‌هایمان سپاسگزاری می‌کنیم.

1. Julien Masanès
2. Gabriel Vasile
3. Radu Pop

دوم، آرشیو وب حاوی اشیای داده ای شاید علاوه بر صفحه های وب را می توان توسط تحلیلگران به طور مؤثرتری نسبت به استفاده صرف آرشیو صفحه های وب استفاده کرد به عنوان مثال یک زبان شناس میتواند روی اصطلاحات جدیدی که در مقاله های روزنامه‌های ظاهر میشوند بدون در نظر گرفتن اصطلاحاتی که در نظر ظاهر میشود تمرکز کند به طور کلی هدف این است که معنای قابل بهره برداری و موقتی در سطوح بهتر به مجموعه ای از وب آرشیوهای رساتر و قابل انطباق با نیازهای کاربران اضافه شود .

تشکر و قدردانی ما از وب آرشیو اروپا (مخصوصاً جولین ماسان(1) گابریل واسل(2) و رادو پاپ(3)) به خاطر بحث و تبادل نظر در مورد عنوان این مقاله و کمکهایشان در به دست آوردن مجموعه دادههای آزمایشهایمان سپاسگزاری میکنیم.

Julien Masanè s -1

Gabriel Vasile -2

Radu Pop -3

7. منابع

1. E. Adar, J. Teevan, S. T. Dumais, and J. L. Elsas. The web changes everything: Understanding the dynamicsof web content. In Proc. WSDM, Barcelona, Spain, Feb.2009.

2. A. Arasu and H. Garcia-Molina. Extracting structured data from Web pages. In Proc. SIGMOD, San Diego, USA, June 2003.

3. J. Cho and H. Garcia-Molina. The evolution of the Web and implications for an incremental crawler. In Proc. VLDB, Cairo, Egypt, Sept. 2000.

4. V. Crescenzi, G. Mecca, and P. Merialdo. Roadrunner: Towards automatic data extraction from large Websites. In Proc. VLDB, Roma, Italy, Sept. 2001.

5. H. V. de Sompel, M. L. Nelson, C. Lagoze, andS. Warner. Resource harvesting within the oai-pmh framework. In D-Lib Magazine, volume 10, number 12,Dec. 2004.

6. P. Dmitriev, C. Lagoze, and B. Suchkov. As we may perveive: Inferring logical documents from hypertext.In Proc. HT, Salzburg, Austria, Sept. 2005.

7. Eddie java feed parser. Website, 2010. http://www.davidpashley.com/projects/eddie.html.

8. D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In Proc. WWW, Budapest, Hungary, May 2003.

9. J. Hunter and S. Choudhury. Implementing preservation strategies for complex multimedia objects. In Proc. ECDL, Trondheim, Norway, Aug. 2003.

10. P. L. B. II, J. Johnson, U. P. Karadkar, R. Furuta, and F. Shipman. Application of kalman filters to identify unexpected change in blogs. In Proc. JCDL, Pittsburgh, USA, June

.2008

.Internet Archive. Website, 2010. http://web.archive.org/collections/ web. html .11

C. Kholschutter, P. Fankhauser, and W. Nejdi. Boilerplate detection using shallow text .12

.features. In Proc. WSDM, New York, USA, Feb. 2010

G. Knight and M. Pennock. Data without meaning: Establishing the significant properties .13

.of digital research. International Journal of Data Curation, 4(1), 2009

,B. Liu, R. Grossman, and Y. Zhai. Mining data records in Web pages. In Proc. KDD .14

.Washington, USA, Aug.2003

.J. Masanè s, editor. Web Archiving. Springer–Verlag, Heidelberg, Allemagne, 2006 .15

A. Ntoulas, J. Cho, and C. Olston. What's new on the web? the evolution of the web from .16

.a search engine perspective. In Proc. WWW, New York, USA, May. 2004

.M. Oita and P. Senellart. Archivage du contenu é phé mè re du Web a l'aide des flux Web .17

.In Proc. BDA,Toulouse, France, Oct. 2010. Conference without formal proceedings

.(Demonstration)

J. Pasternack and D. Roth. Extracting article text from the web with maximum .18

.subsequence segmentation. In Proc. WWW, Madrid, Spain, Apr. 2009

Z. Pehlivan, M. Ben Saad, and S. Gançarski. A novel Web archiving approach based on .19

.visual pages analysis. In Proc. IWAW, Corfu, Greece, Sept. 2009

M. Pennock and R. Davis. ArchivePress: A really simple solution to archiving blog .20

.content. In Proc. iPRES, San Francisco, USA, 2009

./Search4RSS feed search engine. Website, 2010. http://www. search4rss.com .21

K. C. Sia, J. Cho, and H.-K. Cho. Efficient monitoring algorithm for fast news alerts. In .22

IEEE Trans. On Knowl. and Data Eng., vol. 19 nb.7, Piscataway, NJ, USA, 2007. IEEE

.Educational Activities Department

.K. Sigursson. Incremental crawling with Heritrix. In Proc. IWAW, Vienna, Austria, Sept .23

.2005

./Blogger's choice awards. Website, 2010. http://bloggerschoiceawards.com .24

M. Spaniol, D. Denev, A. Mazeika, and G. Weikum. Catch me if you can. Temporal.25

.coherence of Web archives. In Proc. IWAW, Aarhus, Denmark, Sept. 2008

,S. Strodl, P. P. Beran, and A. Rauber. Migrating content in warc files. In Proc. IWAW .26

.Corfu, Greece, Sept. 2009

D. S. Swaney, F. McCown, and M. L. Nelson. Dynamic Web file format transformations .27

.with grace. In Proc. IWAW, Vienna, Austria, Sept. 2005

.H. van de Sompel, M. L. Nelson, R. Sanderson, L. L. Balakireva, S. Ainsworth, and H .28

,Shankar. Extracting structured data from Web pages. In Proc. SIGMOD to be modified

.San Diego, USA, June 2009

S. Yu, D. Cai, J.-R. Wen, and W.-Y. Ma. Improving pseudo-relevance feedback in web .29

,information retrieval using web page segmentation. In Proc. WWW.Budapest, Hungary

.May 2003

چکیده:

امروزه با افزایش اهمیت محتویات آرشیو وب یا حداقل بخشی از محتویات آن نگهداری منابع مهمی که برای مشورت و بررسی ضروری هستند به وظیفه ای حساس تبدیل شده است برای اطمینان از سازگاری آرشیو وب و نگهداری پیوسته ،آن خزشگرها به طور متناوب نسخههای جدید اسناد را از وب بازیابی میکنند. در عین حال خزش به صفحات وب با تغییرات کم اهمیت مانند آگهیها که دائماً صفحه را روزآمد میکنند مکرراً اتفاق میافتد. بنابراین سیستمهای آرشیو وب زمان و فضا را برای شاخص گذاری و ذخیره نسخههای این صفحات کم اهمیت وب به هدر می دهند برای حل این مشکل و اطمینان از آرشیو مؤثر صفحات وب روش جدیدی را معرفی میکنیم که تغییرات مهم بین نسخههای اسناد آرشیو شده را تشخیص میدهد روش ما مفهوم نمایش دیداری صفحات را با مفهوم موجودیت در تشخیص تغییرات بین نسخه ها ترکیب می.کند روش مورد نظر شامل آرشیو ساختار دیداری صفحه وب است که به صورت فرم معنایی بلاکها ارائه میشود در این ،مقاله الگوریتمی را برای تشخیص تغییرات ویژه در ساختار دیداری این صفحات ارائه میدهیم همچنین روشی را برای ارزیابی اهمیت تغییرات مشخص شده معرفی میکنیم تجربیات به دست آمده از نتایج اسناد وب نشان میدهند که روش مورد نظر امیدوار کننده است.

کلیدواژه ها آرشیو وبگاه تشخیص تغییرات تجزیه و تحلیل دیداری صفحات

آرشیو صفحات وب برمبنای تحلیل دیداری وDIFF

نوشته: میریام بن سعد(1) - استفان گانکار سکی(2) | زینب پهلوان(3)

ترجمه: مجیدرضا وحیدی(4)

1.انگیزه

با رشد سریع محتوای وب حفظ و نگهداری منابع پرکاربرد اطلاعات به وظیفه بسیار مهمی تبدیل شده است به همین دلیل انجام این وظیفه برای بسیاری از مؤسسه های ملی آرشیو در سراسر جهان اهمیت زیادی دارد. در عین حال، وب، بسیار پویا و در طول زمان در حال تکامل است (صفحات به طور دائم تغییر میکنند) در بیشتر موارد آرشیو کردن وب(5) به طور خودکار با استفاده از خزشگرهای وب اجرا می شود. خز شگرهای وب صفحات وب را که باید آرشیو شوند مشاهده و یک کپی نمونه(6) و / یا شاخص(7) از صفحات وب ایجاد میکنند برای بهروز نگهداری ،آرشیو خزشگر باید به طور متناوب صفحات را بازبینی کند و آرشیو را با کپی جدید به روز رسانی کند به دلیل این که معمولاً خزشگر، منابع محدودی دارد پهنای باند، فضای ذخیره سازی و مانند آن و با توجه به حجم عظیم صفحاتی که باید آرشیو شوند،

1- Myriam Ben Saad

2- Stephane Gancarski

3- Zeyneb Pahlivan

4- دانشجوی کارشناسی ارشد مهندسی نرم افزار از سازمان اسناد و کتابخانه ملی ج.ا.ا.

5- Web crawlers

6- Snapshot

7- Index

ممکن است خزشگر نتواند در همه زمانها یک سایت را بازبینی کند و نسخه(1) جدید صفحه را بارگذاری کند. در حقیقت نگهداری آرشیو کل وب یا حتی قسمتی از آن که شامل همه نسخه های همه صفحه ها باشد امکان پذیر نیست بنابراین مسئله این طور بیان میشود که خزش برای بارگذاری مهمترین نسخه ها را چگونه باید بهینه کنیم تا ریزش اطلاعات مفید به حداقل برسد. البته این کار باید بدون دخالت مدیران وبگاهها انجام شود از این رو سیستم آرشیو باید رفتار سایت را تخمین بزند تا زمان و میزان تناوب(2) بازبینی صفحه را مشخص کند کارهای متعدد [3,4] بر روی تناوب تغییر متمرکز شده اند تا با کمک آنها بتوان خزشگرهای وب را بهبود بخشید. در عین حال ممکن است خزشگر با ذخیره کردن یک نسخه جدید صفحه با تغییرات کم اهمیت زمان و فضا را به هدر دهد مثالی از این مورد آگهی ها هستند که به طور دائم تغییر میکنند .بنابراین روش مؤثری مورد نیاز است که مشخص کند تغییرات بین نسخه ها دقیقاً چه موقع و چند وقت یکبار صورت می پذیرد. روشهایی که تا به حال مطرح شده اند فقط تناوب تغییرات را تخمین میزنند ولی اهمیت تغییرات را در نظر نمی گیرند اگر بتوانیم تناوب تغییرات مهم را با درستی بیشتری پیش بینی کنیم اثر بخشی سیستم آرشیو وب بهبود می یابد.

برای تخمین تناوب به روزرسانیها باید تغییرات بین نسخههای بازیابی شده اسناد مشخص شود. بسیاری از الگوریتمهای موجود [5, 6] به طور ویژه برای مشخص کردن تغییرات بین اسناد نیمه ساخت یافته(3) (xml و html)طراحی شده اند در عین حال روشی وجود ندارد که تغییرات مرتبط / نامرتبط را از اطلاعات پراستفاده بدون استفاده تمیز دهد کارهای قبلی [2] نشان میدهد که میتوان صفحه را به بخشهای(4) متعدد یا بلاک ها(5) تقسیم(6) کرد. معمولاً بلاکهای موجود در یک صفحه اهمیت متفاوتی دارند در حقیقت در صفحات ،وب نواحی مختلف بر حسب موقعیت ،اندازه و محتوا دارای وزن اهمیت متفاوتی هستند. معمولاً اطلاعات مهمتر در مرکز صفحه قرار دارد. آگهی در بالای صفحه یا سمت چپ و حق نشر(7) در قسمت پانویس قرار دارد با تقسیم بندی صفحه باید به هر بلاک ، یک میزان اهمیت نسبی داده شود. این کار برای نمونه با استفاده از الگوریتم [7] یا در کل با روش یادگیری ماشین به طور خود کار انجام میشود سپس میتوانیم اهمیت تغییرات را بین دو نسخه صفحه محاسبه کنیم این محاسبه برمبنای دو مشخصه انجام میشود:1) اهمیت نسبی بلاکها و 2) اهمیت نسبی عملیات (درج ، حذف ، به روزرسانی) که در این بلاکها انجام شده است و با مقایسه دو نسخه مشخص می شود.

در این تحقیق روشی را برای مشخص کردن تغییرات مهم بین نسخه ها برای آرشیو مؤثر وب پیشنهاد میکنیم این ،روش روی یک انباره(8) برای مؤسسه ملی سمعی و بصری فرانسه (INA) به کار رفته است.

1- Version

2- Frequency

3- Semi-structured documents

4- Segments

5- Block

6- Partition

7- Copyright

8- Repository

یکی از وظایف INA ایجاد ذخایر قانونی(1) است که صفحات وب رادیو و تلویزیون فرانسه و صفحات مرتبط آن را نگهداری می.کند یکی از نیازهای این پروژه نگهداری جنبه دیداری صفحات است. بنابراین ایده ما به کارگیری تحلیل دیداری صفحه برای نسبت دادن اهمیت به بخشهای مختلف صفحه وب بر مبنای موقعیت نسبی آنهاست مفاهیم تحلیل دیداری صفحه و اهمیت بخشهای صفحات وب جدید نیستند ولی تا آنجا که میدانیم برای آرشیو وب به صورت ترکیبی به کار نرفته اند.

ادامه مطالب این مقاله به این صورت است: بخش 2 معماری سیستم آرشیو وب را ارائه میدهد. بخش 3 مدل توسعه یافته تقسیم بندی دیداری صفحه برای صفحات وب HTML را توضیح میدهد. در بخش 4 الگوریتم تشخیص تغییر کافی برای محاسبه اختلاف بین دو نسخه صفحه بازسازی شده دیداری را ارائه میدهیم در بخش 5 روش ارزیابی اهمیت بلاکها تغییرات را ارائه میدهیم. در بخش 6 راهبرد به کار رفته برای زمانبندی مؤثر خزشگرهای وب و در بخش 7 نگاهی به همه مراحل روش مورد نظر داریم. در بخش 8 نیز کارهای آینده را مطرح می کنیم.

عکس

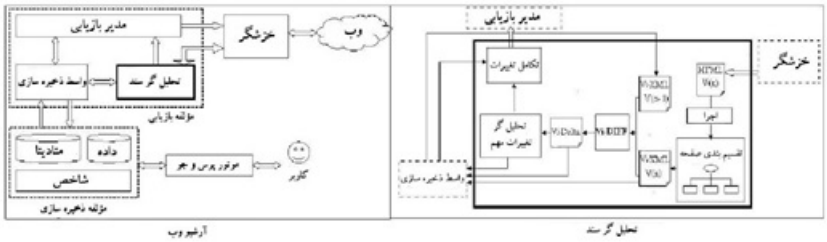
یکی از وظایف INA ایجاد ذخایر قانونی[1] است که صفحات وب رادیو وتلویزیون فرانسه و صفحات مرتبط آن را نگهداری می‌کند. یکی از نیازهای این پروژه، نگهداری جنبه دیداری صفحات است. بنابراین ایده ما به‌کارگیری تحلیل دیداری صفحه برای نسبت دادن اهمیت به بخش‌های مختلف صفحه وب بر مبنای موقعیت نسبی آنهاست. مفاهیم تحلیل دیداری صفحه و اهمیت بخش‌های صفحات وب جدید نیستند، ولی تا آنجا که می‌دانیم برای آرشیو وب به‌صورت ترکیبی به‌کار نرفته‌اند.

ادامه مطالب این مقاله به این صورت است : بخش ۲ معماری سیستم آرشیو وب را ارائه می‌دهد. بخش ۳ مدل توسعه یافته تقسیم‌بندی دیداری صفحه برای صفحات وب HTML را توضیح می‌دهد. در بخش ٤ الگوریتم تشخیص تغییر کافی برای محاسبه اختلاف بین دو نسخه صفحه بازسازی شده دیداری را ارائه می‌دهیم. در بخش ٥ روش ارزیابی اهمیت بلاک‌ها/ تغییرات را ارائه می‌دهیم. در بخش ٦ راهبرد به‌کار رفته برای زمان‌بندی مؤثر خزشگرهای وب و در بخش ۷ نگاهی به همه مراحل روش مورد نظر داریم. در بخش ۸ نیز کارهای آینده را مطرح می‌کنیم.

۲. معماری آرشیو وب

آرشیو وب شامل ٤ مؤلفه اصلی است : خزشگرهای وب، مؤلفه بازیابی[2]، مؤلفه ذخیره سازی[3]، و موتور پرس‌وجو[4]. شکل ۱ شمای کلی سیستم را ارائه می‌دهد.

خزشگر وب: خزشگرهای وب ، با بارگذاری مکرر نسخه‌های جدید صفحات ، مرور و بررسی وب را پوشش می‌دهند.



تصویر ۱. دید کلی از آرشیو وب

مؤلفه بازیابی: امکان نگهداری آرشیو را به‌صورت به‌روز فراهم می کند. که شامل ۳ ماژول است: ۱) مدیر بازیابی[5]، امکان بهینه‌سازی منابع تخصیص یافته را فراهم می‌کند تا اطلاعات کمتری از دست رود. همچنین، اسنادی را که باید به سرعت به روز رسانی شوند انتخاب می کند تا آرشیو در حد ممکن

1 - Legal deposit
2 - Freshness component
3 - Storage component
4 - Query engine
5 - Freshness Manager

2.معماری آرشیو وب

آرشیو وب شامل 4 مؤلفه اصلی است : خزشگرهای وب مؤلفه بازیابی(2) ، مؤلفه ذخیره سازی(3)، و موتور پرس وجو(4). شکل 1 شمای کلی سیستم را ارائه می دهد.

خزشگر :وب خزشگرهای وب با بارگذاری مکرر نسخه های جدید صفحات ، مرور و بررسی وب

را پوشش میدهند .

تصویر 1 دید کلی از آرشیو وب

مؤلفه بازیابی امکان نگهـداری آرشـیو را به صورت به روز فراهم می.کند که شامل 3 ماژول است 1)مدیر بازیابی(5)، امکان بهینه سازی منابع تخصـیص یافته را فراهم میکند تا اطلاعات کمتری از دست رود. همچنین اسـنادی را که باید به سـرعت به روز رسانی شوند انتخاب میکند تا آرشیو در حد ممکن

1- Legal deposit

2- Freshness component

3- Storage component

4- Query engine

5- Freshness Manager

به روز نگهداری شود .2) تحلیلگر سند(1) امکان تشخیص و تحلیل نسخههای صفحه وب بازیابی شده را فراهم می.کند با توجه به اینکه تحلیلگر سند هسته اصلی روش ما را تشکیل میدهد، در اینجا جزئیات بیشتری در مورد آن ارائه میدهیم تحلیلگر سند شامل زیر ماژولهای متعددی مطابق با مراحل مختلف تحلیل صفحه است که در شکل 1 نشان داده شده است. تحلیلگر سند برای به دست آوردن نسخه صفحه HTML خاصی که باید آرشیو شود با خزشگر در تعامل است. سپس برای بازیابی اطلاعات دیداری صفحه توسط موتور تفسیر(2) پردازش می شود. منفعت اصلی تفسیر فراهم کردن یک توضیح دیداری کامل و حقیقی از سند حتی با وجود اسکریپتهای صفحه مانند جاوا اسکریپت است پس از آن، صفحه تفسیر شده تقسیم بندی میشود و ساختار طرح بندی دیداری(3) صفحه ساخته میشود الگوریتم VIPS [2] برای تقسیم بندی صفحه وب به بلاکهای سلسله مراتبی معنایی(4) به کار میرود و قسمتهای کافی برای صفحه وب در نظر می گیرد. ما این الگوریتم را برای استخراج ،پیوندها تصاویر و متنهای هر بلاک توسعه داده ایم الگوریتم توسعه یافته VIPS، یک سند Vi-XML را به عنوان خروجی تولید می کند که ساختار محتوای سلسله مراتبی صفحه را توضیح میدهد در پایان فرآیند تقسیم بندی الگوریتم تشخیص تغییر Vi-DIFF، از تغییرات ایجاد شده بین نسخه جدید Vi-XML تولید شده (Vn) و آخرین نسخه آرشیو شده V(n-1) توضیحی فراهم میکند تغییرات در یک فایل delta XML که Vi-Delta نام دارد ذخیره می شوند. این فایل اعمال اتفاق افتاده بین دو سند (درج ، حذف و مانند آن) را توضیح میدهد. پس از آن، فایل Vi-Delta توسط زیر ماژول تحلیلگر تغییرات مهم(5) تحلیل میشود تا اهمیت تغییرات مشخص شده ارزیابی شود ماژول 3)تکامل تغییرات(6) ، نتیجه این ارزیابی تغییر را برای بهینه سازی بیشتر منابع مدیریت شده توسط مؤلفه ،بازیابی مورد استفاده قرار میدهد در پایان Vi-Delta و نسخه فعلی Vi-XML از طریق واسط ذخیره سازی(7) با اطلاعات اضافی مانند URL و تاریخ زمان خزش در پایگاه داده ذخیره می شوند واسط ذخیره سازی برای ذخیره کردن / شاخص کردن(8) نسخه های صفحه و فراداده آنها، که در طول تحلیل به دست آمده ،است با مؤلفه ذخیره در تعامل است.

مؤلفه ذخیره سازی مؤلفه ذخیره شامل واحدهای ذخیره داده و متادیتاست. همچنین شامل شاخصی است که پرس و جو از آرشیو را تسهیل می کند.

موتور پرس وجو کاربران میتوانند از طریق موتور پرس و جو بین نسخه ها و نسخه های صفحه

آرشیو شده درخواستی حرکت(9) کنند.

1- Document Analyzer

2- rendering engine

3- visual page layout structure

4- semantic hierarchical blocks

5- Importance Changes Analyzer

6- Changes Evolution

7- Storage Interface

8- Index

9- Navigate

3.تقسیم بندی دیداری صفحه

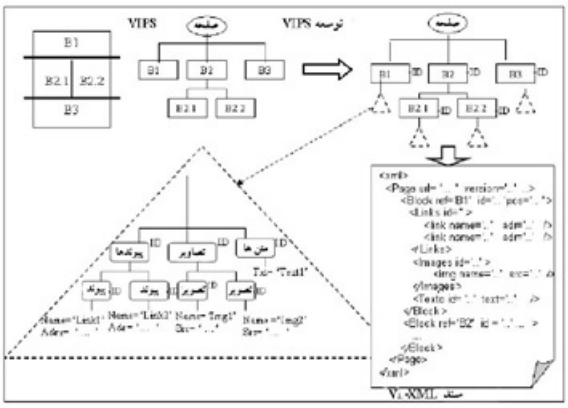
همانطور که قبلا ذکر شـد ، [2] VIPS برای تقسـیم بنـدی صـفحه وب به بلاکهای معنایی تو در تو بر مبنای گره های مناسب در درخت HTML DOM در صـفحه به کار میرود و جداکننـده های(1) عمودی و افقی در صـفحه را مشـخص میکنـد بر مبنای این جداکننـده ها درخت معنایی صـفحه و بی را که به بلاکهای متعدد تقسـیم شده است تولید می.کند. ریشه درخت کل صـفحه است. هر بلاک به عنوان گرهی از درخت در نظر گرفته میشود که در شکل 2 نشان داده شده است.

شکل .2 الگوریتم VIPS توسعه یافته

عکس

۳. تقسیم‌بندی دیداری صفحه

همانطور که قبلا ذکر شد ، [۲] VIPS برای تقسیم‌بندی صفحه وب به بلاک‌های معنایی تو در تو بر مبنای گره‌های مناسب در درخت HTML DOM در صفحه به‌کار می‌رود و جداکننده‌های[1] عمودی و افقی در صفحه را مشخص می‌کند. بر مبنای این جداکننده‌ها، درخت معنایی صفحه وبی را که به بلاک‌های متعدد تقسیم شده است، تولید می‌کند. ریشه درخت، کل صفحه است. هر بلاک، به عنوان گرهی از درخت در نظر گرفته می‌شود که در شکل ۲ نشان داده شده است.



شکل ۲. الگوریتم VIPS توسعه یافته

برای تکمیل درخت معنایی کل صفحه، الگوریتم VIPS را با استخراج پیوندها، تصاویر، و متن برای هر بلاک توسعه دادیم. همانطور که در شکل ۳ نشان داده شده است، هر گرهِ بلاک، دارای گره‌های فرزند اضافه‌ای است. گره‌های پیوندها، تصاویر، و متون هر کدام به‌ترتیب پیوندها، تصاویر، و متون هر بلاک را جمع‌آوری می‌کند. همه گره‌های صفحه، به‌طور منحصر به‌فرد، با یک خصوصیت ID شناسایی می‌شوند. این ID یک مقدار درهم‌سازی شده[2] است که با استفاده از محتوای گره و محتوای گره‌های فرزند آن محاسبه می‌شود. اگر گره‌های همتا[3] (گره‌های در موقعیت یکسان در دو نسخه متوالی) چند ID متفاوت داشته باشند، لزوماً محتویات آنها به‌روزرسانی شده است. گره‌های برگ، خصوصیات دیگری مانند نام و نشانی برای ابر پیوند دارند. الگوریتم VIPS توسعه یافته ما، یک سند Vi-XML را به‌عنوان خروجی تولید می‌کند. این سند، ساختار سلسله مراتبی صفحه وب را توضیح می‌دهد. که در شکل ۲ نشان داده شده است.

1 - Separators
2 - Hash value
3 - Matched nodes

برای تکمیل درخت معنایی کل صفحه الگوریتم VIPS را با استخراج پیوندها، تصاویر، و متن برای هر بلاک توسعه دادیم همانطور که در شکل 3 نشان داده شده است هر گره ،بلاک دارای گره های فرزند اضافه ای است گره های ،پیوندها ،تصاویر و متون هر کدام به ترتیب ،پیوندها تصاویر و متون هر بلاک را جمع آوری می کند. همه گره های صفحه، به طور منحصر به فرد، با یک خصوصیت ID شناسایی می شوند. این ID یک مقدار در هم سازی شده(2) است که با استفاده از محتوای گره و محتوای گره های فرزند آن محاسبه می شود. اگر گره های همتا(3)(گره های در موقعیت یکسان در دو نسخه متوالی) چند ID متفاوت داشته باشند، لزوماً محتویات آنها به روزرسانی شده است. گره های برگ، خصوصیات دیگری مانند نام و نشانی برای ابر پیوند دارند. الگوریتم VIPS توسعه یافته ما یک سند Vi-XML را به

عنوان خروجی تولید میکند این سند ساختار سلسله مراتبی صفحه وب را توضیح میدهد که در شکل 2 نشان داده شده است.

ص: 33

1- Separators

2- Hash value

3- Matched nodes

4.تشخیص تغییر

الگوریتمهای تشخیص تغییر متفاوت [5 ,6] برای اسناد xml منظور شـده اند. با توجه به اینکه آنها الگوریتمهای عمومی هسـتند، به طور کامل نیازمندیهای ما را برآورده نمی کنند. ما میخواهیم بعضـی ضوابط ویژه را برای مقایسه گرههای خصوصـیت مانند تشخیص پیوندهای به روز رسانی شده در صورت تغییر نام یا نشانی یک پیوند که خصوصـیات پیوند (هستند اضافه کنیم. همچنین می خواهیم یک متن به روز رسانی شده در دو بلاک همتا برمبنای امتیاز شباهت فاصله متنی تعداد کلمات مشترک را مشخص کنیم با الگوریتمهای عمومی این گره ها به صورت حذف شده از نسـخه قدیمی و افزوده شده به نسـخه جدید در نظر گرفته می شوند. ویژگی دیگر روش ما این است که لازم اسـت عناصر تغییر یافته درون یک بلاک و عناصر تغییر موقعیت داده از یک بلاک به بلاک دیگر را تشخیص دهیم. ولی عنصر تغییر موقعیت داده درون یک بلاک بـدون اهمیت است؛ زیرا در این صورت نه اطلاعاتی به بلاک اضافه شده و نه اطلاعاتی از آن حذف شده است. همچنین میخواهیم تغییر ساختار صـفحه در سـطح بلاکها از نسخه ای به نسخه دیگر تشخیص داده شود بلاک حذف شده درج شده بنابراین دلایل الگوریتم تشـخیص تغییرات به نام Vi-DIFF [1] را ارائه میـدهیم این ،الگوریتم اختلافات بین دو نسـخه سـند Vi XML- را محاسبه میکند و یک سند Vi-Delta تولید میکند که اختلاف (دلتای) بین دو نسخه را ارائه میدهد Vi-DIFF مورد نظر امکان مشخص کردن دو نوع تغییر را فراهم میکند : تغییرات ساختاری و تغییرات محتوایی.

تغییرات ساختاری معمولاً ساختار سـند XML را اصـلاح میکننـد در سـطح بلاکها)؛ در صورتی که تغییرات محتوایی، محتوای متنی را اصـلاح میکننـد در سـطح پیوندها ، تصاویر، و مانند آن. اگر فرض کنیم که تغییری در ساختار وجود ندارد پیچیدگی الگوریتم Vi-DIFF لگـاریتمی - خطی ((nlog(n) .است در اینجـا n تعـداد کـل گرهها است. اگر تغییرات ساختاری وجود داشـته باشـد در بـدترین حالت (حالتی که همه ساختار تغییر یابد) پیچیدگی الگوریتم از درجه دوم O(n) است. لازم به ذکر است که n همیشه کوچک باقی میماند.

5.اهمیت تغییرات

بر مبنای Vi-Delta تولید شده توسط الگوریتم Vi-DIFF تابعی را در نظر میگیریم که اهمیت تغییرات مشخص شده را ارزیابی می.کند این ،کار وظیفه زیر ماژول تحلیلگر تغییرات مهم (شکل 1)است این تابع، به به 3 پارامتر اصلی وابسته است:

اهمیت بلاـک به روز رسانی شـده معمولاً مهمترین اطلاعات در مرکز، و آگهی ها در بالای صـفحه قرار دارنـد. سانگ و همکارانش ،[7] نسـبت دادن مقـادیر اهمیت برای بلاکهای مختلف در صـفحه وب به طور خودکار را پیشـنهاد میدهنـد همچنین میتوانیم سایر پارامترها را برای ارزیـابی اهمیت یک بلاک، با توجه به تاریخچه تغییرات این بلاک در نظر بگیریم. برای نمونه، بلاکی که با تناوب بیشتری تغییر میکند اهمیت کمتری دارد مطالعه بیشتر برای یافتن بهترین روش به منظور تخمین

اهمیت بلاکها ضروری است

اهمیت عملیات اهمیت عملیات به نوع عمل جابه جایی درج و از این قبیل) و عنصر تغییر یافته پیوند، تصویر، و مانند آن بستگی دارد برای نمونه ممکن است عملیات درج و حذف اهمیت بیشتری نسبت به جابه جایی داشته باشد ،همچنین ممکن است درج تصویر از درج یک پیوند یا متن مهم تر باشد بنابراین برای انتخاب بهترین مقادیر پارامترها برای هر نوع عملیات، مطالعه روشهای یادگیری ماشین در برنامه ما قرار دارد.

میزان تغییرات در بلاک میزان عملیات تغییر درج و حذف و از این قبیل) روی داده در یک بلاک برای هر عنصر (پیوند) تصویر (و متن از Vi–Delta تولید شده استنباط میشود این میزان درصد عملیات تغییر مشخص شده برای هر بلاک تقسیم بر تعداد کل عناصر بلاک را مشخص می کند.

بر مبنای این پارامترها ، تابع E(V،V) را ارائه می دهیم که اهمیت تغییرات بین نسخه هایV و Vرا)که هر کدام از آنها از بلاکهای BK تشکیل شده است) تخمین میزند.

در این فرمول :

{درج، حذف، به روزرسانی، جابه جایی } – OP –

}پیوند ،تصویر متن { – El –

به ترتیب تعداد بلاکها تعداد نوع عملیات و تعداد نوع عناصر هستند = Na, NO, NB –

مقدار اهمیت X که میتواند یک بلاک یا یک عملیات تغییر را مشخص میکند = (x)

El

عکس

اهمیت بلاک‌ها ضروری است.

- اهمیت عملیات. اهمیت عملیات به نوع عمل(جابه‌جایی، درج، و از این قبیل) و عنصر تغییر یافته (پیوند، تصویر، و مانند آن) بستگی دارد. برای نمونه، ممکن است عملیات درج و حذف اهمیت بیشتری نسبت به جابه‌جایی داشته باشد. همچنین، ممکن است درج تصویر از درج یک پیوند یا متن مهم‌تر باشد. بنابراین، برای انتخاب بهترین مقادیر پارامترها برای هرنوع عملیات، مطالعه روش‌های یادگیری ماشین در برنامه ما قرار دارد.
- میزان تغییرات در بلاک. میزان عملیات تغییر (درج و حذف، و از این قبیل) روی داده در یک بلاک برای هر عنصر (پیوند ، تصویر، ومتن) از Vi-Delta تولید شده استنباط می‌شود. این میزان، درصد عملیات تغییر مشخص شده برای هر بلاک تقسیم بر تعداد کل عناصر بلاک را مشخص می‌کند.

بر مبنای این پارامترها ، تابع E(v۱,v۲) را ارائه می دهیم که اهمیت تغییرات بین نسخه‌های v۱ و v۲ را (که هر کدام از آنها از بلاک های Bk_i تشکیل شده است) تخمین می‌زند.

در این فرمول :

$$E = \sum_{i=1}^{N_{Bk}} I(Bk_i) * [\frac{1}{N_{Op}} \sum_{j=1}^{N_{Op}} I(Op_j) * \frac{1}{N_{El}} \sum_{k=1}^{N_{El}} \frac{N(Op_j, El_k)}{N(El_k, bk_i)}]$$

- Op_j = {درج، حذف، به‌روزرسانی، جابه‌جایی}
- El_k = {پیوند، تصویر، متن}
- N_{El} , NO_p, NB_k = به‌ترتیب، تعداد بلاک‌ها، تعداد نوع عملیات، و تعداد نوع عناصر هستند
- I(x) = مقدار اهمیت X که می‌تواند یک بلاک یا یک عملیات تغییر را مشخص می‌کند
- $N(Op_j, El_k)$ = تعداد عملیات تغییر j را که روی عنصر k روی میدهد، مشخص می‌کند
- $N(El_k, Bk_i)$ = تعداد کل عناصر k در بلاک i را مشخص می‌کند

برای نرمال سازی نتیجه تابع()E، محدودیت زیر را روی اهمیت بلاک‌ها اضافه می‌کنیم:

$$\sum_{i=1}^{N_{Bk}} I(Bk_i) = \quad 1; 0 \leq I(Op) \leq 1$$

تابع()E با ضرب درصد تغییرات، برای هر عمل (Op_j) و بلاک Bk_i، در اهمیت عملیات($I(Op_j$) و بلاک‌ها ($I(Bk_i$) محاسبه می‌شود و مقدار نرمالی بین ۰ و ۱ بر می‌گرداند.

٦. زمان‌بندی خزش وب

یکی از اهداف روش ما، بارگذاری مهم‌ترین نسخه‌ها بر مبنای زمان‌بندی خزش است. مهم‌ترین وظیفه زمان‌بند عبارت است از خزش به مهم‌ترین نسخه اسناد برمبنای تاریخچه تغییرات. زمان‌بند، فهرستی از اسناد مرتب شده توسط یک تابع ضرورت بازیابی[1] را مدیریت می‌کند. این تابع، برای هر صفحه

1 - Freshness urgency function

تعداد عملیات تغییر ز را که روی عنصر روی میدهد مشخص میکند = NOP),El –

تعداد کل عناصر در بلاک i را مشخص می کند = (NEBK –

برای نرمال سازی نتیجه تابع ()E، محدودیت زیر را روی اهمیت بلاکها اضافه میکنیم

تابع ()E با ضرب درصد تغییرات برای هر عمل (Op) و بلاک BK در اهمیت عملیات(OP) و

بلاکها (BK) محاسبه میشود و مقدار نرمالی بین 0 و 1 بر می گرداند.

6.زمان بندی خزش وب

یکی از اهداف روش ما بارگذاری مهمترین نسخه ها بر مبنای زمان بندی خزش است. مهمترین وظیفه زمان بند عبارت است از خزش به مهمترین نسخه اسناد برمبنای تاریخچه تغییرات زمان بند، فهرستی از اسناد مرتب شده توسط یک تابع ضرورت بازیابی(1) را مدیریت می.کند این تابع برای هر صفحه

1- Freshness urgency function

میزان ضـروری بودن بازیابی آن در یک زمان داده شـده را مشـخص میکند این تابع اهمیت تغییرات را به حساب می آورد تخمین زده شده توسط تابع (E() که بین نسـخه اصـلی و آخرین نسـخه آرشـیو شـده روی داده است. تابع ضـرورت بازیابی برای زمانبنـد را به صورت زیر تعریف میکنیم:

که در آن :

اولویت صفحه =p

متوسط اهمیت تغییرات تخمین زده شده بین دو نسخه اسناد = AvgE

زمان آخرین نسخه بازیابی شده -DATE

زمان اولین نسخه اسناد-DATE

این تابع به موارد زیر بستگی دارد :

1 - اولویت صفحه،

2-اهمیت تغییرات نسخههای آرشیو شده قبلی و

3- زمان آخرین بازیابی صفحه

برای تولیـد به روزرسانیهـای ایجاد شـده با اهمیت تغییرات متفاوت روی اسـناد شبیه سازی شـده، نوعی شبیه ساز تولیـد کرده ایم. سـپس راهبرد زمان بندی با استفاده از تابع ضرورت توسعه داده میشود و با دو سیاست خزش موجود مقایسه می شود.

([Round Robin t, Cho [4)

نتایج اولیه امیدوار کننده هسـتند. در واقع راهبرد به کارگیری تابع ضـرورت نسبت به سایر سیاستهای موجود، بازیابی نسخه های مهم تر را امکان پذیر می کند.

عکس

میزان ضروری بودن بازیابی آن در یک زمان داده شده را مشخص می‌کند. این تابع، اهمیت تغییرات را به حساب می‌آورد (تخمین زده شده توسط تابع E()) که بین نسخه اصلی و آخرین نسخه آرشیو شده روی داده است. تابع ضرورت بازیابی برای زمان‌بند را به صورت زیر تعریف می‌کنیم:

$$U(doc, date) = p * \frac{Avg_E}{date_{lastUpd} - date_{OrigDoc}} * (date - date_{lastUpd})$$

که در آن :

p= اولویت صفحه

AvgE = متوسط اهمیت تغییرات تخمین زده شده بین دو نسخه اسناد

$date_{lastUpd}$= زمان آخرین نسخه بازیابی شده

$date_{OrigDoc}$= زمان اولین نسخه اسناد

این تابع به موارد زیر بستگی دارد :

۱- اولویت صفحه،

۲- اهمیت تغییرات نسخه های آرشیو شده قبلی، و

۳- زمان آخرین بازیابی صفحه.

برای تولید به‌روزرسانی‌های ایجاد شده (با اهمیت تغییرات متفاوت) روی اسناد شبیه‌سازی شده، نوعی شبیه‌ساز تولید کرده‌ایم. سپس راهبرد زمان‌بندی با استفاده از تابع ضرورت توسعه داده می‌شود و با دو سیاست خزش موجود مقایسه می‌شود.

(Round Robin †, Cho [٤] ‡)

نتایج اولیه، امیدوار کننده هستند. در واقع، راهبرد به‌کارگیری تابع ضرورت نسبت به سایر سیاست‌های موجود، بازیابی نسخه‌های مهم‌تررا امکان‌پذیر می‌کند.

۷ . جمع بندی

در این قسمت، همه مراحل روش را بیان می‌کنیم: تقسیم‌بندی دیداری تشخیص تغییرات و ارزیابی اهمیت تغییرات . این موارد بر روی صفحات مختلف وب HTML رادیو و تلویزیون اعمال می شود.

- تقسیم‌بندی دیداری

ابتدا نشان می‌دهیم که صفحات وب، چگونه به بلاک‌های معنایی دیداری تقسیم‌بندی می‌شوند و چگونه پیوندها، تصاویر، و متن‌ها برای هر بلاک استخراج می‌شوند. یک سند Vi-XML، به‌عنوان خروجی تولید می‌شود که ساختار دیداری سلسله مراتبی صفحه را نشان می‌دهد. نتایج تجربیِ (از لحاظ زمان اجرا و اندازه خروجی) تقسیم‌بندی دیداری ارائه می‌شوند.

- تشخیص تغییرات

در این مرحله، نشان می‌دهیم که الگوریتم Vi-DIFF، تغییرات بین نسخه‌های مختلف اسناد را

7. جمع بندی

در این قسمت همه مراحل روش را بیان میکنیم تقسیم بندی دیداری تشخیص تغییرات و ارزیابی اهمیت تغییرات این موارد بر روی صفحات مختلف وب HTML رادیو و تلویزیون اعمال میشود

- تقسیم بندی دیداری

ابتـدا نشان میدهیم که صـفحات ،وب چگونه به بلاکهای معنایی دیداری تقسـیم بندی می شوند و چگونه پیوندها ،تصاویر و متنها برای هر بلاک استخراج میشوند یک سند ViXML، به عنوان خروجی تولید میشود که ساختار دیداری سلسله مراتبی صفحه را نشان میدهد. نتایج تجربی (از لحاظ زمان اجرا و اندازه خروجی) تقسیم بندی دیداری ارائه میشوند.

- تشخیص تغییرات

در این مرحله نشان میدهیم که الگوریتم Vi-DIFF تغییرات بین نسخه های مختلف اسناد را

مشـخص می.کند آزمایشـها روی نسخه های گوناگون اسناد Vi-XML بـا تغییرات متفـاوتی اعمال می شونـد:تغییرات محتوا و تغییرات ساختاری.

Vi-Delta خروجی عملیات تغییر روی داده بین دو نسـخه از اسـناد را توضـیح می دهد که با موقعیت تغییرات در سند اصلی HTML به صورت شهودی در آمدهاند پس از آن کارآیی Vi-DIFF از لحاظ زمان اجرا ارائه میشود.

اهمیت تغییرات

برمبنای Vi-Delta تولید شده، اهمیت تغییرات مشخص شده را با استفاده از تابع E() را که، در بخش ه توضیح داده شد، ارزیابی میکنیم از طریق شبیه سازی نشان میـدهیم که چگونه این تابع را میتوان برای زمان بندی خزشـگرها به کار برد. همانطور که در بخش 6 مطرح شد زمان بند با اسـتفاده از تابع ضـرورت، ضـروری ترین سـندی که باید بازیابی شود را انتخاب می.کند. مطالعه مقایسـهای راهبردهای زمان بندی شبیه سازی شده ارائه میشود.

8. کارهای آینده

کارهـای آینـده ، مرتبـط به مـدیر بازیابی و تخمین اهمیت است. ما در حال حاضـر بهترین روش یادگیری ماشـین را برای به دست آوردن خودکار اهمیت نسبی بلاکها و اهمیت عملیات تغییر جست وجو می کنیم. کار ،دیگر مشـخص کردن ،انتقال جداشدن و پیوستن بلاکها به عنوان تغییرات ساختاری است. همچنین قصد داریم روشـمان را به منظور آشکارسازی و تحلیل تغییرات بین دو نسخه از سایت به جای دو صفحه گسترش دهیم.

منابع

[1] M. Ben Saad, S. Gançarski, and Z. Pehlivan. A Novel Web Archiving Approach based on Visual Pages Analysis.

In IWAW '09: 9th International Web Archiving Workshop, Corfu, Greece, 2009.

[2] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. VIPS: a Vision-based Page Segmentation Algorithm, Technical report,

Microsoft Research, 2003.

[3] J. Cho and H. Garcia-Molina. The Evolution of the Web and Implications for an Incremental Crawler. In VLDB

Proceedings of the 26th International Conference on Very Large Data Bases, 2000 :00".

[4] J. Cho and H. Garcia-Molina. Estimating frequency of change. ACM Trans. Interet Technol., 3(3),

.2003

G. Cobena, S. Abiteboul, and A. Marian. Detecting changes in XML documents. In ICDE '02: [5] Proceedings of

.18th International Conference on Data Engineering, 2002

R. La-Fontaine. A Delta Format for XML: Identifying Changes in XML Files and Representing the [6] Changes in

.XML. In XML Europe, 2001

R. Song, H. Liu, J.-R. Wen, and W.-Y. Ma. Learning block importance models for web pages. In WWW [7] :'04

.Proceedings of the 13th international conference on World Wide Web, 2004

چکیده

آرشیو داران وب برای گردآوری منابع ویدئویی وب بیش از همیشه با ابزارها و پروتکلهای غیراستاندارد میزبانی میشوند این ،مقاله به وضعیت فعلی فناوری در این حوزه میپردازد. در این نوشتار، توجه به تجربیات چندین ساله گردآوری محتوای وب ویدئو به ذکر نمونه های مفصلی میپردازیم که به درک مسائل و راه حلهای محتوای ویدئویی در وب کمک میکنند ،همچنین به معرفی چارچوب معماری که برای مقیاس بندی گردآوری محتوای ویدئویی وب می پردازیم که بخشی از پروژه تحقیقاتی اتحادیه اروپا موسوم به LIWA(1)است.

ص: 38

1- http://liwa-project.eu

آرشیو منابع ویدئویی وب

نوشته: رادو پاپ،(1) گابریل واسیلی،(2) ژولین ماسانه(3)

ترجمه: فروزان رضائی نیا(4)

مقدمه

ویدئو، امروزه به عنوان بخش مهمی از وب شناخته شده است فناوری گسترش یافته به ویدئو کمک میکند که همواره بر نیاز صنعت رسانه مسلط باشد به ویژه زمانی که دسترسی مستقیم به فایلها توسط کاربران منع شده باشد. به عنوان مفهوم ،دیگر وظیفه جمع آوری مطالب آرشیو وب، بسیار سخت است و به توسعه رویکردها و ابزارهای خاص نیاز دارد. هدف این مقاله بررسی مشکلات اصلی آرشیو منابع ویدئویی وب است. براساس تجربیات به دست آمده در بنیاد آرشیو اروپا که در سالهای اخیر بر طیف متنوعی از این نوع محتوا در وب کار کرده است انواع حالتها در سال گذشته مشکلات مختلفی در ضبط بارگذاری و گردآوری ویدئو بوده که به دو دسته اصلی تقسیم و با استفاده از برخی مثالهای مناسب چندین راه حل فنی شرح داده می شود.

دسته نخست شامل وبگاههایی است که با استفاده از پروتکل استاندارد HTTP به ارائه محتوای

ویدئویی مبادرت میکنند.

1- Radu Pop

2- Gabriel Vasile

3- Julien Masanes

4- کارشناس سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

مشکل آنها شامل ناشی از فنون مختلفی است که از پیوندهای مبهم برای فایل ویدئویی استفاده میکنند (برای مثال 2 یا 3 مسیردهی مجدد(1) یا پرش(2)). نمونهٔ گویای این دسته مانند نماینده وبگاهYouTube است.

دسته دوم مشکلات در وبگاههایی نمودار میشوند که با استفاده از پروتکلهای انتقال دادهای غیر از حمل و نقل از HTTP نشان داده شده است از بین پروتکلهای مختلف پخش زنده ای(3) که در حال حاضر مورد استفاده قرار میگیرند جدیدترین و مشکل ترین آنها یعنی پروتکل RTMP را برگزیدیم. ذکر این نکته مهم است که فناوریهای استفاده شده برای میزبانی منابع ویدئویی در وب بسیار سریع دستخوش تحول میشود و به احتمال زیاد جزئیات مواردی که در این مقاله ارائه میشوند به سرعت در شرف تغییر است همچنین در این نوع نیز به احتمال زیاد تغییر سریع در جزئیات آنها ارائه شده است. با این حال قصد این است که جزئیات به اندازه کافی ذکر ،شوند تا درک منطق این فنون میسر .گردد. حتی اگر این جزئیات به سرعت در شرف کهنه شدن باشند نتیجه آنچه که امروز به یک بازی موش و گربه تبدیل شده آن است که ابزارهای ما هم در سطح گردآوری و هم در سطح دسترسی، باید به طور مداوم بهبود یافته و روزآمد شود.

در بخش دوم مقاله، طرحی برای معماری آرشیو منابع ویدئویی وب پیشنهاد میکنیم که قادر به انطباق سریع و همچنین مقیاس پذیری براساس تفکیک فرآیندهای مربوط به بارگذاری منابع ویدئویی از خزشگر است. این طرح، باعث بهبود کارآمدی هم از لحاظ مقیاس پذیری و هم از لحاظ انعطاف پذیری میشود چرا که ابزارهای بارگذاری را که توسط متخصصان چندرسانه ای توسعه یافته اند، آسانتر میتوان ادغام و روزآمد کرد بالاخره با استفاده از خزشگرها به صورت غیر همزمان بهتر میتوان کنترل خطا و مدیریت فرآیند را پشتیبانی کرد.

2 – نمونه هایی برای بارگذاری ویدئو

پیچیدگیهای مختلفی در بارگذاری منابع ویدئویی وجود دارد که در دو نمونه زیر با تفصیل بیشتری

بیان شده اند.

2-1- تغییر مسیر HTTP به YouTube.com

هر ویدئوی YouTube منحصراً با یک شناسه در هم سازی مشخص شده است(یک رشته از 11 کاراکتر) و عموماً میتواند در یک صفحه HTML با نشانی شبیه به زیر دسترس پذیر باشد

http://www.youtube.com/watch?v=uniqueID

احتمالاً سختترین چالش در برداشت ویدئوهای YouTube روزآمد کردن مکرر مکانیسمهایی است

که YouTube برای دسترسی ویدئوها در اختیار قرار میدهد و عبارت است از تلاش مستمربرای

Redirect -1

مخفی کردن نشانی مستقیم فایلهای ویدئویی.

با استفاده از روشهای گردآوری ،کلاسیک خزشگر باید 5 مرحلهٔ متمایز از پیوندهای مستقیم

یا غیر مستقیم را به منظور دسترسی به مطالب ویدئویی طی کند الگوهای کلی از نشانیهای حد واسط در جدول 1 خلاصه شده است.

جدول . الگوهای YouTube برای دسترسی به فایلهای ویدئویی

اگر شناسگر یک ویدئو (foobar) و آدرس صفحه 1) ouTube)،باشد خزشگر ابتدا URL فلش پلیر (2) صفحه را کشف می.کند با عبور از فهرست پارامترهـا به ،پلیر خزشـگر ممکن است عبـارت پرس وجـوی HTML را برای شـناسایی به کـار بـبرد که برای درخواست یک محتوای ویدئویی ارسال می،شود قبل از به دست آوردن URL فایل ویدئویی (5) خزشگر باید یک واسطه(1) تغییر مسیر را دنبال کند که حاوی توکن های(2) رمزگذاری شده باشد مانند نشانیهای IP از میزبان و برچسب زمان(3) در پرس وجو.

عکس

مخفی کردن نشانی مستقیم فایل‌های ویدئویی.

با اســـتفاده از روش‌های گردآوری کلاسیک،خزشـــگر باید ۵ مرحلهٔ متمایز ازپیوندهای مستقیم یا غیرمستقیم را به‌منظور دسترسی به مطالب ویدئویی طی کند. الگوهای کلی از نشانی‌های حد واسط در جدول ۱ خلاصه شده است.

جدول ۱. الگوهای YouTube برای دسترسی به فایل‌های ویدئویی

#	URL	mime-type	HTTP response
1	http://www.youtube.com/watch?v=foobar	text / html	200
2	http://s.ytimg.com/yt/swf/watch- vfl118818.swf	application /x-shockwave-flash	200
3	http://www.youtube.com/get_video?video_id=foobar&t=...	text/html	204 (OK no content)
4	http://v1.lscache4.c.youtube.com/videoplayback?ip=0.0.0.0&sparams=id...	Redirect	302
5	http://v1.cache2.c.youtube.com/videoplayback?ip=0.0.0.0&sparams=id...	video / x-lv	200

اگر شناسگر یک ویدئو (foobar) و آدرس صفحهٔ YouTube (#۱) باشد، خزشگر ابتدا URL فلش پلیر(#۲) صفحه را کشـــف می‌کند. با عبور از فهرســـت پارامترها به پلیر، خزشگر ممکن است عبارت پرس‌وجوی HTML (۵#) را برای شناسایی به‌کار ببرد، که برای درخواست یک محتوای ویدئویی ارسال می‌شـــود، قبل از به دست آوردن URL فایل ویدئویی (۵#)، خزشگر باید یک واسطهٔ[1] تغییر مسیر (۵#) را دنبال کند. که حاوی توکن‌های[2] رمزگذاری شده باشد، مانند نشانی‌هایIP از میزبان و برچسب زمان[3] در پرس‌وجو.

URL ویدئـــو(# ۵) در پارامترهای فلش، که هنگام بارگذاری صفحه(# ۱)به‌صورت دینامیک تولید شده‌اند یکپارچه سازی می‌شوند.

مسئله‌ای که خزشگر با آن مواجه است، شناسایی صحیح این URL است، چراکه کاراکترهای مختلفی از آن حذف شده‌اند و پارامترهای مختلف دیگری توسط شیء فلش به آن الصاق شده‌اند.

به‌عنوان نمونه، در نســـخه فعلی صفحه‌های YouTube ، پارامتر«flashvars» رشته‌ای[4] به‌طول ۶۳۷۴ کاراکتر اســـت، در حالی‌که، URL آن می‌تواند ۳۹۲ کاراکتر داشـــته باشد. در تجزیه و تحلیل دقیق‌تر از

1 Intermediary.
2. Tokens
3.Time-stamp
4. String

URL ویدئو(5) در پارامترهای فلش که هنگام بارگذاری صفحه به صورت دینامیک تولید

شده اند یکپارچه سازی میشوند.

مسـئله ای که خزشگر با آن مواجه است شناسایی صحیح این URL است چراکه کاراکترهای مختلفی از آن حـذف شده اند و پارامترهای مختلف دیگری توسط شیء فلش به آن الصاق شده اند.

به عنوان نمونه در نسخه فعلی صفحه های YouTube پارامتر «flashvars» رشته ای(4) به طول 6374 کاراکتر است در حالی که URL آن میتواند 392 کاراکتر داشته باشد در تجزیه و تحلیل دقیق تر از

1- Intermediary

2- Tokens

3- Time-stamp

4- String

صفحه (1) URL ویدئو (5) در قطعه(1) پردازه جاوا(2) شناسایی میشود که پارامتر فلش را ایجاد میکند در این نسخه صفحه های YouTube متنی که باید تجزیه شود حاوی رشته PLAYER_CONFIG » و به دنبال آن دارای فهرستی از URLها به صورت تصادفی است. در بین دو کارکتر «[» میتوان URL ویدئو را برداشت ،کرد که توکنهای محاسبه شده را براساس نشانی IP و برچسب زمان در خود دارد.

همان طور که ملاحظه میشود مسیری که برای برداشت فایلهای ویدئویی باید طی کرد، بسیار پیچیده است و URLهای منابع ویدئویی با تغییر مسیرهای مجدد و اضافه شدن توکنهای موقتی پیچیده شده اند.

در خصوص پارامترهای خزشگر میتوان خزش را برای 5 سطح در صفحه YouTube تنظیم کرد.

علاوه بر این URLها در و 5 به زیر دامنه های مختلف باز میشوند یا ارجاع میدهند، بنابراین باید با صراحت به دامنههای خزشگر اضافه شوند.

بالاخره مشکل بزرگی که به علت URLهای تغییر مسیر داده به وجود می آید، به ابزارهای دسترسی آرشیو مربوط است. حتی اگر هر فایل ویدئویی در YouTube یک URL متفاوت به صورت پویا در هر بارگذاری تولید می شود. به همین دلیل، باید به نمایه URLها برای ردیابی بین صفحه اصلی (1) و URL فایل ویدئو (5) اضافه شود.

در عمل، دو روش برای ضبط فایلهای ویدئویی در YouTube وجود دارد: 1) ضبط ویدئویی برخط، 2) فنون بارگذاری غیر برخط فایلهای ویدئویی

با توجه به فرآیند ،خزش در روش برخط فایلهای ویدئو در زمان خزش بارگذاری می شوند و این کار با استفاده از یک پردازنده اضافی به خزشگر Heritrix صورت می پذیرد عنوان مثال اسکریپت BeanShell که توسط آدام تیلور(3) نوشته شده، همه URLهای واسطه (در 5-2) را به فهرست خزش(4) میفرستد چرا که اغلب خارج از دامنه مطلوب برای خزش هستند(s.ytimg.com,v.cache2.c.youtube.com.) فایلهای ویدئویی به فایلهای WARC اضافه میشوند که توسط خزشگر ایجاد شده.

در روش غیر بر خط فایلهای ویدئویی پس از مرحله پردازش بر اساس URL صفحه های YouTube می شوند. در این روش از یک بارگذاری کننده بیرونی استفاده می کنند، که یک نمونه آن بارگذاری کننده توسعه یافتهها توسط ریکاردو گارسیا گونزالس(5) است که محتوای ویدئویی را بارگذاری و به فایلهای flv منتقل می.کند. کاربرد ابزار(6) WARC ، فایلهای flv به فایلهای مجزای WARC رابسته بندی میکنند.

هر دو روش نیاز به ایجاد یک پیوند از URL اصلی ویدئو دارند همان گونه که در صفحه وب ظاهر میشود و نام فایل یا URL جدیدی که به برای محتوای ویدئو باز میشود چون فایل ویدئویی باتعقیب

JavaScript –2

http://webarchive.jira.com/wiki/display/Heritrix/BeanShell+Script+For+Downloading+Video –3

Frontier –4

http://bitbucket.org/rg3/youtube–dl/wiki/Home –5

http://code.google.com/p/warc–tools –6

URLهای واسطه در آرشیو ذخیره میشود).

مزیت روش برخط این است که بارگذاری فایلهای ویدئویی توسط خود خزشگر انجام میشود و

هیـچ ابزار بیرونی برای پایش و همزمانی وجود ندارد. علاوه بر این همه سـرآیندهای HTTP در آرشـیو در تعامل با سرورهای YouTube ذخیره می.شوند اشکال این روش این است که URL نهایی محتوای ویدئویی (5) دیگر شناسگر اولیه ویدئو را در خود ندارد (1) مدیریت ردیابی شناسـگر ویدئویی دشوار است؛ همچنین آرشیوها به همه URLهای پرشی(1) آلوده میشوند(این URLها دیگر معتبر نیستند چرا که توکنهای بارگذاری اعتبار موقت دارند) .

از سوی دیگر رویکرد غیر برخـط در پـایش و مـدیریت بارگـذاری کننـدهای بیرونی انعطـاف پذیرترنـد به عنوان مثال کنترل خطا) فایلهای ویـدئویی نـام شناسـگرهای اولیه خود را حفـظ میکننـد و URLهـای پرشـی در آرشـیو ذخیره نمیشونـد لازم است یک سـرآیند HTTP ساختگی(2) (برای هر فایلflv)در فایلهای WARC وارد شود زیرا بارگذاری کننده بیرونی پاسخ سرور را حفظ نمیکند.

RTMP streaming on SWR.de -2-2

-2-2-1- مروری بر پروتکلهای داده در جریان

Streaming، که با اسـتانداردهای کارگروه مهندسـی اینترنت (IETF) تطـابق دارد اجازه میدهـد تا سـرور تبادلات را کنترل کنـد و برای نگهـداری موجودیتها در وضـعیت بهینه سازی شـده است. لازم نیست کاربران فایلهای عظیم را بارگذاری کنند و این رویکرد به خصوص برای اطلاع رسانی و پخش زنده مناسب است.

در واقع پروتکل داده در جریان از دو نوع پروتکل داده در جریان به سـرعت اسـتفاده میکند:(3) [RTP 3550 برای ارسال بسته های داده رسانه ای و(4) RTSP [RFC2326] برای کنترل اطلاعات RTP از UDP اسـتفاده میکند که بسته های گمشده را دوباره منتقل نمیکند ،بنابراین حمل بر این است که همه طرفها پـذیرفته اند که هنگام انتقال بعضـی از بسـته ها ناپدید شوند این بدان معنی است که کاربران باید موقرانه عدم دریافت همه داده مربوط به یک ویدئو و یا محتوای شنیداری را بپذیرند و خودشان مدیریت کنند. این نسبت به رویکرد مبتنی بر TCI/IP ترجیـح داده میشود که گرفتن بسـته های گـم شده ممکن است مجبور باشد دفعات نامعینی تلاش کند و بنابراین زمان نامعینی نیز وقت لازم خواهـد بود. RTSP نوعی پروتکل کنترل شـبکه ای برای اسـتفاده در صـنعت سـرگرمی و سیسـتمهای ارتباطی به منظور کنترل سرورهای رسانه streaming .است این ،پروتکـل برای ایجـاد و کنترل تراکنش رسـانه ای بین کـاربران است کاربران سـرورهای رسانه ای فرمانهای VCR مانند پخش و توقف برای تسهیل کنترل زمان بلادرنگ پخش فایلهای رسانه از سرورها را ارسال میکنند.

1- Redirect URLs

2- Fake

3- Real-Time Sreaming Protocol

4- Real - Time Transport Protocol

پروتکل RTSP به HTTP شباهت دارد با این تفاوت که RTSP درخواستهای جدیدی را اضافه می.کند HTTP فاقد هویت است. حال آنکه RTSP کاملاً دارای هویت است شناسه تراکنش برای ردگیری در مواقع لزوم مورد استفاده قرار میگیرد از این رو هیچ گونه ارتباط TPC دائمی مورد نیاز نیست. پیامهای RTSP از کاربر به سرور فرستاده میشود هر چند استثنائاتی وجود دارد و گاه سرور به کاربر پیام می فرستد.

سرویس پیام رسانی چندرسانه ای (1)(MMS)

یک استاندارد ارتباطات از راه دور برای ارسال پیام به وسیله ابزارهای چند رسانه ای (عکس، صدا، تصویر و متن است. MMS توسعه یافته استاندارد SMS است که پیامهای طولانی تر را با استفاده از WAP برای نمایش محتوای پیامها را امکان پذیر می.کند پیامهای MMS در یک روشی تقریبا مشابه SMS تحویل داده میشوند اما محتوای چندرسانه ای ابتدا کدگذاری و در یک پیامی متنی به شیوه ای شبیه به ارسال ایمیلMIME درج میشود.

RTMP یک پروتکل دارای حق مالکیت است که توسط شرکت سیستمهای Adobe برای جریان دادههای صوتی، تصویری و اطلاعات موجود در اینترنت بین فلش پلیر و سرور گسترش یافته است. این پروتکل برای تضمین تحویل جریان داده صوتی و تصویری در عین حفظ ارسال حجم بیشتری از ،اطلاعات داده و تصاویر را از هم مجزا می.کند اندازه قطعات میتواند به صورت پویا به وسیله کاربر و سرور تعیین شود، و حتی این امکان را میتوان پشت صفحه در صورت تمایل غیر فعال کرد.

شرکت سازنده Adobe مشخصات پروتکل RTMP را در 15 ژوئن 2009(2) در دسترس عموم قرار ،داد اما این به نظر میرسد در این مشخصات بسیاری از جزئیات مربوط به پیاده سازی پروتکل ارائه

نشده است.

2 2 – 2 – ضبط داده تصویری در جریان

بر اساس یک خزش در اینترنت که توسط European Archive برای ضبط تصاویر در حال پخش انجام شده است جزئیات فنی ای را که از وبگاه SWR.de گرفته شده اند ارائه میکنیم.

صفحه آرایی صفحه های ویدئو

ساختار نمایشی صفحه وبی که ویدئو نمایش میدهد از الگویی مشابه که مثال قبلی در YouTube.com پیروی می.کند صفحه HTML شامل پانلهای اصلی از جمله پخش ویدئو (یعنی JW فلش ویدئو پلیر با کد منبع باز از ویدئو Long Tail) است. محتوای هر صفحه HTML به صورت پویا به وسیله سرور ایجاد و در عناصر خاص HTML ذخیره میشود مانند:

1- Multimedia Messaging Service

2- http://www.adobe.com/devnet/rtmp

<

«type=>>application/x-shockwave-flash

<<id=>>player46

/data=http://www.swr.de/static

<"jwplayer/player46.swf

...

فلش پلیر در یک عنصر object> جاسازی شده و از طریق سرور وب بارگذاری میشود همه پارامترهایی که برای فلش پلیر مورد نیاز هستند به وسیله JavaScript آماده و (با استفاده از پارامترهای فلش به شیء فلش منتقل میشوند از اینجا همه تعاملات با سرور داده در جریان به طور مستقیم به وسیله فلش پلیر انجام میشود

شکل –1 وبگاه زنده – ویدئوی ATMP

بارگذاری فیلم در حال پخش

در این مثال ،ویژه URL فایل تصویری به صورت واضح در پردازهٔ جاوا نوشته شده است. از دید خزشگر، این نمونه مناسبی است چون برداشت کننده(1) پردازهٔ جاوا قادر به شناسایی و برداشت URL فایلهای تصویری است اما این URL برای خزشگر URL معتبری نیست زیرا برنامه های پروتکل HTTP/HTTPS را پشتیبانی نمیکند.

از این رو URL مربوط به RTMP به عنوان یک URL غیر معتبر در رخدادهای خطای خزشگر ارائه گزارش داده میشود برای بارگیری بهتر فایلهای تصویری (URL (rtmp://.../foobar-video.flvها RTMP از رخدادهای خطا خارج و به بارگذاری کنندگان بیرونی واگذار میشود (مانند FLVStreamer). بارگذاری کننده RTMP محتوای تصاویر را در فایلهای flv که در فایل WARC بسته بندی شده اند کپی می کند.

عکس

شکل ۱- وبگاه زنده ـ ویدئوی ATMP

بارگذاری فیلم در حال پخش

در این مثال ویژه، URL فایل تصویری به صورت واضح در پردازهٔ جاوا نوشته شده است. از دید خزشگر، این نمونه مناسبی است چون برداشت‌کنندهٔ[1] پردازهٔ جاوا قادر به شناسایی و برداشت URL فایل‌های تصویری است. اما این URL برای خزشگر، URL معتبری نیست، زیرا برنامه‌های پروتکل / HTTP HTTPS را پشتیبانی نمی‌کند.

از این رو، URL مربوط به RTMP به‌عنوان یک URL غیرمعتبر در رخدادهای خطای خزشگر ارائه گزارش داده می‌شود. برای بارگیری بهتر فایل‌های تصویری (rtmp://.../foobar-video.flv)، URLها RTMP از رخدادهای خطا خارج و به بارگذاری‌کنندگان بیرونی واگذار می‌شود (مانند FLVStreamer). بارگذاری‌کننده RTMP، محتوای تصاویر را در فایل‌های flv که در فایل WARC بسته‌بندی شده‌اند، کپی می‌کند.

دسترسی به آرشیو محتوای ویدئویی

از نقطه نظر دسترسی، تفاوت اساسی بین صفحه وب زنده و صفحهٔ وب آرشیو شده، استفاده از پروتکل انتقال برای تحویل تصاویر است. **قرار دادن سرور** برای دادهٔ جریان در زیرساخت آرشیو، مستلزم تلاش فراوان و راه‌حلی پرهزینه برای توسعه و نگهداری از آن است. همچنین، سرورهای مختلفی برای پروتکل‌های مختلف مورد نیاز خواهد بود. به‌طور کلی، دسترسی به آرشیو محتوای ویدئویی از طریق پروتکل HTTP و فایل‌های flv، به‌طور مستقیم از فایل‌های WARC و صفحه های HTML و دیگر

1. Extractor

دسترسی به آرشیو محتوای ویدئویی

از نقطه نظر دسترسی تفاوت اساسی بین صفحه وب زنده و صفحه وب آرشیو شده، استفاده از پروتکل انتقال برای تحویل تصاویر است قرار دادن سرور برای داده جریان در زیرساخت آرشیو مستلزم تلاش فراوان و راه حلی پرهزینه برای توسعه و نگهداری از آن است. همچنین سرورهای مختلفی برای پروتکلهای مختلف مورد نیاز خواهد بود به طور کلی دسترسی به آرشیو محتوای ویدئویی از طریق پروتکل HTTP و فایلهای flv ، به طور مستقیم از فایلهای WARC و صفحه های HTML و دیگر

Extractor -1

منابع ایجاد می شود. در حال حاضـر دسترسـی به کارکردهای خاص جریان داده وجود ندارد (مانند fastforwar یا بارگیری حجیم) اما دسترسی پایه به محتوا قطعی است و آن را میتوان بدون هزینه اضافی برای نمونه های مختلف تنظیم کرد.

ایراد مهم در جـایگزینی داده در جریان با پروتکل HTTP ساده در ویـدئوهای بزرگ قابل مشاهـده است و اگر هیچ گونه تنظیمی صورت نگیرد اجراکننده باید تمام فایلهای flv را قبل از اجرای تصاویر بارگذاری کند ما ابزارهای دسترسـی را در EA برای کمتر کردن این مشکل را به وسـیله بارگـذاری زیـادی از فایلهـای ویـدئویی از آرشـیو تنظیم کردیم. حجم این تصاویر بهینه سازی می شود، اما بایـد تعادلی بین دسترسی سریع برای اجرای ویدئوها و پیچیدگی روشهایbuffering انجام شود.

در زمان دسترسـی، تنظیم اصـلی که باید صورت گیرد جایگزینی فلش پلیر اصـلی است. چون فایل ویدئویی آرشـیو شده دیگر در جریان نیست اجراکننده اولیه در صـفحه نمیتواند برای نسـخه آرشـیو شده مورد استفاده قرار گیرد؛ از این رو قالب HTML باید به تناسب روزآمد شود:

flashvars=file=http://collection.europarch

.ive.org/swr/...7.1.AEVA/rtmp://fcondemand

swr.de/at/e/foobar-video.flv

src=http://collections.europarchive.org/me

dia/player.swf

type-application/x-shockwave-flash

<id-video645568_plr

در مقـــایسه بـــا صـــفحه ،زنـــده در یـــک نســـخه آرشـــیوی عناصـــر script و object> را بـــا عنصـــر خـــاص

که حاوی اجراکننده خاص آرشیو است و URL تصاویر آرشیوی جایگزین میکنیم:

"http://collections.europarchive.org/media/player.swf"

http://collection.europarchive.org/swr/20100601084708/rtmp://.../foobar-video.flv

ص: 47

شکل 2

توجه داشته باشیم که URL فایل ویدئویی آرشیوی به HTTP URL که به فایل flv در آرشیو باز

می شود تبدیل شده اند.

جایگزینی عناصر HTML با شیوه های خاصی که در کد دسترسی در سرور پیاده سازی صورت می گیرد. فایلهای WARC همیشه نسخه اصلی صفحه ها را نگه میدارد پیدا کردن یک الگوی مشترک برای شناسایی صحیح و جایگزینی عناصر حاوی اجراکننده هنوز چالش برانگیز ست، چرا که ساختار و ویژگیهای یک عنصر object> ممکن است از یک وبگاه به وبگاه دیگر متفاوت باشد.

در این بخش ما به گرفتن نمونهٔ فیلمهای RTMP ،پرداختیم زیرا فرآیند بارگذاری مستلزم پروتکل خاص داده در جریان است. با این حال فنون جایگزینی اجراکننده که در کد دسترسی پیاده می شوند در مورد فیلمهای بارگذاری شده در پروتکل HTTP به همان شیوه انجام میگیرد.

عکس

شکل ۲

توجه داشته باشیم که URL فایل ویدئویی آرشیوی به HTTP URL که به فایل flv در آرشیو باز می‌شود تبدیل شده‌اند.

جایگزینی عناصر HTML، با شیوه‌های خاصی که در کد دسترسی در سرور پیاده‌سازی صورت می‌گیرد. فایل‌های WARC همیشه نسخه اصلی صفحه‌ها را نگه می‌دارد. پیدا کردن یک الگوی مشترک برای شناسایی صحیح و جایگزینی عناصر حاوی اجراکننده هنوز چالش برانگیزست، چرا که ساختار و ویژگی‌های یک عنصر <object> ممکن است از یک وبگاه به وبگاه دیگر متفاوت باشد.

در این بخش ما به گرفتن نمونهٔ فیلم‌های RTMP پرداختیم، زیرا فرآیند بارگذاری مستلزم پروتکل خاص داده در جریان است. با این حال، فنون جایگزینی اجراکننده که در کد دسترسی پیاده می شوند در مورد فیلم‌های بارگذاری شده در پروتکل HTTP به همان شیوه انجام می‌گیرد.

۳- ضبط ویدئو با استفاده از بارگذاری‌کننده خارجی

به‌عنوان جزئی از فناوری‌های جدید در پروژه LiWA[1]، ماژول خاصی برای افزایش قابلیت‌های ضبط توسط خزشگر با توجه به انواع محتوای چندرسانه‌ای طراحی شد [در مورد نخستین تلاش‌هایی که در این مورد در EA انجام شد نگاه کنید به Baly ۲۰۰۶] . نسخه فعلی Heritrixمبتنی بر پروتکل HTTP/ HTTPS است و نمی‌تواند پروتکل‌های دیگر را که در چندرسانه‌ای‌ها (نظیر داده در جریان) مورد استفاده قرار می‌گیرند، اجرا کند.

ماژول Liwa Rich Media Capture[2] بازیابی محتوای چندرسانه‌ای را به یک برنامه کاربردی خارجی

1. http://www.liwa-project.eu/
2. The LiWA Rich Media Capture module, http://code.google.com/p/liwatechnologies/source/browse/rich-media-capture

3- ضبط ویدئو با استفاده از بارگذاری کننده خارجی

به عنوان جزئی از فناوریهای جدید در پروژه(1) LIWA ، ماژول خاصی برای افزایش قابلیتهای ضبط توسط خزشگر با توجه به انواع محتوای چندرسانه ای طراحی شد در مورد نخستین تلاشهایی که [در این مورد در EA انجام شد نگاه کنید به Baly 2006] . نسخه فعلی Heritrixمبتنی بر پروتکل HTTP HTTPS است و نمیتواند پروتکلهای دیگر را که در چندرسانه ایها (نظیر داده در جریان) مورد استفاده قرار میگیرند اجرا کند.

ماژول(2) Liwa Rich Media Capture بازیابی محتوای چندرسانه ای را به یک برنامه کاربردی خارجی

1- http://www.liwa-project.eu/

2- The LiWA Rich Media Capturemodulehttp://code.google.com/p/liwatechnologies/source/browse/rich-media-capture

واگذار می کند (مانند MPlayer (1)یا(2) FLVStreamer) که قادر است طیف وسیعی ازپروتکلهای انتقال را مدیریت کند.

این ماژول به عنوان پلاگین(3) خارجی برای Heritrix ساخته شده است. در این رویکرد، شناسایی و بازیابی دادههای در جریان کاملاً از هم جدا شده اند و امکان استفاده از ابزارهای کارآمدتر برای تجزیه و تحلیل محتوای ویدئویی و شنیداری فراهم شده است.

-3-1- معماری

ماژول از چندین جزء فرعی ساخته شده است که از طریق پیام ارتباط برقرار میکنند ما از پروتکل ارتباط استانداردی استفاده میکنیم که(4) AMQP نامیده میشود.

تلفیق ضبط رسانه های غنی(5) با خزشگر در تصویر 3 نشان داده شده و گردش کار پیامها را به صورتی که ذکر میشود میتوان خلاصه کرد پلاگین متصل به هریتریکس URL منابع دادههای در جریان را شناسایی و برای هر کدام از آنها یک پیام AMQP ایجاد میکند این پیام به یک سرور پیام مرکزی منتقل میشود و نقش سرور پیام این است که هریتریکس را از بارگذاری کننده های خوشه بندی شده داده در جریان مجزا کند یعنی همان ابزارهای بارگذاری کنندۀ خارجی سرور پیام یوآرال را در صف(6) میکند و زمانی که یکی از بارگذاری کننده ها در دسترس باشد URL بعدی را برای پردازش میفرستد.

در معماری ماژول سه ماژول فرعی را شناسایی میکنیم:

- ماژول کنترل اولیه مسئول دسترسی به سرور ،پیام آغاز کارهای جدید متوقف کردن آنها و ارسال هشدار؛
- ماژول دوم که برای شناسایی و بارگذاری دادههای در جریان استفاده میشود (از یک ابزار خارجی مانند MPlayer استفاده میشود)
- ماژول سوم که داده در جریان بارگذاری شده را در فرمتی که به وسیله ابزارهای دسترسی قابل شناسایی باشند دوباره بسته بندی میکند (ضبط کننده WARL)

1- http://www.mplayerhq.hu

2- http://savannah.nongnu.org/projects/flvstreamer/

3- Plugin

4- Advanced Message Queuing Protocol(AMQP),http://www.amqp.org/confluence/display/AMQP/Advanced+Message+Queuing+Protocol

5- Rich Media

6- Queue

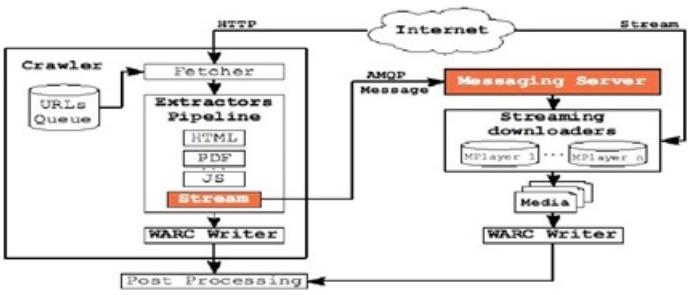
تصویر3 – ماژول ضبط داده در جریان در تعامل با خزشگر

بارگذاری کننده ،فیلم به سرور پیام دهی متصل میشود تا URL فیلم جدیدی را برای ضبط کردن درخواست کند به محض دریافت URL جدید یک تحلیل اولیه به منظور شناسایی بعضی پارامترها انجام می شود از جمله نوع و طول زمان جریان داده ،البته اگر فیلم زنده باشد ممکن است زمان ثابت و قابل پیکر بندی در نظر گرفت.

بعد از یک شناسایی ،موفق بارگذاری واقعی شروع میشود ماژول کنترل برنامه کاری ای را ایجاد میکند که همراه با تمهیدات حفاظتی به MPlayer منتقل میشود تا مدت زمان بارگذاری از تخمین اولیه طولانی تر نشود بعد از ضبط موفقیت آمیز آخرین مرحله بسته بندی فیلمهای ضبط شده در فایل WARC است که بعد از آن به مرحلهٔ ذخیره سازی نهایی هدایت میشود.

2-3- بهینه سازی(1)

عکس



تصویر ۳- ماژول ضبط دادهٔ در جریان در تعامل با خزشگر

بارگذاری‌کنندهٔ فیلم، به سرور پیام دهی متصل می‌شود تا URL فیلم جدیدی را برای ضبط کردن درخواست کند به محض دریافت URL جدید، یک تحلیل اولیه به منظور شناسایی بعضی پارامترها انجام می‌شود از جمله نوع و طول زمان جریان داده. البته، اگر فیلم زنده باشد ممکن است زمان ثابت و قابل پیکربندی در نظر گرفت.

بعد از یک شناسایی موفق، بارگذاری واقعی شروع می‌شود. ماژول کنترل، برنامهٔ کاری‌ای را ایجاد می‌کند که همراه با تمهیدات حفاظتی به MPlayer منتقل می‌شود تا مدت زمان بارگذاری از تخمین اولیه طولانی‌تر نشود. بعد از ضبط موفقیت‌آمیز، آخرین مرحله بسته‌بندی فیلم‌های ضبط شده در فایل WARC است که بعد از آن به مرحلهٔ ذخیره‌سازی نهایی هدایت می‌شود.

۲-۳- بهینه‌سازی[1]

مسائل اصلی که از آزمایش‌های اولیه مطرح می‌شوند، همزمان کردن خزشگر و ماژول ضبط خارجی است. در مورد مجموعه‌ای از فیلم‌های پرحجم در وبگاه، بارگذاری پی‌درپی فیلم قطعاً از فرآیند جست‌وجوی صفحه‌های متن طولانی‌تر خواهد بود. و خزشگر باید منتظر باشد تا ماژول بیرونی بارگذاری فیلم را تمام کند. با افزایش بارگذاری‌کننده‌ها می‌توان سرعت بارگذاری تصاویر را افزایش داد. از سوی دیگر، اگر بخواهیم فرآیند به‌صورت موازی انجام شود.

راه‌حل دیگر برای کنترل بارگذاری‌کننده‌های فیلم این است که ماژول ضبط فیلم را از خزشگر جدا کنیم و آن را در مرحله بعد از پردازش به کار بیندازیم. این به معنای جایگزینی پلاگین خزشگر با یک ثبت‌کننده ورود به سیستم و یک مدیر مستقل برای بارگذاری‌کننده‌های فیلم است.

از مزایای این روش(که برای نمونه در EA استفاده شده است) عبارت‌اند از:

- نگرشی همه جانبه از تعداد کل URLهای فیلم

1. Optimizations

مسائل اصلی که از آزمایشهای اولیه مطرح می شوند، همزمان کردن خزشگر و ماژول ضبط خارجی است. در مورد مجموعه ای از فیلمهای پرحجم در وبگاه بارگذاری پی در پی فیلم قطعاً از فرآیند جست و جوی صفحه های متن طولانی تر خواهد بود و خزشگر باید منتظر باشد تا ماژول بیرونی بارگذاری فیلم را تمام کند. با افزایش بارگذاری کننده ها میتوان سرعت بارگذاری تصاویر را افزایش داد. از سوی دیگر، اگر بخواهیم فرآیند به صورت موازی انجام شود.

راه حل دیگر برای کنترل بارگذاری کننده های فیلم این است که ماژول ضبط فیلم را از خزشگر جدا کنیم و آن را در مرحله بعد از پردازش به کار بیندازیم. این به معنای جایگزینی پلاگین خزشگر با یک ثبت کننده ورود به سیستم و یک مدیر مستقل برای بارگذاری کننده های

فیلم است.

از مزایای این روش (که برای نمونه در EA استفاده شده است) عبارت اند از:

- نگرشی همه جانبه از تعداد کل URLهای فیلم

Optimizations -1

• مـدیریت بهتر منابع تعـداد بارگذاری کننده های فیلم که مشترکاً از پهنای باند اسـتفاده می کنند). اشـکال اصـلی این روش ناهماهنگی است که ممکن است بین زمان خزش از وبگاه و ضبط فیلم در مرحله بعد از پردازش ظاهر شود.

• برخی محتویات فیلم ممکن است ناپدید شود (با فاصله یک یا دو روز)

• بارگذاری فیلم در انتظار پایان یافتن خزش متوقف شود.

بنـابراین بایـد تعـادلی هنگـام مـدیریت بارگـذاری فیلم برقرار کرد کوتاه کردن زمان بارگـذاری کامل کنترل اشـتباهات (برای فیلمهایی که سرورهایشان کند است) و بهینه سازی پهنای باند که توسط بارگذاری کننده های مختلف استفاده میشود.

4- نتیجه گیری و چشم انداز

همانگونه که ملاحظه شـد ارائه راه حلی کلی برای پرداختن به تمام وبگاهها که محتوای ویـدئویی دارند مشـکل است. براساس روش ارائه شـده در این مقاله، بایـد برای هر مورد خاص تکنیک گردآوری خاصـی به کار گرفته شود. مهندسـی خزش برای تنظیم با ابزارها بسـتگی به پیچیدگی وبگاه دارد.

کارهایی که در این زمینه میتوان انجام داد در سه گروه جای می گیرد:

• مقیاس بندی ضبط ویدئو به احتمال زیاد با جدا کردن آن از خزش و مدیریت بهتر خطاهای متعدد و وقفه هایی که سرورهای ویدئویی به طور کلی و سرورهای جریان به طور خاص ایجاد میکنند.

• بهبود کشف خودکار از پیونـدهای سـردرگم و تعقیب آنها با قواعد خاص (allowing off-domains کشف منـابع در قالبهایی غیر از Html و....).

• توسعه روشی عام برای دسترسی و ارائه این مستلزم تشخیص ویژگی نمایش دهنده ها به طور خودکار ،است به گونه ای که آنها را به شیوه ای عام جایگزین کند دسترسی بهتر را مدیریت و گزینه هایی برای فایلهای بزرگ و غیره در اختیار قرار دهد.

5- تقدیر و تشکر

بخشی از هزینه های این کار توسط کمیسیون اروپا زیر نظر LIWA تامین اعتبار شده است.

منابع

Baly, N., Sauvin, F. (2006). Archiving Streaming Media on the Web, Proof of Concept

.and First Results. International Web Archiving Workshop (IWAW 06), Alicante, Spain

چکیده

توسعه خدمات اطلاعاتی و محتوایی در وب و از دیگر سو توسعه فناوریهای ارتباطی و معرفی زیر ساختهای نرم افزاری سخت افزاری و مدلهای جدید مدیریتی در سطوح مختلف سازمانها باعث افزایش روزافزون نیاز به ایجاد زیرساختهای لازم جهت پشتیبانی از ذینفعان مختلف این حوزه شده است نیاز به توسعه دسترسی در سطح ،وب درخواست جستجوهای ،پیشرفته وابستگی به خدمات معنایی تقاضای بهبود رابطها و محیطهای تعاملی و مبتنی به مدل برای کاربران رشد رویکردهای بهینه سازی روشهای گردآوری و ذخیره سازی اطلاعات در وب از جمله مؤلفه هایی است که توجه به زیر ساختهای متناسب و نوین در این حوزه را به سرعت مطرح مینماید گسترده شدن نوع خدمات اختصاصی در سازمانها که منجر به گسترده شدن زنجیره های ارزش شده است و همچنین توسعه سازمانهای مجازی و افزوده شدن حجم تعاملات الکترونیکی درون و برون سازمانهای مختلف باعث شده استکه قابلیت تعامل پذیری در سطوح مختلف ،مدیریتی فرایندها، خدمات و دادگان به عنوان یکی از حوزه های اصلی پژوهش مطرح شده و از جنبه های مختلف مورد بررسی قرار گیرد نگاه سلسله مراتبی فوق به این تعاملات علاوه بر تحقق خدمات مشترک یکپارچگی لازم در تمام فضای زنجیره ارزش را تا فیزیکی ترین تعاملات داده ای و نرم افزارهای کاربردی ایجاد مینماید در حوزه تعاملات نرم افزاری به عنوان یکی از زیر ساختهای فنی مهم در سازمانها پشتیبانی از نیازمندیهای سیستمی نظیر پروتکلهای ارتباطی واسطهای ارتباطی دسترسی به دادگان گونه های ،اطلاعاتی معناشناسی پارامترها ایجاد قابلیتهای تحرک و انجام وظایف کاربردی به عنوان اولویت اصلی شناخته شده است. یکی از شیوه های تحقق زیر ساخت نرم افزاری مناسب برای تعاملات سازمانی بکارگیری عاملهای هوشمند است.

عاملهای هوشمند، سیستمهای نرم افزاری خود ،مختار واکنش گرا پیش فعال و با توانایی برقراری ارتباط هستند که می توانند به عنوان اجزاء کلیدی خدمات اطلاعاتی و محتوایی مبتنی بر وب در سامانه های مختلف مورد استفاده قرار گیرند. بسیاری از خدمات محتوایی تحت وب از جمله فرایندهای مدیریتی توزیع شده اشتراک گذاری محتوا و یکپارچه سازی خدمات و محتوا از مخازن و سامانههای مدیریت محتوای مختلف با استفاده از عاملها محقق میشوند و علاوه بر صرفه جویی در زمان و هزینه انعطاف لازم جهت پشتیبانی از رخدادهای غیر منتظره را فراهم می آورند به کارگیری عاملهای نرم افزاری همانگونه که در محیط درون سازمان زیر ساخت مناسبی برای انجام مؤثر و کارای تعاملات ایجاد مینماید در فضای برون سازمانی نیز تأثیرات زیادی خواهد داشت به دلیل تفاوت در دامنه کاری و فناوریهای مورد استفاده در هر سازمان لازم است که روشهایی که برای تبادل و به اشتراک گذاری داده میان آنها استفاده میشود مستقل از هر نوع فناوری خاصی بوده و امکان تولید و به اشتراک گذاری اطلاعات سازگار و با معنی را داشته باشند برای تبادل مؤثر دادهها باید درک مشترکی از داده میان سازمانها وجود داشته باشد که این امر با استفاده از روشهای هستان شناسی محقق می.شود پیشرفتهای وب 2 و امکان به اشتراک گذاری و درک داده ها توسط موتورهای جستجو از جمله دستاوردهای استفاده از روشهای هستان شناسی است. از مجموع مباحث فوق میتوان نتیجه گرفت که استفاده از عاملها نه تنها در فرایندهای درون سازمان منجر به افزایش کارایی و اثربخشی میشوند؛ بلکه در افزایش کیفیت رابطه با نهادهای برون سازمانی از قبیل تأمین کنندگان، مشتریان، رقبا یا شرکای احتمالی تأثیر بسزایی دارند و استفاده از معماری مبتنی بر عامل میتواند ضمن ایجاد صرفه اقتصادی برای سازمانها به عنوان مزیت رقابتی ماندگار محسوب شود. در این مقاله ضمن معرفی کاربرد عاملهای نرم افزاری در سازمانها، به معرفی چارچوبی مبتنی بر عاملها برای یکپارچه کردن تعاملات اطلاعاتی در آنها پرداخته میشود همچنین در انتها با استفاده از یک استاندارد شناخته شده در کنسرسیوم جهانی ،توسعه همگرایی و پذیرش خدمات تحت وب و کسب و کار الکترونیکی(1) OASIS ، نشان داده خواهد شد که چگونه میتوان چارچوب پیشنهادی را برای پشتیبانی تعامل پذیری در سیستمهای مدیریت محتوا و مخازن اطلاعاتی استفاده کرد و ضمن بهره مندی از تعامل گسترده میان منابع و خدمات مختلف در سطح ،وب، از قابلیتهای عاملهای هوشمند برای بهبود این گونه خدمات استفاده کرد.

کلیدواژه تعامل ،پذیری خدمات اطلاعاتی عامل نرم افزاری زنجیره ،ارزش سامانه تعامل پذیری مدیریت محتواCMIS

Organization for Advancement of Structured Information Standard -1

استفاده از عاملهای هوشمند نرم افزاری جهت ایجاد قابلیت تعامل پذیری درخدمات محتوایی و اطلاعاتی سازمانها

محمود خراط(1) | مائده مشرف(2) | فتانه تقی یاره(3)

1. مقدمه

استفاده از فناوری اطلاعات در نهادهای مختلف منجر به رخداد تغییرات بسیاری در درون سازمانها و در طول زنجیره ارزش شده .است اگرچه این فناوری مزیتهای بسیار زیادی برای سازمانها به همراه دارد، ولی مسائلی را در بر میگیرد که میتواند مدیریت ارشد را با مشکلاتی روبرو کند تغییرات زیر در نتیجه ورود فناوری اطلاعات و تحولات فضای سازمانها در زنجیره ارزش رخ داده است:

- تغییر تمرکز مدیریت زنجیره ارزش از تولیدات خاص به درخواستهای کاربر و همزمان سازی

فعالیتها

- به دست آوردن دیدگاههای جدید راجع به شبکه گسترده شرکا و روابط میان آنها
- افزایش عدم قطعیت در زنجیره ارزش و به طبع در اجزای کوچکتر زنجیره و سازمانها
- افزایش اطلاعات و دادههای ورودی و خروجی سازمان و پیچیده شدن نگهداری و کنترل این داده ها

1- عضو هیئت علمی پژوهشگاه ارتباطات و فناوری اطلاعات kharrat@itrc.ac.ir
2- دانشجوی کارشناسی ارشد مهندسی فناوری اطلاعات دانشگاه تهرانm.mosharraf@ut.ac.ir
3- استادیار دانشکده مهندسی برق و کامپیوتر دانشگاه تهران ftaghiyar@ut.ac.ir

- افزایش سیستمهای اطلاعاتی سازمانها توزیع شدن این سیستمها در شبکه و تغییر مدیریت آنها از مرکزی به توزیع شده

تبدیل شدن امکان بازیابی اطلاعات و استفاده مجدد از امکانات موجود به عنوان یک ضرورت

- انجام تعاملات سازمانها از طریق شبکههای اینترنتی و به صورت مجازی

- ضروری شدن توانایی پیکر بندی پویا و عملکرد سریع در برابر تغییرات

- نیاز به ایجاد یکپارچه سازی هم شکلی و اتحاد میان سازمانها و سامانه های موجود برای ارائه خدمات با توجه به نکات اشاره شده لازم است راهکارهای گوناگونی برای تحقق فرایندهای سازمانی و روند مدیریت جدید در این حوزه مد نظر قرار بگیرد؛ به طوریکه علاوه بر ایجاد زنجیره ارزش مورد نظر تعاملات لازم در این خصوص را نیز پشتیبانی کند.

ظرفیتها و قابلیتهای تعامل میان سازمانها و سامانه ها حوزه جدیدی را تحت عنوان تعامل پذیری(1) وارد ادبیات سازمانی نموده است. مفهوم تعامل پذیری در بردارنده موارد و نکات متنوعی است که تعریف آن را با پیچیدگیهای خاصی روبرو می.کند از جمع بندی تعاریف معتبر و در عین حال متفاوتی که محققان و دست اندرکاران حوزه تعامل پذیری ارائه کرده اند میتوان تعریف ساده و در عین حال نسبتاً کامل ذیل را برای تعامل پذیری قبول کرد تعامل پذیری عبارت است از توانایی دو سیستم به منظور شناخت و استفاده از کارکردهای یکدیگر از دیدگاه فناوری رایانه تعامل پذیری نشان دهنده توانایی دو سیستم رایانهای ناهمگن برای کارکرد مشترک و همچنین برای دادن دسترسی به منابع یکدیگر به صورت متقابل و دو طرفه است. در زمینه سازمانهای شبکه ای نیز تعامل پذیری اشاره به توانایی تعاملات (تبادل اطلاعات و خدمات میان سیستمهای سازمانی دارد در صورتی که تعاملات میان دو طرف حداقل درسه سطح داده، خدمات و فرایندها و در زمینه کسب و کاری مشخص صورت پذیرد، آنگاه تعامل پذیری حساس و مهم جلوه میکند (David et al.2008).

به طور کلی، تعامل پذیری به معنای همزیستی خودمختاری و محیط متحد است؛ در حالی که یکپارچگی بیشتر اشاره به مفاهیمی نظیر ،هماهنگی وابستگی و متحدالشکل شدن دارد. مقایسه رایانه های به هم پیوسته قوی(2) و پیوسته ضعیف(3) میتواند به درک بهتر تفاوت این دو مفهوم کمک کند. پیوستگی قوی یعنی ارتباط اجزا به گونه ای است که جدا شدن آنها غیر ممکن بوده و به هم وابستگی دارند. این تعریف قابل مقایسه با مفهوم یکپارچگی است. در مقابل، پیوستگی ضعیف بیانگر گونه ای از ارتباط است که اجزا میتوانند در عین اتصال به یکدیگر با حفظ منطق عملیاتی خاص خود با یکدیگر تبادل خدماتی برقرار .نمایند این گونه از ارتباط از درجه به هم پیوستگی قابل قیاس با تعامل پذیری است. یکپارچگی شرط لازم و کافی برای تعامل پذیری است ولی رابطه معکوس آن برقرار نیست.

دیدگاه دیگری در این خصوص که توسط ISO 14258 بیان شده عنوان میکند که دو سیستم در صورتی یکپارچه در نظر گرفته میشوند که قالبی استاندارد و تفصیلی برای تمامی اجزا مؤلفهها وجود داشته

1- Interoperability

2- Tightly Coupled

3- Loosely Coupled

باشد (ISO14258،1999) رویکرد همشکل سازی در تعامل پذیری منوط به وجود ساختاری فراسطحی و مشترک (ارائه ابزاری برای برقراری تعامل معنایی) در تمامی مدلهای مؤلفهها است. از سوی دیگر در محیطی که لازم است مدلها به صورت پویا و مستمر با یکدیگر تطابق یابند، رویکرد تعامل پذیری از ضروریات آن محیط قلمداد می شود لذا اگر ضرورت فوق در محیطی نباشد با توجه به اینکه شرایط اولیه تحقق تعامل پذیری پیچیدگی بالایی دارد تعهد به تعامل پذیری توصیه نمیشود (David et al.2008).

1. تعامل پذیری میتواند در حوزههای مختلف سازمانی مورد توجه قرار گیرد که این حوزه ها عبارتند از: .1 تعامل پذیری داده این مفهوم اشاره به توانایی سازگار کردن مدلهای داده ای متفاوت و زبانهای جستجو دارد تعامل پذیری داده مشتمل بر یافتن و به اشتراک گذاری اطلاعات بدست آمده از پایه های غیریکنواخت که توسط سیستمهای کاربردی و مدیریتی بانکهای اطلاعاتی مختلف استخراج می شوند، است.

2. تعامل پذیری خدمات اشاره به شناسایی و سازگار کردن کارکردهای بسیاری از خدماتی که مجزا طراحی و اجرا شده اند .دارد واژه خدمات تنها به برنامههای کاربردی مبتنی بر رایانه محدود نشده و کارکردهای شرکتها و سازمانهای شبکه ای را نیز شامل میشود.

3. تعامل پذیری فرایندها هدف از تعامل پذیری فرایندها، هماهنگ کردن فرایندهای متعدد سازمانی برای کار کردن با یکدیگر است یک فرایند سلسله مراتب خدمات را با توجه به برخی نیازهای مشخص سازمان تعریف می.کند همچنین بررسی چگونگی اتصال فرایندهای داخلی دو سازمان به منظور به وجود آوردن فرایندی مشترک ضروری است.

4. تعامل پذیری کسب و کار اشاره به کارکرد هماهنگ و یکنواخت در سطح سازمانی و پرهیز از سبکهای مختلف و متفاوت تصمیم گیری روشهای مختلف ،کاری قانونی و رویکردهای تجاری به منظور توسعه کسب و کار مشترک و هماهنگ در داخل سازمان و با سازمانهای دیگردارد(David2008.et al)

بر اساس استاندارد ISO 14258 سه رویکرد اساسی و پایه برای مرتبط کردن موجودیتها(سیستمها)به یکدیگر به منظور برقراری تعامل پذیری وجود دارد (ISO 12581999):

.1 رویکرد ایجاد یکپارچگی(1): در این رویکرد یک قالب مشترک برای تمامی مدلها وجود دارد این قالب مشترک، لزوماً یک استاندارد نیست بلکه شیوه ای است که باید توسط تمامی ذی نفعان در امور ساخت مدلها و سیستمها مورد تأیید قرار گیرد.

2. رویکرد ایجاد یک شکلی(2): رویکردی است مشترک که تنها در فراسطح(3) وجود دارد. فرامدل(4) موجودیتی قابل اجرا نیست؛ بلکه وسیله ای برای تبادل معانی است که تبدیل و تفسیر مدلها راامکان پذیر می کند.

1- Integrated Approach

2- Unified Approach

3- Meta-level

4- Meta Model

3. رویکرد ایجاد اتحاد(1): در این ،رویکرد قالب مشترکی وجود ندارد برای برقراری تعامل پذیری هر یک از ذی نفعان باید خود را با شرایط محیطی تعامل پذیری تطبیق دهد استفاده از این رویکرد این نکته را در بردارد که هیچ یک از ذی نفعان نمیتوانند مدلها ،زبانها و روشهای کاری خودش را به دیگران تحمیل کند این مطلب نشان از آن دارد که باید برای ترسیم مفاهیم در سطح معنایی، یک هستان شناسی مشترک ایجاد گردد (Berre et al. 2004).

هر یک از سه رویکرد فوق امکان تعامل پذیری میان سیستمهای سازمانی را فراهم میکنند اما رویکرد ایجاد اتحاد، نسبت به رویکردهای دیگر از مقبولیت بیشتری برای ایجاد تعامل پذیری برخوردار است.

به طور کلی میتوان گفت که انتخاب رویکرد مناسب بستگی زیادی به نوع نیازها و زمینه مورد نظر دارد. اگر هدف از تعامل پذیری ادغام سازمانها باشد استفاده از رویکرد ایجاد یکپارچگی مناسب است. در چنین موردی تنها نیاز به ایجاد یک قالب مشترک برای طرفین بوده و همه مدلها بر اساس این قالب مشترک میبایست ساخته و تفسیر شوند. اگر نیاز به تعامل پذیری برای ایجاد همکاری بلندمدت باشد، رویکرد ایجاد یک شکلی رویکردی مناسب است در این حالت یک فرا مدل مشترک برای تمامی طرفها تعریف شده و سپس براساس آن امکان همترازی معنایی و ایجاد نگاشت میان مدلهای مختلف فراهم میشود در صورتی که نیاز به تعامل پذیری برای ایجاد همکاری در پروژه های کوتاه مدت باشد میتوان از رویکرد ایجاد تعهد و همکاری استفاده کرد همچون سازمانهای مجازی در این حالت طرفها باید برای کسب توافق به صورت پویا با هم سازگاری لازم را ایجاد کنند.

تحقق راهکارهای ذکر شده نیازمند زیر ساختهای گوناگونی از جمله زیر ساختهای فنی است. در همین راستا سامانه های الکترونیکی و نرم افزارهای تحلیلی – مدیریتی، امکانات بسیاری را برای سازمانها فراهم کرده اند که هر یک وجوهی از نیازهای مربوطه را محقق مینمایند. استفاده از هستان شناسی در رویکرد ایجاد اتحاد به سازماندهی دانش می انجامد و همزمان قالب مناسبی را پدید می آورد که از آن رهگذر چگونگی استفاده از دانش قابل درک میشود. علاوه بر این استفاده از هستان شناسی نه تنها باعث شناسایی عناصر دانش می شود، بلکه به شناسایی و اصلاح ناهمخوانیهای واژه ای میان واحدهای مختلف کمک میکند (Navarretta et al.2006). شباهت و تناظر ویژگیهایی چون پیمانه ای بودن(2)، توزیع شدگی تغییر پذیری ساختار یافتگی و پیچیدگی در سازمانها با همین خصوصیات در عاملهای هوشمند آنها را به گزینه مناسبی برای پشتیبانی از تغییرات و حل مشکلات سازمان تبدیل کرده است.

به دلیل اهمیت اقتصاد دانش محور استفاده از عاملهای هوشمند در سازمانها با معرفی مدلهایی در سازمانهای تجاری شروع شد. در سال 1997 مدلی مبتنی بر عاملها برای نشان دادن اهمیت یکپارچگی در سازمانهای تجاری توسط Chu و همکارانش پیشنهاد شد (1997). اگرچه در فرایندی که توسط آنها مطرح شده بود، روند یک سازمان به صورت جامع و از طراحی تا تولید پوشانده نشده و بیشترین تمرکز بر معماری نرم افزاری عاملها قرار گرفته بود ولی به عنوان یکی از نخستین گامهای برداشته شده

1- Unified Approach

2- Modularity

در این موضوع مورد توجه است در همین سال Shen و همکارانش مدلی برای زیر ساختهای مبتنی بر ،عامل به نام(1) ABMEI ارائه دادند که به عنوان شبکه ای از واسطها بین زیر بخشهای سازمان قرار گرفته و به حل مشکل واگرایی آنها میپرداخت (1997) این سیستم به دلیل مشکلاتی که در مذاکرات عاملها وجود داشت با موفقیت زیادی همراه .نبود در ادامه این روند دیدگاه مشابه دیگری توسط Maturana و همکارانش مطرح و سیستمی با نام Metamorph پیشنهاد شد (1999). Metamorph نیز مانند ABMEIاز عاملها برای برقراری ارتباط استفاده میکرد در سال 2000 این دیدگاه با گسترش دامنه کار از طراحی تا اجرا ادامه پیدا کرده و Metamorph II پا به عرصه نهاد (Shen et al.2000) در سیستم جدید نیز کاستیهایی مانند عدم پشتیبانی از بخشهای کیفیت و تحقیق و توسعه نتوانست انتظارات پیش بینی شده را به خوبی محقق کند به این منظور محققان سامانه ای را برای اعمال کنترل توزیع شده به نام ManAge پیشنهاد کردند (Heikkila et al. عملکرد این سیستم مبتنی بر چهار عامل اصلی عملگر کنترل کننده اجراکننده و ناظر بنا شده بود. مشکل اصلی این مدل نیز قرار گرفتن تمرکز اصلی کار بر ساختار درونی خود عاملها به جای فرایندهای سازمان بود.

زنجیره ارزش یکپارچه مبتنی بر سیستمهای چند عاملی با تمرکز بر تولید، برنامه ریزی، کنترل و حرکت هدف محور توسط Frey و همکارانش ارائه شد (2003). بعد از آن قالب دیگری مبتنی بر سیستمهای چند عاملی برای انجام یکپارچه فرایند برنامه ریزی و زمان بندی تولید در سازمانها مطرح شد(Lim et al. 2004). هدف این ،سیستم ایجاد یکپارچگی به منظور افزایش انعطاف و پویایی سازمان در مقابل رقیبان بود Feng و همکارانش به ارائه مدل دیگری به منظور پشتیبانی از برنامه ریزی، پیش بینی و کنترل در سازمان پرداختند (2005) در این مدل عاملها به پایگاه دانش پایگاه داده سازمان سیستم کنترل و طراحی کامپیوتری دسترسی داشته و از اطلاعات آنها استفاده میکردند.

برای ایجاد یکپارچگی در سازمان قالبی مفهومی به نام MIBIS توسط Kishore و همکارانشارائه شد (2006). در این مدل سیستمهای چند عاملی به عنوان ابزاری برای شبیه سازی یکپارچگی در سیستمهای اطلاعاتی – تجاری مورد استفاده قرار میگرفتند عاملها در این مدل با استفاده از هشت کلمه کلیدی عامل ،نقش هدف تعامل کار ،منابع اطلاعات و دانش ورودیها را درک کرده و با برقراری یکپارچگی در فرایندهای مربوط به این سیستمها زیر ساخت بهینه ای را برای سازمانهای تجاری فراهم میکردند به منظور پوشش فرایندهای طراحی محصولات و خدمات ارزیابی سازمان برنامه ریزی زمان بندی و مدیریت تولید بلادرنگ، عاملها در سامانه ای به نام(2) MMAمطرح شدند.(Mahesh et al،2007MMA)به عنوان کنترل کننده مرکزی با ارسال پیام به دیگر عاملها باعث سازماندهی آنها و حفظ اطلاعات سیستم میشد. در ادامه تحقیقات این دامنه عاملهای چند ،رفتاره، در زنجیره ارزش و با استراتژیهای برنامه ریزی مختلف توسط Forget و همکارانش پیشنهاد شدند (2008). تمرکز این مدل بر عاملهای برنامه ریزی ،تولید شامل ،تحویل ،نگهداری انبار و حمل و نقل قرار گرفته بود.

1- Agent- Based Manufacturing Enterprise Infrastructure

2- Manufacturing Management Agent

در ادامه این مقـاله ابتـدا عاملهای هوشـمند و ساختار اسـتنتاجی آنها را معرفی خواهیم کرد. در بخش سوم به بیان کاربرد اصـلی عاملهای هوشـمند در مـدیریت اطلاعـات سازمانها پرداخته و عملکرد آنها از منظر تعاملات درون سازمانی و برون سازمانی را بررسـی میکنیم در بخش چهار، چارچوبی مبتنی بر عاملهای هوشـمند برای یکپارچگی سازمانها و زنجیره ارزش معرفی شده و از جنبه های مختلفی بررسی میشود. در ادامه نیز ملاحظـات لازم جهت انطبـاق چارچوب پیشـنهادی برای پشتیبانی تعامل پـذیری در سیسـتمهای مـدیریت محتوا و مخازن اطلاعاتی بیان میگردد در بخش پایانی نیز جمع بندی و نتیجه گیری ارائه شده است.

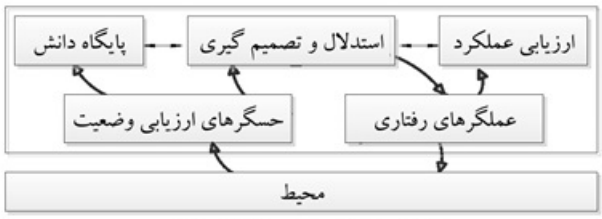
2. عاملهای هوشمند

عامل هوشـمند سیسـتم نرم افزاری خود مختار با تواناییهای اجتماعی واکنش گرا و پیش فعال است تواناییهای تعاملی این عاملها و قـدرت پشتیبـانی آنهـا در برخورد بـا تغییرات پیش بینی نشـده، آنهـا را از دیگر نرم افزارها متمایز کرده است.عاملها به عنوان جزئی از سیسـتمهای پیچیـده و برای نیازمنـدیهای ،توزیعی تعـاملی موازی ،سازی مقیاس ،پـذیری هماهنگی و همکاری بین اجزا اسـتفاده می شونـد (2005) .Cao et al. 2009; Zhang et al) به همین دلیـل در دیـدگاههایی که بر مبنای عاملهای هوشـمند نرم افزاری ارائه میشونـد راه حلها به صورت توزیع شـده در نظر گرفته شده و اسـتانداردهایی برای برقراری تعامل میان آنها تعریف میشود که قدرت مذاکره و ایجاد فهم مشترک را برای آنهـا ممکن میکنـد. شبیه سـازی اجتمـاع انسانهـا در سیسـتمهای چنـد عاملی تعریف سـلسه مراتب برای آنها و تنظیم نظارت در ساختارشان، و استفاده از مکانیزمهای حراجیها در شکستن کارها و تقسیم خرده کارها بین عاملها به طوریکه با افزایش کیفیت باعث بهبود رونـد اجرای کارهـا نیز شود از دیگر ویژگیهای این عاملها هسـتند. به کار گیری عاملها به عنوان زیر ساخت نرم افزاری در تمام کاربردهای مبتنی بر فنـاوری توانسـته است بستری یکپـارچه برای تعامل و هماهنگی همه اجزای آنها فراهم آورده و بهبود چشـمگیری در سامانه های مرتبـط بـا آنهـا و کیفیت ارائه خـدمات ایجـاد کنـد. علاـوه بر این سیسـتمهای چنـد عـاملی به عنوان ابزاری برای پشتیبـانی و بهینه سازی پایگاههای اطلاعاتی سیسـتمها نیز میتواننـد اسـتفاده شونـد. در این دیدگاه عاملها میتوانند به عنوان کاربردهای انحصاری و ناهمگن که در فرایند تصـمیم سازی شرکت می کنند مطرح شوند در ادامه ساختار استنتاجی یک عامل هوشمند نشان داده شده است. همانطور که شکل 1 نشان میدهـد یک عامل میتواند با دریافت دادههای محیط و پردازشـهایی که روی این دادهها انجام میدهد، قوانین اسـتنتاجی خود و در نتیجه عملکردش را در در طول زمان بهبود بخشد(Oztemel et al. 2009)

شکل :1 ساختار استنتاجی عامل هوشمند

استفاده از عاملها به عنوان قدرت اجرایی در تعامل با کاربران یک سیستم الکترونیکی و نیز در ارتباط سیستمهای الکترونیکی با همدیگر از ویژگی تعامل پذیری عاملها و عملکرد اجتماعی آنها ناشی شده است. استفاده از بسترهای مختلف برای پیاده سازی عاملها و ارائه واسط کاربری متناسب با کاربر باعث شده است کاربران سیستمهای الکترونیکی هنگام استفاده با انگیزه و اشتیاق بیشتری عمل کرده و اعتماد بیشتری نیز به امنیت تعاملات داشته باشند.

عکس



شکل ۱: ساختار استنتاجی عامل هوشمند

استفاده از عامل‌ها به عنوان قدرت اجرایی در تعامل با کاربران یک سیستم الکترونیکی و نیز در ارتباط سیستم‌های الکترونیکی با همدیگر، از ویژگی تعامل پذیری عامل‌ها و عملکرد اجتماعی آنها ناشی شده است. استفاده از بسترهای مختلف برای پیاده سازی عامل‌ها و ارائه واسط کاربری متناسب با کاربر، باعث شده است کاربران سیستم‌های الکترونیکی هنگام استفاده، با انگیزه و اشتیاق بیشتری عمل کرده و اعتماد بیشتری نیز به امنیت تعاملات داشته باشند.

۳. کاربرد عامل‌های هوشمند در مدیریت اطلاعات سازمان‌ها

یکپارچگی در یک سازمان و در بعد وسیع‌تر در زنجیره ارزش به همبستگی بین بخش‌های مختلف زنجیره که از دریافت نیازمندی‌ها شروع شده و تا تحویل آنها ادامه می‌یابد اشاره می‌کند. این همبستگی در تمام قسمت‌ها، شامل عناصر پیش پردازش، بخش ورودی، پردازش، خروجی، سنجش و تعامل با محیط وجود دارد. در صورت وجود یکپارچگی در سیستم‌های اطلاعاتی سازمان، خدمات ترتیبی هر بخش طوری زمان بندی می‌شوند که به محض پایان کار بخش پیشین، آغاز شده و بلافاصله پس از تمام شدن(به موقع و یا پیش از موعد) آنها بخش بعدی شروع به کار کند. مشکلی که در این باب وجود دارد،این است که منابع لازم برای این بخش‌ها به طور قطعی مشخص نیست. نمود اصلی این مشکل در درخواست‌های ارائه شده به سازمان (نه در فرایندهای معمول سازمان) است. راهکار ارائه شده برای حل این مشکل،تقسیم بلادرنگ درخواست ارائه شده به چند زیر مسئله و تخمین زدن منابع برای هر کدام است. نکته قابل توجه در اینجا این است که باید بین زیرمسئله‌ها، راه حل‌ها و منابع ارائه شده برای هرکدام هماهنگی وجود داشته باشد(Wang et al. ۲۰۰۸). به همین جهت ضمن انجام کارها به صورت موازی، لازم است بخش‌های مستقل نیز تعاملات نسبی با هم داشته و بین آنها زیرساخت یکپارچه‌ای وجود داشته باشد.

۳-۱. درون سازمانی

با قرار گرفتن زیرساخت سازمان‌ها بر سیستم‌های کامپیوتری و نیز در نظر گرفتن این موضوع که با گسترده شدن ابعاد سازمان و گسترش حجم داده‌ها، این سیستم‌ها نیز باید از قابلیت توزیع شدگی

3. کاربرد عاملهای هوشمند در مدیریت اطلاعات سازمانها

یکپارچگی در یک سازمان و در بعد وسیعتر در زنجیره ارزش به همبستگی بین بخشهای مختلف زنجیره که از دریافت نیازمندیها شروع شده و تا تحویل آنها ادامه می یابد اشاره می کند. این همبستگی در تمام قسمتها شامل عناصر پیش پردازش بخش ورودی پردازش، خروجی، سنجش و تعامل با محیط وجود دارد. در صورت وجود یکپارچگی در سیستمهای اطلاعاتی سازمان، خدمات ترتیبی هر بخش طوری زمان بندی میشوند که به محض پایان کار بخش پیشین آغاز شده و بلافاصله پس از تمام شدن به موقع و یا پیش از موعد آنها بخش بعدی شروع به کار کند مشکلی که در این باب وجود دارد، این است که منابع لازم برای این بخشها به طور قطعی مشخص نیست. نمود

اصـلی این مشـکل در درخواسـتهای ارائه شده به سازمان نه) در فرایندهای معمول (سازمان .است راهکار ارائه شده برای حل این مشکل تقسیم بلادرنگ درخواست ارائه شده به چند زیر مسئله و تخمین زدن منابع برای هر کدام است. نکته قابل توجه در اینجا این است که باید بین زیر مسـئلهها، راه حلها و منابع ارائه شده برای هر کدام هماهنگی وجود داشـته باشد (Wang et al. (2008 به همین جهت ضمن انجام کارها به صورت موازی لازم است بخشـهای مسـتقل نیز تعاملات نسبی با هم داشته و بین آنها زیر ساخت یکپارچه ای وجود داشته باشد.

3-1 درون سازمانی

بـا قرار گرفتن زیر سـاخت سازمانهـا بر سیسـتمهای کامپیوتری و نیز در نظر گرفتن این موضوع که با گسترده شـدن ابعاد سازمان و گسترش حجم داده ها این سیستمها نیز باید از قابلیت توزیع شدگی

برخوردار باشند ضرورت این نکته که نه تنها عناصر یک سامانه باید به طور مستقل به خوبی مدیریت شوند، بلکه روابط بین آنها نیز از اهمیت ویژه ای برخوردار خواهد شد دو چندان میشود Sheory (2006). مقیاس پذیری سازمانها و ایجاد قابلیت توزیع شدگی در زیر ساختهای آن، با پشتیبانی از مسائلی چون همزمانی ناسازگاری و سربار اطلاعاتی ممکن خواهد بود با استفاده از عاملهای هوشمند می توان مدیریت توزیع شده را در چهار گام جمع آوری داده ها مشخص کردن پارامترهای وابستگی، انجام تحلیلها به صورت توزیعی تجمیع و همکاری پیاده سازی کرد (Wang et al.2009).

عاملهای هوشمند میتوانند در سطوح مختلف زیر ساختهای یک سازمان را تشکیل دهند بسته به سطحی که عاملها در آن قرار میگیرند و وظایفی که در آن به عهده دارند قابلیتهای آنها و در نتیجه پیچیدگیهای آنها تفاوتهای اساسی خواهد داشت. معمولا زیر ساختهای فراهم شده به وسیله عاملها هم سطح نبوده و دارای سلسله مراتب معنی داری است (Pawlewski et al، 2009 Wang et al.2009). این سلسله مراتب حداقل سه لایه زیر را در بر خواهد گرفت:

- نظارت
- تحلیل و کاربرد
- داده

بر حسب شرایط و اندازه سازمان پیچیدگیهای امنیتی و تسهیلاتی که در دسترسیها فراهم می کند، لایه های دیگری نیز میتوانند به این ساختار افزوده شده و یا لایه های معرفی شده خود به چندین زیر لایه تقسیم شوند.

معمولا پایین ترین سطح به پایگاههای اطلاعاتی و ساختارهای داده ای سازمان می پردازد. عاملهای مربوط به این سطح ضمن مدیریت دسترسی به ساختارهای اطلاعاتی سازمان به عنوان واسطی برای تبدیل اطلاعات و ایجاد هماهنگی و سازگاری در اطلاعات توزیع شده عمل می کنند (Wang et al. 2006).

لایه تحلیل گر معمولاً در بخش اصلی این ساختار قرار گرفته و میتواند به عنوان یک عنصر نرم افزاری یک پارچه یا مجموعه ای از عناصر توزیع شده عمل کند این لایه دادههای دریافت شده از لایه پایین تر را بازیابی کرده و ضمن انجام پالایش آن دادهها تحلیلهای مرتبط را بر آنها انجام خواهد داد. نتایج تحلیل میتواند به عنوان گزارشی برای به روز کردن سیاستهای برنامه ریزی و فعالیتهای سازمان به بخش مدیریت ارسال شود. این لایه میتواند بنا به اقتضا از تعدادی عامل کاربردی نیز تشکیل شود که بخش فیزیکی کار را انجام دهند (Yang et al.2010)

لایه نظارت از ابزارهایی برای نمایش روند کارها شامل توصیف فرایندهای تجاری، قوانین و منطق آنها تشکیل شده است کاربردهای مناسب برنامه ریزی و زمان بندی برای لایه های پایینتر در این لایه مشخص می شود. علاوه بر این روند کار با دیگر سیستمهای تجاری عملگرهای انسانی، مشتریها و ماشین ها و دیگر کاربردها از وظایف این لایه خواهد بود(Wang et al. 2006).

اگرچه عاملها به عنوان نرم افزارهایی که دارای قابلیتهای اجتماعی هستند، برای برقراری تعامل و ارتباط بین بخشهای مستقل یا توزیع شده معرفی شدند؛ ولی باید این نکته در نظر گرفته شود که استفاده

از عاملها تنها به همین مورد محـدود نمی شود نکته ای که در ارتباط عاملها بایـد در نظر گرفته شود این است که این ارتباطات میتواننـد از طریق شبکه و به صورت ناهمگام نیز انجام شود و همین مورد میتواند استقلال عملکردی آنها را افزایش دهد.

3-2 بین سازمانی

موفقیت یک سازمان در برقراری سـریع ارتباط با دیگر سازمانها به طرز اشتراک گذاری اطلاعات وابستگی مستقیم دارد به این دلیل طراحی و ساخت ابزارهای ارتباطی و نحوه تبادل دادهها اهمیت بسیار زیادی خواهد داشت فناوریهای ارتباطی و اطلاعاتی میتوانند با فراهم آوردن مواردی نظیر مـدیریت شـبکه مـدیریت و نظارت بر اجتماع ،شـرکا پیکر بنـدی سازمان مجازی و کنترل مشارکتی زمان هزینه و کیفیت از تعاملات مجازی سازمانها پشتیبانی کنند.

وب سرویسـها به عنوان راه حلی برای سیسـتمهای کاربردی سازمانی و به دلیل قابلیت پیکر بنـدی بالا و پشتیبانی از عملکرد سازمانها برای رسـیدن به سطوح بالاتری از توزیع شدگی مورد توجه قرار گرفتند. اگر چه این نرم افزارها به عنوان پارادایم مناسب برای کاربردهای با حجم بزرگ ماننـد زنجیره ارزش و معماریهای مبتنی بر سـرویس به عنوان شـیوه ای پیشـرو در معماریهای سازمانی مطرح شـده انـد، با این حال نتوانسـته انـد انتظـارات سازمانها در تعاملات برون سازمانی را محقق کننـد دلایلی چون توصـیف سـخت فراینـدهای سازمانی برای وب سرویسـها کشف زمان بر ساختار معنایی نابالغ تجمیع سـخت و نبود تضمین برای امنیت دادههای سازمانها از جمله دلایل عدم موفقیت وب سرویسها در برقراری ارتباط بین سازمانی محسوب میشوند (Shen et al. (2007).

بـا در نظر گرفتن قابلیتهایی که اجتماع عاملهای هوشـمند در توزیع شـدگی برای سازمانها فراهم کننـد و نیز ویژگیهای واکنش گرایی و پیش فعـالی ،عاملهـا ترکیب آنهـا بـا وب سرویسـها می توانـد نقایص آنها را جبران کنـد عاملهای هوشـمند با تحلیل نیازمنـدیها ساخت پویای سرویس، مـذاکره بـا سازمانهـایی که از سـوی وب سرویسـها کشف شدهانـد و ایجـاد قرار داد در صورت حاصـل شـدن توافق به برقراری ارتباطهای بین سازمانی کمک میکنند.

در فرایند مذاکره که به صورت بین سازمانی انجام میشود علاوه بر اینکه ممکن است دو عامل مذاکره کننده دارای درک مشترکی از مفاهیم نبوده و دچار مشـکل شوند در مذاکرات کاربر انسانی با عاملها نیز ممکن است بحرانهایی رخ دهد به همین دلیل برای استفاده از عاملها در سطوح پایین تعاملات داشـتن زبان مشترک برای برقراری ارتباط میان آنها کفایت میکند؛ ولی در سـطوح بالا وجود هستان شناسی مشترک برای فراهم کردن در ک مشترکی از مفاهیم لازم است.

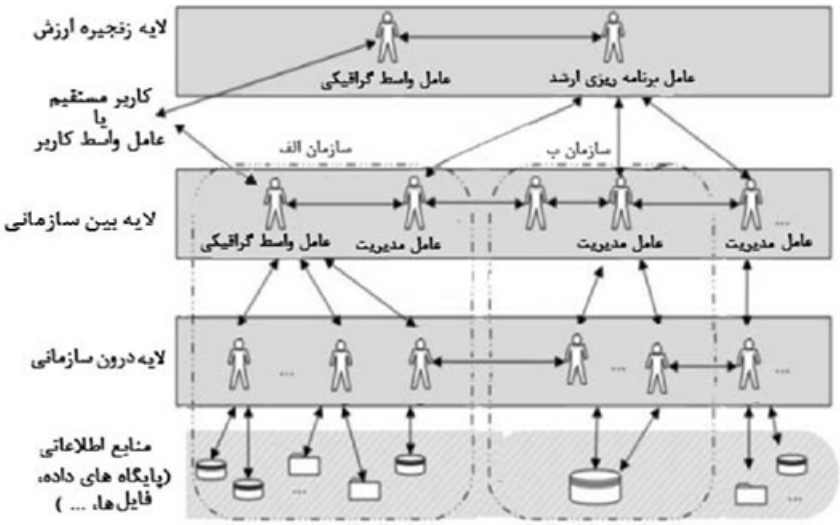
4. پیشنهاد چارچوبی مبتنی بر عامل در یکپارچه سازی مدیریت اطلاعات سازمانها ایجاد یکپارچگی و امکان تعامل پذیری در سازمانها نیازمند فراهم کردن چارچوبی منطقی برای جریان یافتن اطلاعات در میان بخشـهای مختلف و نیز تصمیم گیریهای مشارکتی است به این منظور لازم

است ضـمن ارائه زیر ساخت پیوسـته در تمامی بخشـهای سازمان امنیت اطلاعات و ساختارهای نظارتی نیز در نظر گرفته شود استفاده از شیوه های مطرح در معماری نرم افزار مانند معماریهای چند لایه، می تواند برای این منظور مفید واقع شود. در این معماریها ضـمن فراهم کردن ساختار سلسله مراتبی و اعمال کنترل از بالا به پایین امنیت اطلاعات نیز از پایین به بالا تأمین می شود.

عکس

است ضمن ارائه زیرساخت پیوسته در تمامی بخش‌های سازمان، امنیت اطلاعات و ساختارهای نظارتی نیز در نظر گرفته شـود. استفاده از شـیوه‌های مطرح در معماری نرم افزار مانند معماری‌های چند لایه، می‌تواند برای این منظور مفید واقع شود. در این معماری‌ها ضمن فراهم کردن ساختار سلسله مراتبی و اعمال کنترل از بالا به پایین، امنیت اطلاعات نیز از پایین به بالا تأمین می‌شود.

بـا در نظر گرفتن این موضوع که تعاملات سـازمان‌ها علاوه بر اطلاعـات، لایه فرایند را نیز در بر می‌گیرد،معماری‌های سلسله مراتبی نمی‌تواند جوابگوی نیاز آنها باشد. از طرفی برای برقراری امنیت در تعاملات سـازمان و اعمال دیدگاه‌های نظارتی، این معماری‌های سلسله مراتبی مفید هستند. با افزودن تعاملات در سطوح فرایندی به ساختار سلسله مراتبی، می‌توان بر مشکل مطرح شده غلبه کرد. شکل ۲ چارچوب چهار لایه پیشنهاد شده مبتنی بر عامل‌های هوشمند در سازمان‌ها را نشان می‌دهد. در چارچوب ارائه شده سه لایه پایین در سطح سازمان و لایه چهارم به منظور اعمال مدیریت در سطح زنجیره ارزش خواهد بود.زنجیره ارزش شـبکه‌ای از نهادهایی است که در تولید و توزیع مواد اولیه با ارزش افزوده و تحویل نهایی آن به مشتری همکاری دارند. عملکرد مؤثر زنجیره ارزش را می‌توان در رساندن به موقع و با کیفیت خدمات، اطلاعات یا محصولات تولیدی به مشتری‌ها اندازه گیری نمود. به همین دلیل زنجیره ارزش بایـد بتواند در فرایندهای برنامه ریزی، زمان‌بندی و کنترل با پیشـامدهای غیر قطعی داخلی و خارجی مقابله کند. مدیریت زنجیره ارزش یک رویکرد یکپارچه سازی برای برنامه ریزی و کنترل مواد و اطلاعات است که از تأمین کنندگان تا مشتریان جریان دارد.



شکل۲: چارچوب مبتنی بر عامل در یکپارچه کردن مدیریت اطلاعات سازمان‌ها

بـا در نظر گرفتن این موضوع که تعاملات سازمانها علاوه بر اطلاعات لایه فراینـد را نیز در بر می گیرد، معماریهای سلسـله مراتبی نمیتواند

جوابگوی نیـاز آنهـا باشـد. از طرفی برای برقراری امنیت در تعاملات سازمان و اعمال دیـدگاههای نظارتی این معماریهای سلسـله مراتبی مفید هسـتند با افزودن تعاملات در سـطوح فرایندی به ساختار سلسـله مراتبی میتوان بر مشـکل مطرح شده غلبه کرد. شـکل 2 چارچوب چهار لایه پیشـنهاد شده مبتنی بر عاملهای هوشـمند در سازمانها را نشان میدهد در چارچوب ارائه شده سه لایه پایین در سـطح سازمان و لایه چهارم به منظور اعمال مدیریت در سـطح زنجیره ارزش خواهد بود زنجیره ارزش شبکه ای از نهادهایی است که در تولید و توزیع مواد اولیه بـا ارزش افزوده و تحویـل نهـایی آن به مشتری همکـاری دارنـد عملکرد مؤثر زنجیره ارزش را میتوان در رسانـدن به موقع و با کیفیت ،خـدمات اطلاعات یا محصولات تولیـدی به مشتریها انـدازه گیری نمود به همین دلیل زنجیره ارزش بایـد بتواند در فرایندهای برنامه ریزی زمان بنـدی و کنترل با پیشامـدهای غیر قطعی داخلی و خارجی مقابله .کند مدیریت زنجیره ارزش یک رویکرد یکپارچه سازی برای برنامه ریزی و کنترل مواد و اطلاعات است که از تأمین کنندگان تا مشتریان جریان دارد.

شکل :2 چارچوب مبتنی بر عامل در یکپارچه کردن مدیریت اطلاعات سازمانها

ص: 62

به دلیل اهمیتی که منابع اطلاعاتی و دادههای سازمان دارند این بخش در پایین ترین لایه چارچوب پیشنهادی قرار گرفته و دسترسی به آن به وسیله عاملهای کنترل کننده محدود می شود. بخشهای مختلف سازمان به فراخور نیاز به منابع اطلاعاتی دسترسی داشته و در لایه بعدی قرار می گیرند. بخش مدیریت در بالاترین قسمت قرار گرفته و بر تعاملات تمام بخشهای درونی سازمان نظارت دارد. به همین ترتیب مدیریت زنجیره ارزش بر کل سازمانهایی که در زنجیره قرار میگیرند نظارت خواهد داشت.لایه های مطرح شده در این چارچوب به این شرح هستند:

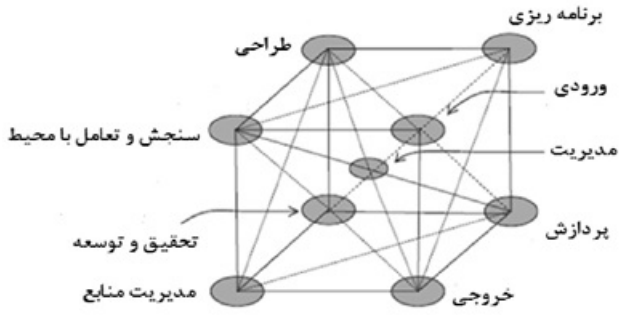
- لایه زنجیره ارزش: وظیفه این لایه ایجاد یکپارچگی در طول زنجیره ارزش و برقراری ارتباط با کاربر نهایی است. با توجه به درخواست کاربر و موقعیت سازمانها این لایه وظایف سطوح پایینتر را مشخص کرده و به آنها اعلام میکند عامل برنامه ریزی ارشد که وظیفه اصلی برنامه ریزی را بر عهده دارد در طول زنجیره ارزش یکتا است. این عامل با جمع آوری دادههای سازمانها در طول زنجیره و اتصال این داده ها در مورد زمان بندی تولید، توزیع کننده ها و تأمین کننده ها تصمیمهای لازم را کند. عامل برنامه ریزی ارشد مسئول برنامه ریزیهای کلان برای عاملهای مدیریت است.

عکس

به دلیل اهمیتی که منابع اطلاعاتی و داده‌های سازمان دارند، این بخش در پایین‌ترین لایه چارچوب پیشـــنهادی قرار گرفته و دسترسی به آن به وســـیله عامل‌های کنترل کننده محدود می‌شود. بخش‌های مختلف سازمان، به فراخور نیاز به منابع اطلاعاتی دسترسی داشته و در لایه بعدی قرار می‌گیرند. بخش مدیریت در بالاترین قسمت قرارگرفته و بر تعاملات تمام بخش‌های درونی سازمان نظارت دارد. به همین ترتیب مدیریت زنجیره ارزش بر کل سازمان‌هایی که در زنجیره قرار می‌گیرند نظارت خواهد داشت.

لایه‌های مطرح شده در این چارچوب به این شرح هستند:

- لایه زنجیره ارزش: وظیفه این لایه ایجاد یکپارچگی در طول زنجیره ارزش و برقراری ارتباط با کاربر نهایی اســـت. با توجه به درخواست کاربر و موقعیت سازمان‌ها، این لایه وظایف سطوح پایین‌تر را مشخص کرده و به آنها اعلام می‌کند. عامل برنامه ریزی ارشد که وظیفه اصلی برنامه ریزی را بر عهده دارد در طول زنجیره ارزش یکتا است. این عامل با جمع آوری داده‌های سازمان‌ها در طول زنجیره و اتصـــال این داده‌ها، در مورد زمان بندی تولید، توزیع کننده‌ها و تأمین کننده‌ها تصمیم‌های لازم را اتخاذ می‌کند. عامل برنامه ریزی ارشد مسئول برنامه ریزی‌های کلان برای عامل‌های مدیریت است.
- لایه مدیریت سازمان: وظیفه این لایه ایجاد یکپارچگی در سازمان، تصمیم گیری‌های مدیریتی، برنامه ریزی و نظارت بر لایه‌های درونی سازمان است. عامل مدیریت که عاملی یکتا در سازمان است؛ بر روابط ســـایر عامل‌ها در درون سازمان نظارت کرده و مسئول اجرای طرح‌های محول شده از سوی عامل برنامه ریزی ارشد است. این عامل می‌تواند تا زمانی که در طرح اصلی و برنامه‌های استراتژیک ســـازمان تناقضی پیش نیامده است، با تدوین برنامه‌های محلی، زمان بندی سایر بخش‌های سازمان را مشخص کند.
- لایه درونی سازمان: این لایه از بخش‌های مختلفی تشکیل شده که هر کدام وظایف مشخصی دارند. بر حسب نوع و پیچیدگی سازمان، تعداد عامل‌های این لایه و روابط آنها متغیر است. در حالت کلی این لایه از هشـــت عامل تشکیل می‌شود.شـــکل ۳ عامل‌های این لایه را که نماینده بخش مرتبط با نامشان هستند را نشان می‌دهد.



شکل۳: عامل‌های لایه درون سازمان و روابط بین آنها

• لایه مـدیریت سازمان: وظیفه این لایه ایجاد یکپارچگی در سازمان تصـمیم گیریهای مدیریتی، برنامه ریزی و نظارت بر لایههای درونی سازمـان است عامـل مـدیریت که عـاملی یکتـا در سازمـان است؛ بر روابط سایر عاملها در درون سازمان نظارت کرده و مسـئول اجرای طرحهـای محول شـده از سوی عامـل برنـامه ریزی ارشـد است. این عامـل میتوانـد تا زمانی که در طرح اصـلی و برنامه های استراتژیک سازمان تناقضی پیش نیامده ،است با تدوین برنامههای محلی زمان بندی سایر بخشهای سازمان را مشخص کند.

• لایه درونی سازمان: این لایه از بخشـهای مختلفی تشـکیل شده که هر کدام وظایف مشخصـی دارند. بر حسب نوع و پیچیدگی سازمان تعداد عاملهای این لایه و روابط آنها متغیر است. در حالت کلی این لایه از هشت عامل تشـکیل میشود. شـکل 3 عاملهای این لایه را که

نماینده بخش مرتبط با نامشان هستند را نشان میدهد.

شکل 3: عاملهای لایه درون سازمان و روابط بین آنها

ص: 63

رأسهای مکعب فوق نشان دهنده عاملهای درونی سازمان هستند. هر کدام از گره های نشان داده شده می توانند شامل زیر لایه ها و عاملهای دیگری بوده و خودشان مدیریت بخش را بر عهده داشته باشند. پیوندهای رسم ،شده ارتباط و وابستگی اطلاعاتی بخشهای مورد نظر را نشان میدهد. این مجموعه از عاملها علاوه بر اینکه خود ساختار دانش مستقلی دارند با دسترسی به پایگاههای اطلاعاتی مجاز ،سازمان به انجام وظیفه میپردازند گره نشان داده شده در قطر مکعب نشان دهنده عامل مدیریتی سازمان است که در لایه دوم ساختار پیشنهادی قرار دارد.

- لایه اطلاعاتی سازمان: این لایه شامل سیستمهای اطلاعاتی سازمان، اعم از پایگاه داده اطلاعات کاربران و نیز اطلاعات و اسناد سازمان خواهد بود عاملهای مختلف با رعایت قوانین دسترسی مجاز به استفاده از این اطلاعات خواهند بود عاملهای مختلفی مسئولیت مدیریت این سیستمهای اطلاعاتی را بر عهده دارند وظیفه این عاملها کنترل دسترسیهای مجاز به این اطلاعات و نیز ایجاد هماهنگی در ورودی و خروجی پایگاههای داده است. در صورتی که پایگاههای داده توزیع شده ،باشند عامل مرتبط با آن وظیفه برقراری یکپارچگی بین اطلاعات آنها را بر عهده خواهد داشت.

در مدل مطرح شده عامل برنامه ریزی ارشد مسئول شکست وظیفه محول شده به زیر وظایف و اعلام آن به عاملهای مدیریت است عاملهای مدیریت مبادرت به تعیین وظایف بخشهای مختلف سازمان در راستای وظیفه انتسابی از سوی عامل برنامه ریزی ارشد نموده و زمان بندیها را به این بخش اعلام می کنند.

برقراری ارتباطهای بین سازمانی به وسیله عاملهای لایه سوم ممکن می.شود. عامل واسط مسئول برقراری ارتباط در این لایه خواهد بود. این عامل نیز با دریافت سیاستهای مذاکره از سوی عامل مدیریت با همتای خود در سازمان دیگر و یا با کاربر انسانی مذاکره میکند به منظور ایجاد درک صحیح از تعاملات انجام شده لازم است عاملهای مذاکره کننده در سازمانهای مختلف هستان شناسی یکسان و ساختار ارتباطی منطبق بر استانداردهای تعریف شده داشته باشد.

همانند لایه سوم برقراری ارتباط توسط بخشهای مختلف یک سازمان با بخشهای داخلی سازمانهای دیگر توسط عاملهای واسط در لایه دوم ممکن می.شود تعیین سیاستهای ارتباطی عامل واسط ، بر عهده عامل مسئول بخش مرتبط با او است نکته قابل توجه این است که قوانین تعیین شده توسط عاملهای بخشهای مختلف در راستای سیاستهای تعیین شده توسط عامل مدیریت است به دلیل نزدیکی این سطح به منابع اطلاعاتی و دسترسی به جزئیات برنامه های سازمان، تعاملات انجام گرفته در این لایه با کنترل و اعمال قوانین امنیتی بیشتری صورت می گیرد. تعاملات در سطح مدیریت ارشد در این چارچوب بر عهده عاملهای دو لایه مختلف قرار داده شده است. با توجه به اینکه دیدگاه زنجیره ارزش نگاهی راهبردی است و ممکن است سازمان به دلیل پویایی و چابکی در سطح راهبرد با تغییراتی در این سطح مواجه شود ملاحظات درون و برون سازمانی به عاملهای مختلف در لایه های متفاوت واگذار شده است. در چارچوب پیشنهادی در این مقاله آن بخش از رویکردها و تصمیمات حوزه مدیریت ارشد که بر اساس راهبردها و مقررات موجود محقق می شود به

عاملهای لایه مـدیریت سازمان سپرده شـده است عاملهای لایه زنجیره ارزش تعاملات بیرونی زنجیره (عمـدتاً جریان بالایی و پایینی) را با توجه مجموعه قوانین و اهـداف دراز مـدت محقق مینماینـد و مرجعی برای یکپـارچگی میان عملکرد عاملها در لایه های پایین تر خواهند بود.

چارچوب مطرح شده ضـمن فراهم آوردن زیر ساخت یکپارچه برای سازمان و زنجیره ارزش، قابلیت برقراری ارتباط با دیگر سازمانها را نیز فراهم می کند. تنها شرط موفقیت تعاملات برای هر دو سازمان وجود هستان شناسی مشترک و استفاده از استانداردهای تعاملی است. در این چارچوب با برقراری یکپارچگی در سازمان علاوه بر ایجاد دسترسـی به اطلاعات برای بخشـهای مختلف، با حفظ سـطوح دسترسی، امنیت اطلاعـات نیز تـأمین می شود. ساختـار لایه ای ضـمن اعمال قوانین مـدیریتی و کنترل بر تعاملات لایه های پایینتر جهت تصـمیم گیریها را به سرعت از بالاترین سطح در زنجیره ارزش به پایین ترین سطح سازمان هدایت میکند.

4-1. مثالی از به کارگیری چارچوب پیشنهادی در سامانه تعامل پذیری مدیریت محتوا

همانطور که در مقـدمه ذکر شـد، امروزه یکی از دغدغه های خدمات محتوایی امکان برقراری ارتباط میان سازمانها و ذینفعانی است که از گونه های مختلف مـدیریت محتوا و مخازن اطلاعاتی بهره میبرند سامانه تعامل پذیری مدیریت محتوا (CMIS)(1) استاندارد بازی است که در یک سـطح انتزاع بالا جهت ایجاد ظرفیت ارتباط و تعامل مناسب میان بنگاههای مختلف محتوایی ایجاد شده است. این استاندارد مورد قبول و پشتیبـانی کنسرسـیوم اسـتانداردهای وب OASIS نیز میباشـد (OASIS Committee 2010). هـدف اصـلی از ارائه این اسـتاندارد ارائه ویژگی اختصاصـی برای سامانه های مـدیریت محتوا یا ارائه یک معماری جامع برای ارتباط انواع مخازن اطلاعات و مانند آنها نیست بلکه هدف ایجاد ظرفیتی است که سامانههای مختلف مدیریت محتوا و مخازن اطلاعات در سـطح وب بتوانند ضمن تعامل از امکانـات و ظرفیتهای یکـدیگر اسـتفاده نماینـد این اسـتاندارد از یک هسـته مـدل داده جهت تعریف هسـته های اطلاعاتی موجود و یک مجموعه خدمات پایه تشکیل شده است.

در اینجا به منظور بررسی قابلیتهای چارچوب پیشنهادی نحوه انطباق آن در بافتار ذخیره سازی و ارائه خدمات محتوایی بررسی شده است. بـا عنـایت به این واقعیت که موجودیتهایی که توسط CMIS مـدیریت می شونـد در قالب انواع اشـیاء بیان میشوند برای انطباق چارچوب پیشنهادی در این مقاله و CMIS مدل اشیاء این استاندارد مدنظر قرار گرفته است. بدین منظور نحوه انطباق به شرح زیر خواهد بود:

مدل سازی اشـیاء سیاستهای(2) مـدیریتی در لایه زنجیره ارزش سـیاست گـذاری و کنترل بر اشـیاء اصـلی توسط عاملهای این لایه انجام میشود. این اشـیاء میتوانند بطور توزیع شده در سازمانهای مختلف قرار داشـته و از طریق Object Identity تعریف شده در سرویسهای CMIS و مبتنی بر دانش سیاستگذاری موجود در عاملهای این ،لایه مورد مدیریت قرار گیرند.

1- Content Management Interoperability System

2- Policy Object

مدل سازی اشیاء ارتباطی(1) در لایه مدیریت سازمان این اشیاء در استاندارد CMIS با هدف ایجاد ارتباط میان سایر اشیاء تعریف شدهاند و سرویسهای مربوط به آنها به امر یکپارچگی و ارتباط میان سازمانها کمک میکنند عاملهای این لایه میتوانند ضمن ایجاد یکپارچگی در سیاستهای ،مدیریتی امکان مدیریت بر سایر عاملها در لایه های دیگر را هم بر عهده گیرند. ارجاع درخواست مستندات و سایر اشیاء دیجیتالی از سازمانهای مختلف براساس دانش این عامل مسیریابی و رهگیری میشود.

مدل سازی اشیاء پوشه(2) در لایه درونی سازمان اشیاء پوشه در استاندارد CMIS با هدف ارائه اقدامات و تصمیمات منطقی مورد نظر در خصوص اشیاء دیجیتالی تعریف شدهاند این تصمیمات منطقی میتوانند به انواع پردازشها برنامه ریزی تصمیم گیری سنجش و رویکردهای دیگر مدیریتی بر اشیاء دیجیتالی اختصاص .یابند همان طور که در شکل نشان داده شده است عاملهای این لایه به خوبی میتوانند پشتیبان ملزومات این اشیاء باشند.

مدل سازی اشیاء مستندات و سرمایه های دیجیتالی(3) در لایه اطلاعاتی سازمان اشیاء دیجیتالی مورد پشتیبانی در این استاندارد با مشخصات و خصیصه های مختلفی شناسایی میشوند. نکته حائز اهمیت امکان وجود این اشیاء در سامانه های مختلف مدیریت محتوا و مخازن گوناگون دیجیتالی است که در سازمانها و بنگاههای مختلف و بطور توزیع شده وجود دارند عاملهای این لایه میتوانند به شکل خود مختار و هوشمند مراحل پیش پردازش جمع آوری، ذخیره سازی و سایر پردازشهای پسین را به سامانه مدیریت یکپارچه محتوا اعمال نمایند به عنوان مثال "Get Repositories Information" سرویسی است که توسط عاملهای این لایه جهت شناسایی اطلاعاتی نظیر نام ارائه کننده سرویس ،محتوایی نام محتوا، نسخه محتوا مکانهای ذخیره سازی محتوا و مانند آن را در اختیار لایه درونی قرار میدهد تا برنامه ریزی لازم در خصوص آن انجام شود.

5.نتیجه گیری

گسترش منابع مختلف توسعه محتوا و خدمات محتوایی تحت وب امروزه در وضعیتی قرار گرفته است که مرز میان اجزای مختلف زنجیره ارائه خدمات کمی غیر شفاف شده است به طوریکه ضمن از بین رفتن فاصله میان مصرف کننده ها توزیع کنندهها تأمین کننده ها و تولیدکنندگان؛ حوزه های مرتبط با این اجزا و به خصوص در سطح عملکرد خدمات تحت وب گسترش یافته و به ازای سطوح دسترسی ذینفعان قابل بهره برداری شده است. حجم زیاد تنوع و ساختار متفاوت اطلاعات سازمانها و اطلاعات موجود در اینترنت باعث شده است که فرایند مدیریت بازیابی و استخراج آنها اهمیت ویژه ای پیدا کند. از طرف دیگر افزایش سیستمهای اطلاعاتی سازمانها و توزیع شدن بخشهای مختلف زنجیره ارزش

1- Relationship Object

2- Folder Object

3- Documents and Digital Assets

باعث پیچیده و پویا شدن روابط شرکا و عدم قطعیت در زنجیره ارزش شده است. اگرچه سیستمهای نرم افزاری با تغییر و تسریع فرایندها به بالابردن بازدهی زنجیره ارزش کمک کرده اند، با این حال پویایی این زنجیره و گردش اطلاعات مختلف پیش بینی نشده باعث ایجاد مشکلاتی در این امر شده است. عامل های هوشمند با ایجاد یکپارچگی در زیر ساختهای سازمانها و زنجیره ارزش امکان انجام بهینه مدیریت برنامه ریزی ،زمان بندی کنترل و مدیریت عدم قطعیتها را فراهم آورده و بازدهی آنها را بالا میبرند با تعریف هستان شناسی مشترک برای عاملهای تعاملی سازمانها انجام ارتباطهای بین سازمانی نیز به وسیله عاملهای هوشمند ممکن شده و با بهبود کیفیت انعطاف و پویایی در تعاملات خارجی ،سازمان بازدهی و تأثیر آن بالارفته و به مزیت ماندگار منجر می شود.

در این مقاله چارچوبی مبتنی بر عاملهای هوشمند برای مدیریت اطلاعات سازمانها پیشنهاد شده است. این چارچوب با در نظر گرفتن معماریهای مطرح نرم افزار ساختارهای سازمانی و چگونگی تعاملات بخشهای مختلف سازمان؛ یکپارچگی و پویایی را در سطح زنجیره ارزش برقرار کرده و تعاملات سازمان با شرکا را نیز محقق می.کند چارچوب پیشنهادی میتواند ضمن حفظ سطوح دسترسی برای بخشهای مختلف سازمان سازگاری هماهنگی و امنیت اطلاعات را نیز تأمین کند. عامل برنامه ریزی ارشد در زنجیره تأمین با اشراف بر سازمانهای لایه ،پایین ضمن دریافت درخواستهای ارائه شده آن را به بخشهای مختلف تقسیم کرده و با کنترل زمانبندی باعث کم شدن زمان تحویل درخواست و افزایش کیفیت آن خواهد شد. ویژگی دیگر چارچوب پیشنهادی، پشتیبانی و ظرفیت سازی برای تعاملات سازمانی است. وجود عاملها در هر لایه و همچنین رویکرد پیشنهادی در این مقاله سبب شده است که تعاملات سازمانی در سطوح مختلف و با حفظ ملاحظات خاص درون هر سازمان و با ایجاد استقلال لازم به انجام برسند این قابلیت سبب میشود که چارچوب پیشنهادی به گسترش زنجیره ارزشی که متشکل از سازمانهای مختلف است بیانجامد.

منابع

Berre, A. et al. 2004. State-of-the art for interoperability architecture approaches, Model driven and dynamic, federated enterprise interoperability architectures and interoperability for non-functional aspects. Information Society Technology.

Cao, L., V. Gorodetsky, and P.A. Mitkas. 2009.Agent mining: The synergy of agents and data mining. Intelligent Systems, 24:64-72.

Chu, B. et al. 1997.Towards intelligent integrated manufacturing planning-execution. International Journal of Advanced Manufacturing Systems, 1:77-83.

David, C., G.Doumeingts, and F.Vernadat. 2008. Architecture for enterprise integration and interoperability: Past, peresent and future.Computers in Industry, 59:647-659.

Feng, S.C., K.A. Stouffer, and K.K. Jurrens.2005. Manufacturing planning and predictive

,process model integration using software agents. Advanced Engineering Informatics
.19:135-142

Forget, P., S. D'Amours, and J.M. Frayret. 2008.Multi-behavior agent model for planning in supply chains: An application to the lumber industry. Robotics and Computer-Integrated .Manufacturing, 24:664-679

Frey, D., T. Stockheim, P.O. Woelk, and R. Zimmermann 2003. Integrated multi-agent based supply chain management.In Twelfth IEEE International Workshops on Enabling .Technologies: Infrastructure for Collaborative Enterprises (WETICE'03), Washington

Heikkila, T., M. Kollingbaum, P.Valckenaers, and G. Bluemink 2001.An agent architecture .for manufacturing control: manage.Computers in Industry, 46:315-331

.ISO 14258.1999. Industrial automation systems-concepts and rules for enterprise models .ISO TC184/SC5/WG1, April 14, 1999

:Kishore, R., H.Zhang, and R.Ramesh.2006. Enterprise integration using the agent paradigm Foundations of multi-agent-based integrative business information systems. Decision .Support Systems, 42:48-78

Lim, M.K., D.Z. Zhang. 2004.An integrated agent-based approach for responsive control of .manufacturing resources. Computers Industrial Engineering, 46:221-232

Mahesh, M., S.K. Ong, A.Nee, J.Fuh, and Y.F. Zhang. 2007. Towards a generic distributed and ,collaborative digital manufacturing. Robotics and Computer Integrated Manufacturing .23:267-275

Maturana, F., W.Shen, and D.H. Norrie.1999. MetaMorph: An adaptive agent-based

,architecture for intelligent manufacturing. International Journal of Production Research
.37:2159-2173

-Navarretta, C., B.S.Pedersen, and D.H. Hansen. 2006. Language technology in knowledge
.organization systems. New Review of Hypermedia and Multimedia, 12:29-49

OASIS Committee. 2010.The CMIS v1.0 OASIS standard specification. Retrieved from
http://docs.oasis-open.org/cmis/CMIS/v1.0/os/cmis-spec-v1.0.pdf, [Accessed 18 Jan
20131

Oztemel, E., E.K.Tekez. 2009.A general framework of a Reference Model for Intelligent
Integrated Manufacturing Systems (REMIMS). Engineering Applications of Artificial
.Intelligence, 22:855-864

.Shehory, O. 2006. The role of agents in enterprise system management: A position paper
ص: 68

,Fifth International Joint Conference on Autonomous Agents and Multi Agent Systems
.Hokkaido

Shen, W., Q. Hao, S.Wang, Y.Li, and H.Ghenniwa.2007.An agent-based service-oriented integration architecture for collaborative intelligent manufacturing. Robotics and .Computer-Integrated Manufacturing, 23:315-325

Shen, W., F.Maturana, and D.H. Norrie.2000. MetaMorph II: An agent-based architecture for ,distributed intelligent design and manufacturing. Journal of Intelligent Manufacturing .11:237-251

Shen, W., D. Xue, and D.H. Norrie. 1997. An agent-based manufacturing enterprise infrastructure for distributed integrated intelligent manufacturing systems. In Third -International Conference on the Practical Application of Intelligent Agents and Multi .agents, London

Wang, M., J. Liu, H.Wang, W.K. Cheung, and X. Xie.2008. On-demand e-supply chain ,integration: A multi-agent constraint-based approach. Expert Systems with Applications .34:2683-2692

Pawlewski, P., P.Golinska, M. Fertsch, J.A.Trujillo, and Z.J. Pasek. 2008. Multiagent approach for supply Chain integration by distributed production planning, scheduling and control system. International Symposium on Distributed Computing and Artificial .(Intelligence (DCAI

Wang, G., J. Zheng, H. Wu, and Y.Tang. 2009. Research of the enterprise application ,integration platform based on multi-agent. Fifth International Joint Conference on INC

.IMS and IDC, Washington

-Wang, S., W. Shen, and Q. Hao. 2006. An agent-based web service workflow model for inter

.enterprise collaboration. Expert Systems with Applications, 31:787-799

Yang, L., X. Gen, and X. Cao. 2010. Application research of enterprise information integration

based on intelligent agent on ciagent platform. Second International Workshop on

.Education Technology and Computer Science, Wuhan

Zhang, C., Z. Zhang, and L. Cao. 2007.Agents and data mining: Mutual enhancement by

integration. 2nd International Conference on Autonomous intelligent systems: Agents

.and Data Mining, Berlin

چکیده

شبکه جهان گستر وب (WWW) شبکه ای از محتویات و ساختارهای پیوندهای تو در توست که دائماً در حال تکامل و تغییر است از این رو، یک آرشیویست هرگز ممکن نیست که از قوام و صحت محتویاتی که تاکنون جمع کرده با چیزی که بعداً نیاز پیدا خواهد کرد اطمینان حاصل کند بنابراین، پرسشهایی در مورد تشخیص و اندازه گیری این محتویات و در نهایت در مورد درک نقص در انسجام به وجود خواهد آمد به این منظور برخی راهبردهای نموداری ارائه شدهاند که ممکن است در سطوح مختلف به کار روند کار با آخرین تغییرات زمانی صحیح، براساس فراداده های استخراج شده از خزشگرها یا از فایلهای WARC. برای کمک به آرشیویست جهت درک ذات این نارساییها این مقاله به بررسی روشهایی برای نمایش رفتار تغییرات و انسجام آرشیوها می پردازد.

تحلیل انسجام و مصورسازی در آرشیو وب

*(1)

مارک اسپانیول(2) | آرتوراس مازیکا(3) | دیمیتار دنوو(4) | گرهارد ویکوم(5)

ترجمه: عبدالله حسینیان(6)

مقدمه

اگر میتونی منو بگیر عنوان فیلمی براساس یک داستان واقعی درباره یک کلاهبردار معروف به نام فرانک ابا گنیل(7) است. شخصی که در نقش یک خلبان ، دکتر و وکیل ظاهر میشود. این فیلم که سختی های دستگیر کردن یک کلاهبردار را در دنیای واقعی شرح میدهد می تواند با شکل مشابهی در آرشیو وب مقایسه شود.

دنیای جهانگستر وب (www)، میلیونها کاربر را قادر میسازد تا محتویات روی وب را تألیف، تغییر، یا حتی حذف کنند. درست مثل تعقیب یک کلاهبردار حفظ و جمع آوری این دادهها نیز کاری جزئی نیست و میتواند شامل موضوعهای کیفیت داده نیز بشود برای مثال یک سیستم مدیریت محتوا (cms) را در نظر بگیرید که وبگاه مؤسسه ای تحقیقاتی را نگهداری می کند. هر گاه دو محقق به طور مشترک

1- Defects in Web Archiving Catch me if you can": Visual Analysis of Coherence

2- Marc Spaniol

3- Arturas Mazeika

4- Dimitar Denev

5- Gerhard Weikum

6- کارشناس کامپیوتر سازمان اسناد و کتابخانه ملی

7- Frank Abagnale

مقاله ای را منتشر ،کنند cms به صورت خودکار مرجعی برای اتصال مقاله ها در صفحه اصلی دو محقق ایجاد میکند در این حین خزش ممکن است یکی از این صفحه ها را قبل از روز آمدسازی و صفحه دیگر را بعد از آن ملاقات کند. در این صورت آرشیو این صفحه ممکن است غیر منسجم پایان پذیرد. از این رو آرشیویست ممکن نیست که از قوام اطلاعاتی که اکنون جمع آوری کرده برای درخواستهای بعدی اطمینان حاصل .کند به این ترتیب خزش سایت باید برای جلوگیری از بارگذاریهای صفحه بی مورد بین درخواستهای HTTP مکث های قابل توجهی داشته باشد.

در نتیجه ثبات یک وبگاه عظیم ممکن است ساعتها یا حتی روزها به طول انجامد. تغییرات در

خلال این دوره زمانی به طور موقت غیر قابل دسترسی است.

1 – انسجام در واژه نامه آکسفورد چنین تعریف میشود:«عمل یا حقیقت به هم چسباندن» یا «نظم اتصال چندین قسمت یا بخش به منظور ایجاد تمامیت آنها با یکدیگر». در نتیجه زمانی نقص انسجام وجود دارد که بعضی از عناصر به شرایط انسجام حمله ور شوند در مورد آرشیو وب، انسجام، دارای بعد زمانی است محتویات بازه زمانی x یا زمان بین X و Y .

چیزی که به عنوان یک نیاز ساده ظاهر می شود به صورت پیچیده توسعه پیدا میکند و در نهایت غیر ممکن شدن عمل آرشیو را به همراه دارد چون نشر دهندگان نمیتوانند از همه وبگاه خود کپی تهیه کنند، که به صورت بخش بخش است و لزوماً همه با هم کار نمیکنند این مسئله ضمانت کیفیت خزش را محدود میکند شکل 1 اشکال انسجام را در آرشیو وب به تصویر میکشد در این مورد نقص انسجام در مواقعی یک صفحه به صفحه دیگر رجوع میکند که آن صفحه قبلاـ در نسخه اخیر از اعتبار افتاده است، اتفاق می افتد. در این مورد سندهای آرشیو شده در سمت چپ در ارتباط با صفحه ورودی غیر منسجم هستند که با چارچوب قرمز نمایش داده شده اند (با زمان رجوع 2007/2/17) . به هر حال ، پیوندها از صفحه ورودی به صفحه های سمت راست که در تاریخ 2007/2/19 آرشیو شده است – منسجم هستند (که توسط قاب بنر نمایش داده شده اند) چون هر دو صفحه از 2007/2/17 معتبر است و تغییری نکرده اند. اما تشخیص چیزی که به عنوان غیر منسجم برای انسان آسان است برای یک ماشین بسیار سخت خواهد بود . یک رایانه ممکن است تنها سطح محدودی از جنبه های زمانی یک صفحه را تفسیر کند.

با وجود این به دست آوردن آخرین تاریخ اصلاح به عنوان یک نقطه زمانی، ما را قادر می سازد در مورد نقص انسجام بین دو نمونه از یک سند تصمیم بگیریم به این منظور چندین تکنیک در سطوح مختلف جزئیات را برای تعیین نقطه زمان ویرایش محتویات معرفی خواهیم کرد.

تحقیقات بر روی تحلیل تصویری از اشکال انسجام درآرشیو وب در درجه تغییراتی که یا در زمان خزش یا در میان رشته ای از خزشهای سایت رخ خواهد داد کمک بسیاری خواهد کرد. بنابراین، دقت و قدرت تغییر خزش قابل سنجش است؛ دقیقا همانند تعقیب یک کلاهبردار و ما قادر نیستیم که از تغییر شکل یک وبسایت جلوگیری کنیم اما قادر هستیم این تغییرات را مشخص و راهبرد خزش را تنظیم تا در آینده تا حد ممکن منسجم باشد برای فهم بهتر از نقص در انسجام و تغییرات در آرشیو وب ما تحلیل دادهها را با استفاده از چهار روش تصویری پیشنهاد میکنیم:

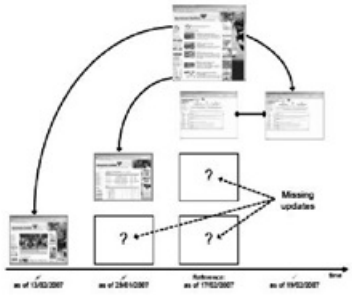
1)تصویر سـازی تغییرات در درخت Spanning خزش با2 ; visone)تصویر سازی Scatterplot در تحلیـل محتوای تغییر؛ 3)منطقه مکانی تصویر سازی سری زمانی تغییرات؛ و 4)تصویر سازی Scatterplot سری زمانی تغییرات.

شکل .1 اشکال انسجام در آرشیو وب

تحلیل تصویری اجازه میدهد که نقص انسـجام در ساختار و محتوا را بررسی کنید سطح تغییرات را به دست آورید و الگوی تغییرات را در میان تک تک صـفحهها در توالی آرشیو وب کشف کنید. UKGOV,MPI از تحلیل تصویری اسـتفاده میکند UKGOV,MPI ،شامل 120 خزش هفتگی از هفـت سـایت دولـتی در انگلسـتان است . MPI شامـل خزش روزانه از سازمانهـای اساسـی برای سایتهای اطلاع رسانی است.

عکس

۱) تصویرسازی تغییرات در درخت Spanning خزش با visone؛ ۲) تصویر سازی Scatterplot در تحلیل محتوای تغییر؛ ۳) منطقه مکانی تصویرسازی سری زمانی تغییرات؛ و ٤) تصویر سازی Scatterplot سری زمانی تغییرات.



شکل ۱. اشکال انسجام در آرشیو وب

تحلیل تصویری اجازه می‌دهد که نقص انسجام در ساختار و محتوا را بررسی کنید، سطح تغییرات را به‌دست آورید، و الگوی تغییرات را در میان تک تک صفحهها در توالی آرشیو وب کشف کنید. UKGOV,MPI از تحلیل تصویری استفاده می‌کند. UKGOV,MPI ،شامل ۱۲۰ خزش هفتگی از هفت سایت دولتی در انگلستان است . MPI شامل خزش روزانه از سازمان‌های اساسی برای سایت‌های اطلاع رسانی است.

این مقاله، براساس روش زیر سازمان‌دهی شده است: ابتدا کارهای مرتبط انجام شده را بازبینی می‌کنیم. در بخش ۳ به بررسی تغییرات می‌پردازیم و مفهوم نقص انسجام را معرفی خواهیم کرد. در بخش ٤ در مورد موضوع جمع‌آوری استخراج و آماده سازی داده‌ها برای تحلیل نقص انسجام صحبت خواهیم کرد. بخش ٤ از تحلیل‌های تصویری را در مورد نقص انسجام معرفی می‌کند در نهایت، در بخش ٦ به نتیجه‌گیری کارهای آینده پرداخته خواهد شد.

۲- فعالیت‌های مرتبط

جامع‌ترین دید در مورد آرشیو وب توسط ماسانه ارائه شده است. این مقاله جنبه‌های مختلف در آرشیو وب را پوشش می‌دهد. همچنین فرآیند دسترسی‌های پشت سر هم را بررسی می‌کند . موضوع نقص در انسجام نیز معرفی می‌شود، اما تنها بعضی از فرآیندهای کاوشی در مورد اینکه چطور عمل آرشیو را توسط خزشگر اندازه‌گیری کنیم و ارتقا دهیم مورد بررسی قرار گرفته است. در این مورد، مفهوم انسجام زمانی به صورت جزئی تر بررسی شده است، اما ذات نقص انسجام در دنیای واقعی را بررسی نکرده است .

این مقاله براساس روش زیر سازماندهی شده است ابتدا کارهای مرتبط انجام شده را بازبینی می کنیم. در بخش 3 به بررسی تغییرات میپردازیم و مفهوم نقص انسجام را معرفی خواهیم کرد. در بخش 4 در مورد موضوع جمع آوری استخراج و آماده سازی داده ها برای تحلیل نقص انسجام صحبت خواهیم کرد. بخش 4 از تحلیلهای تصویری را در مورد نقص انسجام معرفی میکند در نهایت در بخش 6 به نتیجه گیری کارهای آینده پرداخته خواهد شد.

2-فعالیتهای مرتبط

جامع ترین دید در مورد آرشـیو وب توسط ماسانه ارائه شده است. این مقاله جنبه های مختلف در آرشـیو وب را پوشش میدهد. همچنین فرآیند دسترسـیهای پشت سـر هم را بررسـی میکند . موضوع نقص در انسـجام نیز معرفی میشود اما تنها بعضـی از فرآیندهای کاوشی در مورد اینکه چطور عمل آرشـیو را توسط خزشـگر اندازه گیری کنیم و ارتقا دهیم مورد بررسـی قرار گرفته است در این مورد مفهوم انسـجام زمانی به صورت جزئی تر بررسی شده است اما ذات نقص انسجام در دنیای واقعی را بررسی نکرده است .

ص: 73

مک،کاون سیاستهای خزش را که میتواند برای یافتن سایتها استفاده شود ارزیابی کرده است (13,14). نتیجه این تحقیق نشان میدهد که نوع محتویاتی که قرار است دوباره برگردانده شود مهم است و چگونه یک وبگاه دوباره ساخته میشود ،نویسنده ها همچنین تفاوت بین سایت اصلی و سایتی که دوباره ساخته شده است را بررسی میکنند به هر حال آنها راهبردهای ارزیابی های خودکار را برای منابع مبتنی متن پیشنهاد نمی کنند.

در این مقاله نویسندگان به کشف عمر محتویات صفحه ها و منظور از ساخت تاریخچه صفحه ها به صورت پویا در درخواست مشتری می پردازند به طور مشابه Nuneset dl تلاش میکند که سندهای وب را توسط تحلیل همسایه هایش تاریخ گذاری کند (17) . هر دو ،مقاله، مکمل روشهای سنتی بر پایه سرایند HTTP headers HTTP یا تنها روی محتویات فراداده ها هستند به هر حال هیچ یک از دو روش تأثیر تغییرات را روی تمام انسجام سایت اندازه گیری نمیکنند و شامل هیچ تفسیر تصویری از آنها نیستند.

سایر مقاله های مرتبط اغلب روی چیدمان خزشها برای سرعت و تأثیر بیشتر اندیسهای وب تمرکز میکنند. سایبنکو ،(1) تعییرات روی وبگاه را تحلیل میکند و اینکه چگونه باید اندیس گذاری شوند . موضوع اینکه چگونه عمل خزش را مؤثرتر انجام دهیم توسط چو وهمکارانش(5) ارائه شده است.

آنها دلایلی را که نشان میدهند که طراحی یک خزش خوب تا چه اندازه مهم است (برای مثال، ترتیب و توالی نشانی جهت دیده شدن) و نوعی الگوریتم برای به دست آوردن صفحه های مرتبط ارائه می دهند. در مطالعات بعدی چو موهینا و گارسیا توسعه یک خزشگر افزایشی را شرح میدهند(2) آنها به ارتقای مجموعه بی تجربگی ها توسط صفحه های جدید در یک محدوده زمانی کمک می کنند. در همین مطالعات آنها روی سیاستهای تازه کردن مؤثر صفحهها میباشد (3). آنها یک فرآیند پواسون براساس مدل تغییرات منابع داده معرفی کردند در تحقیقی دیگر آنها توالی تغییرات را روی داده های برخط تخمین میزنند به این ،منظور چندین تخمین زننده توالی را با هدف ارتقای خزشگرهای وب و کشهای وب بررسی میکنند در همین مسیر میتوان به تحقیقات اولستون و پندی (18) اشاره کرد که هدفشان تهیه یک جدول زمانبندی Recrawl بر اساس اطلاعات طول عمر به منظور به دست آوردن یک بازده خوب است ایپیروتیس و همکارانش (10) تحلیل بقا را برای بررسی اطلاعات طول عمر به کار میبرند. آنها یک جدول زمانبندی به روزرسانی براساس رگرسیون طول عمر متناسب اختراع کردند. تان و همکارانش (20) ، از نمونه برداری برای بررسی و پیش بینی به روزرسانی صفحه ها استفاده میکنند. آنها ویژگیهای انعکاسی ساختار پیوندها ساختار ، و محتویات صفحه های وب را تعیین میکنند. راهبردهای بارگذاری انطباقی آنها بر اساس بررسی مجموعه صفحه هاست تحقیق دیگری درباره راهبردهای خزشگر توسط ناتورک وویتر (20) ارائه شده است آنها کشف کردند که پهنای بارگذاری صفحه ها در ابتدا بهتر ست اما میانگین کیفیت آنها به مرور زمان کاهش مییابد بنابراین آنها اکیداً جست وجوی breadth first – را به منظور افزایش روی خط بودن برای فراخوانی صفحه ها مهم پیشنهاد کرده اند. تحلیل و فهم

نقص انسجام کاملاً متفاوت و مشکل تر است. ما تغییرات و نقص انسجام را به طور مناسب به تصویر میکشیم و برای تعیین صفحه ها و زیر گرافهای وب خزشهایی که باید در آینده تنظیم شوند، به شما کمک خواهیم کرد.

3-بررسی تغییرات و نقص در انسجام

انسجام یک نوع کاراکتر کیفیتی داده است. به عبارت دیگر در تنظیمات عمومی مجموعه ای از آیتمهای اطلاعاتی هیچ تضادی با محدودیتهای از پیش تعیین شده ندارند در سیستمهای پایگاه داده ای رایج زیر سیستم مدیریت تعاملات(1) مطمئن میشود که کیفیتهای الزامی داده بدون مشکل است. در سیستمهای توزیعی تک تک اجزا باید با یکدیگر همکاری کنند و از الگوریتمهای خاص برای اطمینان از این الزامات استفاده نمایند.

انسجام موضوع پیچیده ای در آرشیو وب است تولید کننده محتویات (ناشر) ممکن است اطلاعاتی را پست کند که با سیستم تضاد دارند برای مثال یک صفحه وب از یک مسابقه فوتبال به عکسی از مسابقه دیگری اشاره کند تهیه کنندگان محتویات وبگاهها تواناییهای محدود شده ای دارند و تمایل به همکاری دارند و منطق در cmsها نیز متفاوت است (صفحه به سرعت میتواند بروز شود در حالی که دیگران ممکن است در تغییرات تأخیر ایجاد کنند محتویات یکی ممکن است به صورت دینامیک ایجاد شود و دیگری خیر).

در این مقاله نقص در انسجام را از دیدگاه زمانی پیگیری میکنیم این مسئله، بخشی از زمان را برای آرشیو تمام وبگاه میگیرد در حالی که اگر آرشیو شامل نسخه صفحه هایی باشد که میتوانند نقطه ای از زمان دیده شوند یا بارگذاری شوند، ما میگوییم که آرشیو بدون نقص در انسجام است. یا اگر بهتر بخواهیم بگوییم اگر یکی از صفحه ها در خلال خزش تغییر کند هیچ ضمانتی وجود ندارد که آن صفحه بتواند در زمان دیگری دیده شود و ما در این مواقع میگوییم که نقص در انسجام وجود دارد.

به منظور اثبات نقص در انسجام دو نمونه از محتویات هم تاریخ و هم محتوا را بررسی می کنیم. به طور رسمی برای ثبت زمان یک صفحه وب از آخرین ویرایش سرایند HTTP استفاده می شود، که متأسفانه غیر قابل اعتماد است (Cf) (11.6) به همین دلیل از روش تاریخ گذاری دیگری بهره می گیریم که تاریخ گذاری معنایی محتوا نام دارد این تکنیک ممکن است یک روش تاریخ گذاری کلی باشد برای نمونه، اولویت تاریخ توسط آخرین ویرایش در پاورقی صفحه وب قرار بگیرد) یا به صورت مجموعه ای از تاریخ تنها روی تک تک آیتمهای صفحه قرار بگیرد (مثل، داستانهای خبری، پستهای بلاگها، کامنتها).

به هر حال استخراج زمان به صورت معنایی مستلزم یک برنامه کاوشی است که در مواردی که به عدم قطعیت در مورد زمان میرسیم به کار برده شود. در نهایت پرهزینه ترین اما صددرصد مطمئن روش مقایسه صفحه ها با نسخه قبلی بارگذاری شده آن است به دلیل پرداخت هزینه ها و دلایل ،مؤثر یک

1- Transaction Management Subsystem

روش چند مرحله ای قوی را دنبال میکنیم:

1.کنترل تاریخ ضمیمه Time Stamp HTTP) اگر ارائه شود و قابل قبول باشد در این مرحله متوقف می شود؛

2.کنترل تاریخ ضمیمه محتویات: اگر تاریخ ارائه شده مطمئن و قابل قبول باشد در این مرحله متوقف می شود؛

3.مقایسه مجموعه صفحهها با مجموعه قبلی که بارگذاری شده است؛

3.حذف تفاوتهای کم اهمیت

تنها مجموعه متنهای محتوا یا متنهای مفید محتوا؛

مقایسه توزیع آن - گرام؛ و

محاسبه مقصد ویرایش از نسخه قبلی.

بر اساس این تکنیکهای تاریخ گذاری قادر هستیم که راهبردهای ارتقای انسجام را گسترش دهیم که به ما اجازه میدهند اطلاعات وابسته به زمان را با چندین خزش یا چندین آرشیو وفق دهیم.

4-استخراج و تهیه داده

این بخش در مورد جمعآوری، استخراج و تهیه داده برای تحلیل نقص انسجام و کشف تغییرات در آرشیو وب صحبت میکند.

الگوهـای تصویری و تحلیـل نقص انسـجام نیازهـای مختلفی را برای ورود داده هـا در نظر می گیرد. در ساده ترین نوع یک تحلیل ممکن است به صـفحه های آرشیو شده یک سایت نیاز پیدا کنیم، در حالی که تحلیلهای وسیع تر ممکن است تغییرات پویا را هم برای صفحه ها محتویات و هم پیوندهای (ساختار) یک سایت بررسی کند.

در این بخش ما یک راهنمایی که یک شـمای بانک اطلاعاتی چگونه بایـد باشـد به شـما میدهیم(بخش 4-1) و اینکه چگونه دادههـا را با Standard SQL وارد یا پاک کنیم.

4-1- شمای بانک اطلاعاتی

به طور خاص شـمای بانک اطلاعاتی شامل صفحهها (Cft_Pages در تصویر10) و پیوندهای (Cft Links در تصویر10) مرتبط با هم است. اطلاعـات صـفحه هـای مرتبـط با یک صـفحه در یک وبگاه شامل نشانی انـدازه حالت کـد و آخرین زمان ویرایش است. به علاوه URL را با URL-id Cf.t_urls) کدگذاری میکنیم به این ترتیب، سـریع تر؛ به طور مؤثر انتخابهایی از صـفحه های مختلف با شـماره سایت خاص و Crawl_id خواهیم داشت و میتوانیم چک کنیم که آیا صـفحه در دو خزش پشت سـر هم تغییر کرده است یا خیر سـپس بایـد به طور مؤثر به پـاک کردن دادههای تکراری بپردازیم. محتویات صـفحههایی که تغییر کرده انـد در صـفت content برای مقایسه با خزش قبلی ذخیره میشود (Cfvs_Page_idSection.)

همچنین اطلاعات پیوندها در جدول t_link ذخیره میشود. هر دو صفت From_url_id و To url_id_ تمامی پیوندهایی را که از یک صفحه به صفحه ای دیگر وجود دارد برای خزش تعیین میکند برای جست و جوی ترتیبی وب ارشیو Parent_Page_id در جدول T_Pageقابل دسترسی است.

دو درخت چند بعدی B روی صفت های Site_id, Crawl_id,Url_id از جدول Tpage- وVisited_TimeStamp,From_Site_id,To_Site_id,Crawl_id,From_url_id,To_url_id) برای T_Link و همچنین یک صفت درخت B برای کلیدهای اصلی وجود دارد. این اصل ساده اما مؤثر در سازماندهی اطلاعات باعث واکشی بسیار سریع دادهها و گزینه ها برای دریافت خزش و شماره (سایت در محاسبه نقص انسجام میشود (بخش5)

2-4- ورود اطلاعات از فایل WARC ورود اطلاعات از ARC و WARC فایلها اصولاً شامل دو وظیفه است:

1) بارگذاری کردن داده ها در بانک اطلاعاتی و (2) حذف اطلاعات تکراری از بانک 8)ARC) و در حالت موفقتر آن 9)WAR) استانداردی کاربردی در آرشیو کردن وبها هستند. آنها برای ذخیره صفحههای آرشیو شده فرادادههای صفحه هایی مثل (نشانی زمان بارگذاری) و کنترل صفحه وب استفاده می.شوند متأسفانه فرمت ARC و WARC اطلاعات مربوط به پیوندهای وبگاهها را پشتیبانی نمی کنند ما این اطلاعات را از DAT فایلها به دست میآوریم یا به طور راحت تر آنها را توسط 15) HERITRIX) در حین آرشیو کردن و استخراج نشانیها ایجاد میکنیم. اگر ARC و WARC قابل دسترسی باشند ساختار پیوند بین صفحه ها میتواند با کمک استخراج نشانی توسط Heritrix از صفحه های HTML آرشیو شده دوباره ایجاد شود.

داده های وب آرشیوها قبل از تحلیل نقص انسجام نیاز به تمیز شدن دارند. رایج ترین مشکل در اینجا بارگذاری چند باره یک صفحه / URL است این ،عمل به چند دلیل اتفاق میافتد بعضی از صفحهها به دلیل سیاستهای بارگذاری وبگاه چند دفعه بارگذاری شده اند مثل Robot.txt)؛ زمانی که بعضی از نیازها در زمان بارگذاری قابل دسترسی نیستند یا به دلیل اینکه آرشیویست برای ارتقای کیفیت پوشش یا کیفیت سایت ممکن یک صفحه را چندین بار بارگذاری کند حتی اگر بخواهیم دلایل بیشتری هم میتوانیم بیاوریم برای مثال صفحهها میتوانند به خاطر فرمت نشانی خود چندین بار بارگذاری شود اگر) وب سرور حروف بزرگ و کوچک در نشانی تشخیص ندهد به خصوص که طراحان صفحه ها تمایل دارند هم از حروف کوچک و هم از حروف بزرگ برای نام گذاری فایلها و مسیرها استفاده کنند) به همین دلایل، قسمت اعظمی از سایت میتواند چندبار خزش شود. حذف اطلاعات تکراری به دلیل تغییر سندها و پیچیده شدن تحلیل تاریخچه این تغییرات برای صفحه ضروری است.

کد SQL برای حذف این داده های تکراری در ضمیمه موجود است این الگوریتم چند زمان بالاتر را در همه گروههای صفحه ها که دارای نشانی یکسان هستند تعیین میکنند (Cf.Line).

حذف دادههای تکراری از فرمت نشانی نیز به طور مشابه انجام میشود همه نشانیها باید به صورت

کوچک و گروه شده توسط نشانیهای یکسان و نشانیهایی با تاریخ جدیدتر گرفته شود. به هر حال، این راه نیز به دلیل مقایسه رشته ها هزینه بر است. به جای آن ما گروهی روی Idهای Urlها (Cf AppendixLine1-AIN Listing) تأسیس میکنیم و کوچکترین مقدار هر گروه را محاسبه می کنیم(Cf.Line10-33).

3-4- جمع آوری دادهها با Heritrix

جمع آوری دادهها توسط Heritrix

در به دست آوردن داده هایی که باید در بانک به طور مستقیم از خزشگر ذخیره شود به جای فرآیند وقت گیر فایلهای WARC استفاده میشود حتی اطلاعاتی از جمله مسیر برای بعضی از صفحه ها به طور مستقیم از خزشگر استخراج میشود اما نیاز به دوباره سازیهای پیچیده فایلهای WARC دارند به علاوه ما نوعی مکانیسم Crawl-Revisit را به منظور کاهش زمان برای تحلیل انسجام توسعه دادیم به صورت تکنیکی انسجام زمانی ما به نسخه ویرایش شده از خزشگر Heritrix بانک اطلاعاتی همراه آن و یک تحلیل و یک محیط تصویری تقسیم میشود. در درون بانک اطلاعاتی فرادادهها و دادههای استخراج شده توسط Heritrix ذخیره میشود و به علاوه، از مکانیسم Crawl-Recrawl نوعی راهبرد مؤثر برای انجام دوباره خزش استفاده می شود و اجازه می دهد که محتویات را بعد از کامل شدن عمل خزش دوباره تست کنیم.

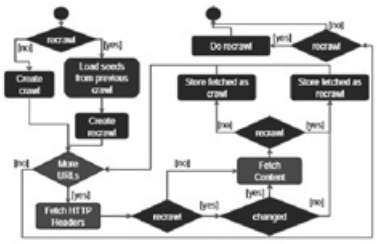
عکس

کوچک و گروه شـــده توسط نشانی‌های یکسان و نشانی‌هایی با تاریخ جدیدتر گرفته شود. به‌هر حال، این راه نیز به‌دلیل مقایســـه رشـــته‌ها هزینه بر اســـت. به‌جای آن، ما گروهی روی Idهـــای Urlها (Cf. AppendixLine۱-۸IN Listing۳) تأســـیس می‌کنیم و کوچک‌ترین مقدار هر گروه را محاســـبه می‌کنیم (Cf.Line۱۰-۳۳).

٤-۳- جمع‌آوری داده‌ها با Heritrix

جمع‌آوری داده‌ها توسط Heritrix در به‌دست آوردن داده‌هایی که باید در بانک به‌طور مستقیم از خزشگر ذخیره شـــود به‌جای فرآیند وقت‌گیر فایل‌هایWARC اســـتفاده می‌شود. حتی اطلاعاتی از جمله مسیر برای بعضی از صفحه‌ها به‌طور مستقیم از خزشگر استخراج می‌شود، اما نیاز به دوباره‌سازی‌های پیچیده فایل‌های WARC دارند. به‌علاوه، ما نوعی مکانیســـم Crawl-Revisit را، به‌منظور کاهش زمان برای تحلیل انســـجام، توسعه دادیم. به‌صورت تکنیکی انســـجام زمانی ما به نسخه ویرایش شده از خزشگر Heritrix، بانک اطلاعاتی همراه آن و یک تحلیل و یک محیط تصویری تقسیم می‌شود. در درون بانک اطلاعاتی فرادادهها و داده‌های اســـتخراج شده توســـط Heritrix ذخیره می‌شودو به‌علاوه، از مکانیسم Crawl-Recrawl نوعـــی راهبرد مؤثر برای انجام دوباره خزش اســـتفاده می‌شـــود و اجازه می‌دهد که محتویات را بعد از کامل شدن عمل خزش دوباره تست کنیم.

به این منظور، Getهای شـــرطی را به‌کار می‌بریم که از محتویات Etagها استفاده می‌کند. در نتیجه، اعتبارسنجی با کاهش پهنای باند و بارگذاری سرور به‌طور همزمان بسیار سریع‌تر خواهد شد. بنابراین، تمام خزش‌ها از Crawl-Recrawlهای مجزا ســـاخته شده‌اند. البته خزش‌های دلخواه می‌توانند به‌عنوان محصولی از جفت Re Crawl–Crawl ترکیب شوند که همان خزش تعریف می‌شود. شکل ۲ فلوچارت قسمت‌های اصلی انسجام زمانی در Heritrix را نشان می‌دهد. عناصر سبز رنگ شامل عناصری هستند که در مقایسه با Heritrix Crawler استاندارد تغییری نکرده‌اند. عناصر آبی رنگ روش خزشگرهای حاضر را که با راهبرد recrawl ما سازگار هستند را نمایش می‌دهد. و در نهایت واحدهای قرمز رنگ یک مرحله اضافه‌ای را که برای شروع Recrawl ها نیاز است نمایش می دهد.



شکل ۲. فلوچارت قسمت‌های اصلی در انسجام زمانی در هر هیتریکس

به این منظور Getهـای شـرطی را به کار میبریم که از محتویات Etagها اسـتفاده میکنـد. در نتیجه، اعتبار سـنجی با کاهش پهنای باند و بارگذاری سـرور به طور همزمان بسـیار سریع تر خواهد شد. بنابراین، تمام خزشها از Crawl–Recrawlهای مجزا ساخته شدهاند البته خزشـهای دلخواه میتواننـد به عنوان محصولی از جفت Re Crawl – Crawl ترکیب شونـد که همـان خزش تعریف می شود. شـکل 2 فلوچارت قسمتهای اصلی انسجام زمانی در Heritrix را نشان میدهد عناصر سبز رنگ شامل عناصری هستند که در مقایسه با Heritrix Crawler اسـتاندارد تغییری نکرده اند عناصـر آبی رنگ روش خزشـگرهای حاضـر را که با راهبرد recrawl ما سازگار هسـتند را نمایش میدهد و در نهایت واحدهای قرمز رنگ یک مرحله اضافه ای را که برای شروع Recrawlها نیاز است نمایش می دهد.

شکل 2 فلوچارت قسمتهای اصلی در انسجام زمانی در هر هیتریکس

ص: 78

تحلیل و محیط تصویری انسـجام زمانی ما به عنوان وسـیله ای برای اندازه گیری کیفیت یک جفت Crawl-Recrawl یا هر جفتی از دو خزش استفاده میشود. به این منظور آمار دادههای هر خزش (از جمله تعداد نقصهای رخ داده و ذخیره شده) از بانک اطلاعاتی مرتبط بعد از اجرای خزش محاسـبه می شود. در این فرآینـد سایت در یک درخت پوشای مشـتق شـده از مسـیرهای خزش درون سایت نمایش داده میشود.

عکس

تحلیل و محیط تصویری انسـجام زمانی ما به‌عنوان وسـیله‌ای برای اندازه‌گیری کیفیت یک جفت Crawl-Recrawl یا هر جفتی از دو خزش اسـتفاده می‌شود. به این منظور آمار داده‌های هر خزش (از جمله تعداد نقص‌های رخ داده و ذخیره شده) از بانک اطلاعاتی مرتبط بعد از اجرای خزش محاسبه می شـود. در این فرآیند، سایت در یک درخت پوشای مشتق شده از مسیرهای خزش درون سایت نمایش داده می‌شود.

تصویرسـازی درخت پوشا[1]، بینشی در مورد موقعیت و ذات تغییرات در محتویات وب در مقایسه با خزش قبلی به ما می‌دهد. به هرحال درخت‌های پوشـا معمولاً بسیار بزرگ هستند و برای بسیاری از ابزارهای تصویرسازی غیرعملی می‌باشند. برای برطرف کردن این مشکل تمرکز روی نقص‌های درخت را فشرده می‌کنیم و تنها قسمت‌های مطرح را به‌صورت تصویر در می‌آوریم (cf.algorithem).

```
input  : CrawlTree tree
begin
    collapseNode (tree.root)
    drawNode (tree.root)
end
```

Algorithm 1: processCrawlTree

الگوریتم ۱. فرآیند درخت خزش

در نخستین گام، محتویات درخت پوشا را تحلیل و آنها را طبق حالتشان تقسیم‌بندی می‌کنیم: سبز اگر بدون تغییر مانده باشند، زرد در مواردی که فقط عدم انسجام از نوع متنی باشد، قرمز در عدم انسجام از نوع ساختار (پیوندها) و در نهایت سیاه برای همزمانی که محتویات در خزش بعدی فراموش شده یا از بین رفته باشـد استفاده می‌شود. در نهایت، راهبرد درهم کردن[2] را به‌کار می‌بریم (cf.Algorithem۳). گره‌های فشرده شده رنگ شده را به‌عنوان گره‌های بررسی شده می‌کشیم. به‌علاوه برای هر زیر درختی که فشرده شده است یک گره پایه به اندازه تعداد گره‌هایی که به این درخت وصل بوده است، رسم می‌کنیم.

```
input  : Node node
begin
    node.collapsing=true
    if hasLinkChange (node) then
        node.color=red
        node.collapsing=false
    else if hasContentChange (node) then
        node.color=yellow
        node.collapsing=false
    else node.color=green
    forall children of node do
        collapseNode (child)
        if child.collapsing=false then node.collapsing=false
        else node.collapsingSize=child.collapsingSize+1
    end
end
```

Algorithm 2: collapseNode

الگوریتم ۲. کنگره درهمکرد

1. Spanning tree
2. Collapsing

تصویر سازی درخت پوشا(1)، بینشی در مورد موقعیت و ذات تغییرات در محتویات وب در مقایسه با خزش قبلی به ما میدهد به هر حال درختهای پوشا معمولا بسیار بزرگ هستند و برای بسیاری از ابزارهای تصویر سازی غیر عملی میباشند برای برطرف کردن این مشکل تمرکز روی نقصهای درخت را فشرده میکنیم و تنها قسمتهای مطرح را به صورت تصویر در می آوریم الگوریتم 1. فرآیند درخت خزش

در نخستین گام محتویات درخت پوشا را تحلیل و آنها را طبق حالتشان تقسیم بندی میکنیم: سبز اگر بدون تغییر مانده باشند زرد در مواردی که فقط عدم انسجام از نوع متنی ،باشد، قرمز در عدم انسجام از نوع ساختار (پیوندها) و در نهایت سیاه برای همزمانی که محتویات در خزش بعدی فراموش شده یا از بین رفته باشد استفاده میشود در نهایت راهبرد درهم کردن(2) را به کار میبریم (cf.Algorithm).

گرههای فشرده شده رنگ شده را به عنوان گره های بررسی شده میکشیم به علاوه برای هر زیر درختی که فشرده شده است یک گره پایه به اندازه تعداد گرههایی که به این درخت وصل بوده است رسم میکنیم.

الگوریتم 2. کنگره در همکرد

1- snanningtree

2- collapsing

برای یک نمایش گرافیکی قبلاً درخت محاسبه شده و در یک فایل graphML ذخیره شده است

(cf.Listing (1) فایل graphML بر پایه استانداردهای XML و یک درخت برای گرافهاست این ،فایل برای شـرح تمام محاسبات قبلی سازگار است و در بسیاری از نرم افزارهای مرتبط با گرافها به کارمی رود.

الگوریتم 3. گره خزش

عکس

برای یک نمایش گرافیکی، قبلاً درخت محاسبه شده و در یک فایل graphML ذخیره شده است (cf.Listing ۱) . فایل graphML بر پایه استانداردهای XML و یک درخت برای گراف‌هاست. این فایل، برای شرح تمام محاسبات قبلی سازگار است و در بسیاری از نرم‌افزارهای مرتبط با گراف‌ها به کار می‌رود.

```
<?xml version="1.0" encoding="UTF-8" standalone="
    no"?>
<graphml xmlns="http://graphml.graphdrawing.org/
    xmlns/graphml" xmlns:xsi="http://www.w3.org
    /2001/XMLSchema-instance" xmlns:y="http://www.
    yworks.com/xml/graphml" xsi:schemaLocation="
    http://graphml.graphdrawing.org/xmlns/graphml
    http://www.yworks.com/xml/schema/graphml/1.0/
    ygraphml.xsd">
 ...
 <graph edgedefault="directed" id="G229">
  <node id="http://www.mpi-inf.mpg.de/index.html">
   <data key="d0">
    <y:ShapeNode>
     <y:Geometry width="10.003" height="10.003"/>
     <y:Fill color="#00FF00" transparent="false"/>
     <y:Shape type="ellipse"/>
    </y:ShapeNode>
   </data>
   <data key="d1">http://www.mpi-inf.mpg.de/index.
       html O&c</data>
  </node>
  ...
  <edge source="http://www.mpi-inf.mpg.de/index.
      html" target="dns:www.mpi-inf.mpg.de"/>
  ...
 </graph>
</graphml>
```

Listing 1: Coherence defect graphML-file (excerpt)

الگوریتم ۳. گره خزش

۵ ـ تحلیل نقص انسجام

تحلیل نقص انسجام، کیفیت یک خزش یا Crawl-recrawl، بین دو خزش یا یک سری از خزش‌ها را اندازه‌گیری می‌کند. به این منظور، روشی را برای تولید آمارهای صحیح و تصویری توسعه داده‌ایم. برای مثال تعداد نقص‌های رخ داده و ذخیره شده توسط انواع کاستی‌ها.

۵-۱- تحلیل تغییر محتویات و ساختار با visone

همان‌طور که در بخش ٤-۳ شرح داده شد، گسترش Heritrix به ما اجازه می‌دهد تا فرآیند خزش را با داده‌های آماری دنبال کنیم و این داده‌ها را برای grapgML ارسال نماییم. با به‌کار بردن graphML نرم‌افزارهای مرتبط قادر خواهند بود درخت پوشا و ظاهر نقص در انسجام را نمایش دهند. این تصویرسازی‌ها به‌عنوان وسیله‌ای اضافی برای آمارهای خودکار جهت کشف مشکلی که در حین ثبت رخ می‌دهد در نظر گرفته می‌شود. قسمت اصلی این برنامه، تحلیل با کیفیت از خزش یک وبگاه است.

شکل ۳ تصویر ساده‌ای از یک خزش از نشانی mpi-inf.mpg.de با نرم‌افزار ویژن را به تصویر کشیده است. بسته به اندازه گره‌ها، شکل و رنگ آنها کاربر یک دید کلی از فرآیندهای موفق و شکست خورده این جست‌وجو به‌دست خواهد آورد. به‌طور خاص اندازه یک گره مبنایی برای تعداد محتویات منسجم سایت (هر چه بزرگ‌تر انسجام در آن بخش بیشتر) در آن زیر درخت است. در همین شکل، رنگ یک گره حالت انسجام آن‌را نمایش می‌دهد. جدی‌ترین نقص از دست دادن محتویات است که به رنگ مشکی نمایش داده شده . در نهایت شکل نت‌ها جنس نقص‌ها را نمایش می‌دهد مثلا دایره

5- تحلیل نقص انسجام

تحلیل نقص انسجام کیفیت یک خزش یا Crawl-recrawl بین دو خزش یا یک سری از خزشها را اندازه گیری میکند به این منظور روشی را برای تولید آمارهای صحیح و تصویری توسعه داده ایم. برای مثال تعداد نقصهای رخ داده و ذخیره شده توسط انواع کاستیها.

-15 تحلیل تغییر محتویات و ساختار با visone همان طور که در بخش 4-3 شرح داده شد گسترش Heritrix به ما اجازه میدهد تا فرآیند خزش را با داده های آماری دنبال کنیم و این دادهها را برای grapgML ارسال نماییم. با به کار بردن graphML نرم افزارهای

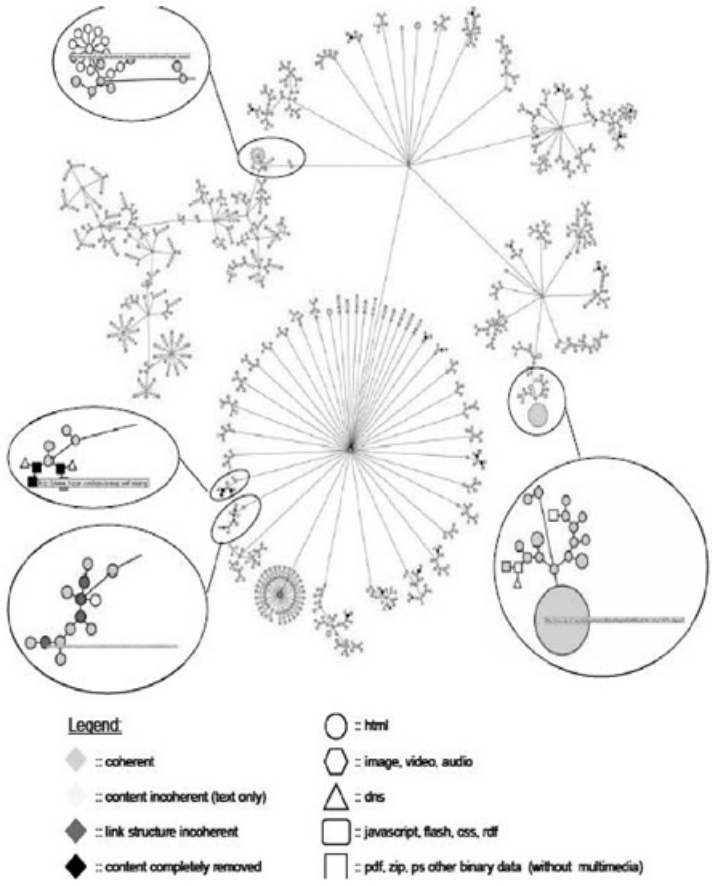
مرتبط قادر خواهند بود درخت پوشا و ظاهر نقص در انسـجام را نمایش دهند. این تصویر سازیها به عنوان وسیله ای اضافی برای آمارهای خودکار جهت کشف مشـکلی که در حین ثبت رخ میدهد در نظر گرفته می.شود. قسـمت اصلی این برنامه تحلیل با کیفیت از خزش یک وبگاه است. شکل 3 تصویر ساده ای از یک خزش از نشانی mpi-inf.mpg.de با نرم افزار ویژن را به تصویر کشیده است. بسته به اندازه گره ها شـکل و رنگ آنها کاربر یک دیـد کلی از فرآینـدهای موفق و شـکست خورده این جست و جو به دست خواهد آورد به طور خاص انـدازه یـک گره مبنایی برای تعـداد محتویات منسـجم سایت هر چه بزرگتر انسـجام در آن بخش بیشتر در آن زیر درخت است. در همین شکل، رنگ یک گره حالت انسـجام آن را نمایش میدهـد جـدی ترین نقص از دست دادن محتویات است که به رنگ مشـکی نمایش داده شده. در نهایت شکل نتها جنس نقصها را نمایش میدهد مثلا دایره

(HTML content)چند وجهی محتویات چندرسانه ای چار گوشه با گوشه های گرد فلش و مشابه آن مربع PDF و دیگر باینریها.

شکل 3

عکس

(HTML content) چند وجهی (محتویات چندرسانه‌ای) چار گوشه با گوشه‌های گرد فلش و مشابه آن، مربع PDF و دیگر باینری‌ها.



شکل ۳

یک روش برای تحلیل یک جفت recrawl-Crawl تحلیل آن براســاس زمان است. ایده بینایی این روش، به‌منظور دنبال کردن نقص در انســجام در میان چند خزش و برای تعیین محتویاتی اســت که کمتر تغییر می‌کنند و احتمالاً منســجم‌تر هستند. شکل ٤ تصویری از نقص انسجام شش خزش پشت ســر هم روی نشانی dmoz.org/news است. هر جفت خزش‌های یک نقص انسجام درست شبیه مثال قبلی انجام می‌شــود فقط برخلاف مورد قبل، اکنون ما خزش‌ها را به جای مقایسه با recrawl-Crawl با

یک روش برای تحلیل یک جفت recrawl-Crawl تحلیل آن براساس زمان است. ایـده بینایی این روش، به منظور دنبال کردن نقص در

انسـجام در میان چند خزش و برای تعیین محتویاتی اس-----ت که کمتر تغییر میکنند و احتمالاً منسجم تر هستند شکل 4 تصویری از نقص انسـجام شش خزش پشت سر هم روی نشانی dmoz.org/news .است هر جفت خزشـهای یک نقص انسجام درست شبیه مثال قبلی انجام میشود فقط برخلاف مورد قبل اکنون ما خزشها را به جای مقایسه با recrawl-Crawl با

ص: 81

خودشان مقایسه می.کنند در انتقالهای هر دو تا از این جفتها همه گره ها مخفی میشوند، آنهایی که در تحلیل دچار نقص انسجام هستند ناپدید میشوند و در مقابل آن محتویاتی نمایش داده می شود که کاراکترهای منسجم یکسانی بین دو خزش حفظ کرده اند و گره هایی که جدید ظاهر میشوند اطراف آنها قرار می گیرند نکته جالبی که در این مثال دیده میشود این است که یک هسته محکم از یک زیر درخت بزرگ منسجم و محتویات غیر منسجم آن در اینجا وجود دارد.

شکل 4

5-2- تحلیل تغییرات محتوا

عکس

خودشـــان مقایسه می‌کنند. در انتقال‌های هر دوتا از این جفت‌ها همه گره‌ها مخفی می‌شوند، آنهایی که در تحلیل دچار نقص انســـجام هستند ناپدید می‌شوند و در مقابل آن محتویاتی نمایش داده می‌شود که کاراکترهای منسجم یکسانی بین دو خزش حفظ کرده‌اند و گره‌هایی که جدید ظاهر می‌شوند اطراف آنها قرار می‌گیرند نکته جالبی که در این مثال دیده می‌شود این است که یک هسته محکم از یک زیر درخت بزرگ منسجم و محتویات غیرمنسجم آن در اینجا وجود دارد.

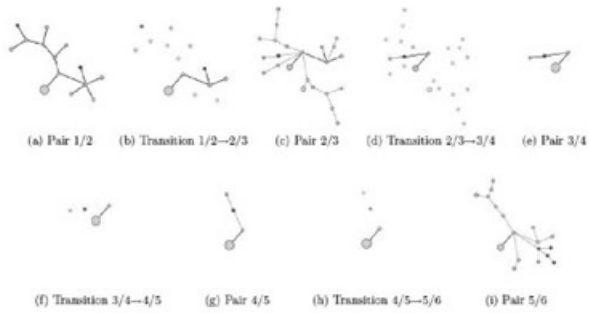


Figure 4: Tracing of coherence defects in crawl-recrawl pairs of the dmoz.org/news subdomain over time

شکل ٤.

۵-۲- تحلیل تغییرات محتوا

نمودارهای پراکندگی دوبعدی و ســـه بعدی می‌توانند برای تصویرسازی مکان و تحلیل تغییرات محتوا به‌کار روند. شکل ۵ و ٦ یک نمودار پراکندگی را برای سایت‌های sabre و royal-navy از بانک اطلاعاتی UKGOVنمایش می‌دهد، در اینجا اندازه و نشـــانی هر صفحه به روی محورهای x,y,z از ســـه بعد یک مکعب قرار گرفته‌اند. در همین حال، رنگ‌ها تغییراتی را که صورت گرفته‌اند نمایش می‌دهند (صفحه‌های جدید به رنگ آبی، صفحه‌هایی که تغییر کرده‌اند قرمز و صفاتی که تغییر نکرده‌اند به رنگ مشـــکی در آمده‌اند). آرشیویســـت باید الگویی از صفحه‌هایی که اضافه شـــده و تغییر کرده‌اند را بیابد. برای مثال از شکل ۵(a) یک نفر می‌تواند ببیند که چند تغییر در فایل‌های HTML در خروجی و زیرشاخه‌های متنی (.cf نطقه‌ها قرمز در شکل) و یک زیر مسیر جدید از فایل‌ها در آرشیو وب اضافه شده است (نقاط آبی در بالای تصویر). تغییرات صفحه‌ها (چهار نقطه قرمز) وابســـتگی‌های بین صفحه‌ها را نمایش می‌دهد. اگر صفحه‌ای در مسیر خروجی تغییر کند، صفحه منطبق با آن در آن مسیر نیز تغییر خواهد کرد. صفحه‌هایی که تازه اضافه شده‌اند نشان می‌دهند که ضوابط ساختاری سایت متحمل تغییرات می‌شوند. اگرچه محتویات سایت خیلی تغییر نکرده باشند.

همچنین، الگوهای عکس‌های اضافه شده برای ســـایت royal-navy در شکل ٦(a)-٦(b) مشابه اســـت. در حالی که، الگوی صفحه‌هاHTML اضافه شـــده و تغییر یافته اندکی متفاوت است که تغییر

نمودارهای پراکنـدگی دوبعدی و سه بعدی میتوانند برای تصویرسازی مکان و تحلیل تغییرات محتوا به کار روند شـکل 5 و 6 یک نمودار پراکندگی را برای سایتهای sabre و royal navy از بانک اطلاعاتی UKGOVنمایش میدهد در اینجا اندازه و نشانی هر صـفحه به روی محورهای x,y,z از سه بعـد یک مکعب قرار گرفته انـد. در همین حال رنگها تغییراتی را که صورت گرفته اند نمایش میدهند صـفحه های جدید به رنگ ،آبی صـفحه هایی که تغییر کرده اند قرمز و صفاتی که تغییر نکرده اند به رنگ مشکی در آمده اند آرشیویست باید الگویی از صفحه هایی که اضافه شـده و تغییر کرده اند را بیابد. برای مثال از شـکل 5(a) یک نفر میتوانـد ببیند که چند تغییر در فایلهای HTML در خروجی و زیرشـاخه های متنی .cf) نطقه ها قرمز در شـکل و یک زیر مسـیر جدید از فایلها در آرشـیو وب اضافه شده است نقاط آبی در

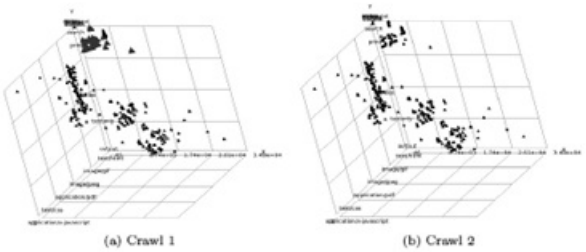
بالای تصویر) تغییرات صـفحه ها (چهار نقطه (قرمز) وابسـتگیهای بین صفحه ها را نمایش میدهد. اگر صفحه ای در مسیر خروجی تغییر کند صـفحه منطبق با آن در آن مسیر نیز تغییر خواهد کرد. صفحه – هایی که تازه اضافه شده اند نشان میدهند که ضوابط ساختاری سایت متحمل تغییرات می شوند. اگرچه محتویات سایت خیلی تغییر نکرده باشند.

همچنین الگوهـای عکسـهای اضافه شـده برای سایت royal-navy در شـکل 6(b)6-(a) مشـابه است در حـالی که الگوی صـفحه ها HTML اضافه شده و تغییر یافته اندکی متفاوت است که تغییر

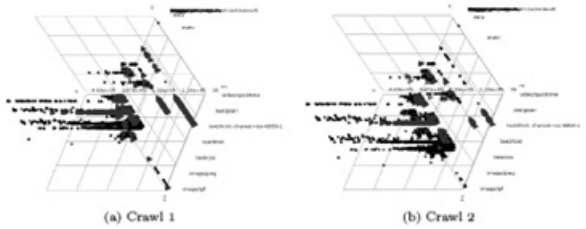
ساختار وبگاه را نشان میدهد.

عکس

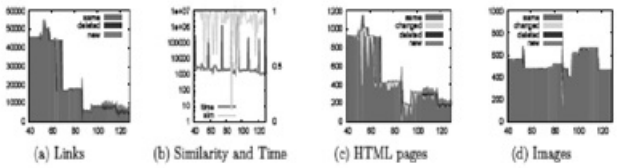
ساختار وبگاه را نشان می‌دهد.



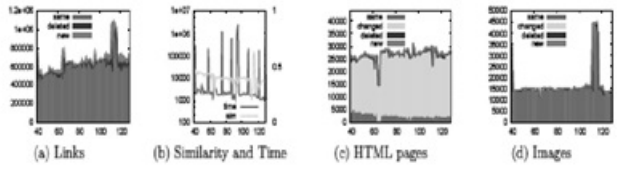
شکل ۵. تحلیل توالی خزش www.sabre.m.d.uksits



شکل ٦. تحلیل توالی خزش در www.royal-navy.mod.uk.site



شکل ۷. تحلیل توالی خزش در www.sabre.mod.us.site



شکل ۸. تحلیل توالی خزش در www.royal-navy.mod.uk.site

شکل .5 تحلیل توالی خزش www.sabre.m.duksits

شکل .6 تحلیل توالی خزش در www.royal-navy.mod.uk.site

شکل7 . تحلیل توالی خزش در www.sabre.mod.us.site

شکل8 . تحلیل توالی خزش در www.royal-navy.mod.uk.site

ص: 83

5-3-تحلیل زمانی با نقشه های محیطی

مجموعه نقشه های محیطی (شکلهای 7 و 8)میتواند برای گرفتن یک دید کلی از درصد تغییرات در خزش آرشیو وبگاهها همان طور که Crawl_id افزایش مییابد استفاده میشود محور X ها شماره خزش و محور لها تعداد صفحه هایی که تغییر کرده اند/ نکرده اند/ اضافه شدهاند حذف شده اند یا درصد زمان بارگذاری را نمایش میدهد. اصولا شکلهای مجزایی از ساختار پیوندها کشیده میشود (گراف ساختار سایت شکل (a)8-7(a) و صفحه های HTML شکل 7(a)7(c)) عکسهای شکل (d) و (d) تغییرات محتوا و به موازات آن کلیه اشکال نیز برای زمان باگذاری مشابه هستند (شکل8(b)7-(b)) بایگانی کننده باید الگوهایی را پیدا کند که باعث تغییرات قابل توجه در نقطه هایی از زمان میشود. برای مثال، یک شخص میتواند که تغییرات قابل توجهی در وبگاه در Crawl52 و 93 ملاحظه کند. زمان خزش پیشنهاد میدهد که در Crawl52 وبگاه متحمل تغییرات قابل ملاحظه ای میشود. به هر حال در Crawl93 کیفیت آرشیو کاهش مییابد و عمل آرشیو لازم به رعابت مسائل خاص می شود.

محاسبه گراف محیطی میتواند در SQL هم بیان شود و توسط بهینه ساز query بهینه گردد (4) cf.Listing) در نتیجه همه این اعمال پیچیدگی به دست آمده (nlogn) یا کمی بهتر خواهد بود. الگوریتم چندین سایت را از Crawl xxx و خزش قبلی آن 10 (Lines. با استفاده از یک outer join برای اتصال خزشها انتخاب می.کند فایلهای که در یک یا دیگر خزشها هستند اما در هر دو وجود ندارند صفحههای جدید یا حذف شده میباشند در حالی که tuples که در نتیجه باقیمانده اند یا تغییر کرده اند و یا بدون تغییر ماندهاند (15-10 tuples of Lines) اضافه شده حذف شده، تغییر یا تغییر نکرده همگی گروه میشوند (cf.Lines1-110)

4-5- الگوهای تغییرات در صفحه های یک سایت

در شکل 9 محور Yها صفحه ها و محور X ها شمار خزش را نمایش میدهد و نقطه تقاطع آنها نشان میدهد که اگر تغییری در صفحه وب در آن خزش وجود داشت با خزش قبلی مقایسه شود. صفحه ها بر روی محور Yها براساس رفتارهای تغییر آنها که مشابه اند مرتب شده اند. شکل به آرشیویست وب برای پیدا کردن و تحلیل صفحههایی که تغییرات مشابهی دارند کمک میکند الگوها و نقص در انسجام را کشف کند ،تصویر به طور واضح صفحههای وبگاه را در بلوکهای مجزایی جدا میکند (مستطیلها در شکل و الگوهای متفاوتی از تغییرات و نقص در انسجام را تعیین می کنند.

شکل .9 الگوی تغییرات صفحه های یک سایت

فهرست 2

عکس

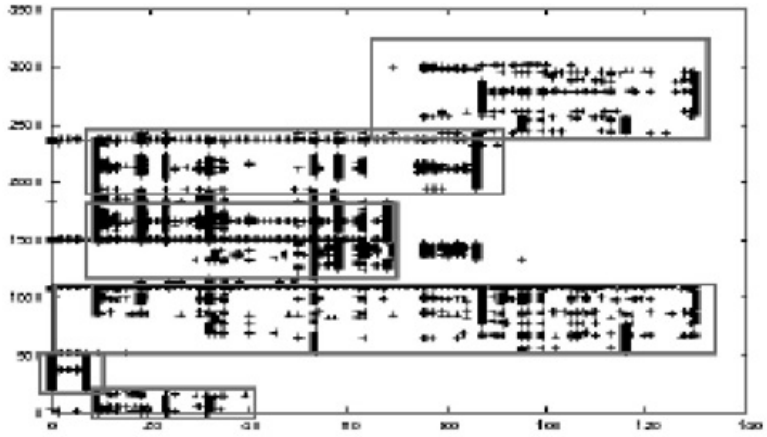


Figure 9: Scatterplot of lines for www.sabre.mod.

شکل ۹. الگوی تغییرات صفحه های یک سایت

```
1  create table t_pages_dd as
2  select t_pages.* from (
3    select crawl_id, url_id, max(visited_timestamp)
4           as latest_timestamp
5    from t_pages
6    group by crawl_id, url_id) as x, t_pages
7  where t_pages.crawl_id = x.crawl_id and
8        t_pages.url_id = x.url_id and
9        t_pages.visited_timestamp
10            = x.latest_timestamp
11
12 create table t_links_dd as
13 select t_links.*
14 from t_pages_dd, t_links
15 where t_pages_dd.url_id = t_links.from_url_id and
16       t_pages_dd.visited_timestamp
17         = t_links.visited_timestamp
```

فهرست ۲

فهرست3 . SQL به کار رفته برای پاکسازی URLهای پایین تر

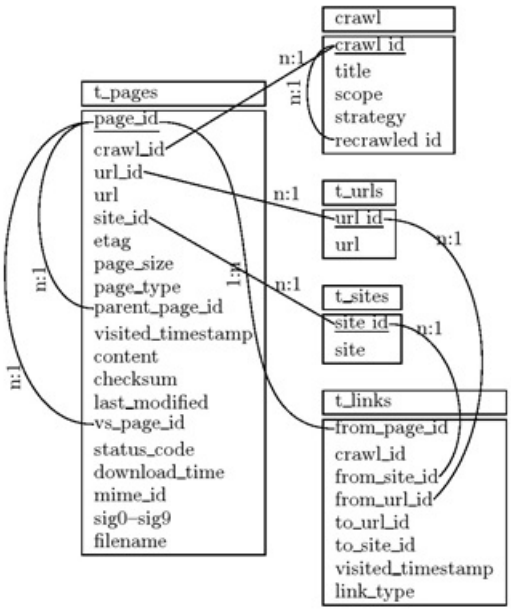
نمودار 10 شمای DB

عکس


```
create table clean_mapping as
select dirty_url.url_id as dirty_url_id,
       clean_url.url_id as clean_url_id
from (
  select min(url_id) as url_id, lower(url) as url
  from t_urls group by lower(url)
) as clean_url, t_urls as dirty_url
where clean_url.url = lower(dirty_url.url);

create table lower_url_dd as
select  crawl_id,
  clean_mapping.clean_url_id as url_id,
  lower(min(url)) as url, site_id as site_id,
  min(etag) as etag,
  min(page_size) as page_size,
  min(page_type) as page_type,
  min(visited_timestamp) as visited_timestamp,
  min(checksum) as checksum,
  min(last_modified) as last_modified,
  min(status_code) as status_code,
  min(download_time) as download_time,
  min(sig0) as sig0, min(sig1) as sig1,
  min(sig2) as sig2, min(sig3) as sig3,
  min(sig4) as sig4, min(sig5) as sig5,
  min(sig6) as sig6, min(sig7) as sig7,
  min(sig8) as sig8, min(sig9) as sig9,
  min(mime_id) as mime_id,
  min(filename) as filename
from t_pages_dd, clean_mapping
where t_pages_dd.url_id
      = clean_mapping.dirty_url_id
group by t_pages_dd.crawl_id, t_pages_dd.site_id,
      clean_mapping.clean_url_id
```

فهرست ۳. SQL به کار رفته برای پاکسازی URLهای پایین‌تر



نمودار ۱۰. شمای DB

ص: 86

6.درسهای آموخته شده و کارهای آینده از نظر یک آرشیویست آرشیو کردن مطلوب ،وب جلوگیری از تغییرات محتویات در حین عمل خزش است. البته این یک توهم و عملاـ نشدنی است در نتیجه ممکن است یک نفر هرگز مطمئن نشود که محتویات که تا کنون جمع کرده است هنوز با محتویاتی که بعدا جمع خواهد شد منطبق است. به هر حال انسجام در آرشیو وب یک موضوع کلیدی برای انسجام خزش جهت دادههای رقومی، در یک حالت قابل تکثیر و تفسیر است به این منظور ما گستره ای از Heritrix که با ارتباطات صحیح و همچنین تاریخ نامناسب محتویات سازگار است را توسعه داده ایم. به علاوه ما قادر هستیم شکل انسجام را مؤثرتر بدون توجه به تکیه بر وب سرور تحلیل کنیم به همین ترتیب توسعه تحلیل و تصویرسازی ویژگیها در کمک به مهندسان خزش برای درک بهتر ذات نقص در انسجام درون و بین وبگاهها و سازگاری راهبردهای Crawling برای خزشهای آینده مفید است در نتیجه مقاله به افزایش انسجام آرشیو هم کمک خواهد کرد.

در حالی که اکنون نقص در انسجام به ما در درک عدم تطابق به صورت سیستمی تر کمک میکند، تحقیقات آینده نیازمند یک بینش مولد و تولید کننده است. به علاوه، تحقیقات در حال پیشرفت به عمل خزش ناتمام و افزایش پوشش آرشیو کمک میکند حتی ترکیب بخشی از Recrawl در ترکیب با یک بخش افزایشی خزش ممکن است جذاب و مؤثر شود. به علاوه نتیجه به دست آمده از تحلیل خزشهای حقیقی برای ایجاد محیطهای شبیه سازی پیشرفته مفید خواهد بود همچنین قادر خواهیم بود رفتارهای تغییرات را در دنیای واقعی وبگاه در یک محیط شبیه سازی شده مشاهده کنیم.

سپاسگزاری

این کار توسط هفتمین برنامه FrameworkIST از E توسط تمرکزی کوچک یا متوسط پروژه های تحقیقی (STREP) روی وب آرشیوهای زنده (LiWA) با شماره 2162670 حمایت شده است. همچنین ما از همکارانمان برای مباحث امید بخششان متشکریم.

منابع

[1] Brian E. Brewington and George Cybenko. Keeping up with the changing web. Computer, 33(5):52{58}, May 2000.

[2] Junghoo Cho and Hector Garcia-Molina. The evolution of the web and implications for an incremental crawler. In VLDB 400: Proceedings of the 26th International Conference on Very Large Data Bases, pages 200{209, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[3] Junghoo Cho and Hector Garcia-Molina. E_ective page refresh policies for web crawlers. ACM Transactions on Database Systems, 28(4), 2003.

.Junghoo Cho and Hector Garcia-Molina. Estimating frequency of change. ACM Trans [4]

.Inter. Tech., 3(3):256{290, August 2003

Junghoo Cho, Hector Garcia-Molina, and Lawrence Page. E_cient crawling through url [5]

ordering. In WWW7: Proceedings of the seventh international conference on World Wide

Web 7, pages 161{172, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier

.Science Publishers B. V

L. Clausen. Concerning etags and datestamps. In A. Rauber J. Masan_es, editor, 4th [6]

.International Web Archiving Workshop (IWAW'04), 2004

Dimitar Denev, Arturas Mazeika, Marc Spaniol, and Gerhard Weikum. Share: Framework [7]

for qualityconscious web archiving. In VLDB '09: Proceedings of the 35th international

.conference on Very Large Data Bases. VLDB Endowment, 2009

.International Internet Preservation Consortium. Arc ia, internet archive arc le format [8]

.http://www.digitalpreservation.gov/formats/fdd/fdd000235.shtml

//:International Internet Preservation Consortium. Warc, web archive_le format. http [9]

.www.digitalpreservation.gov/formats/fdd/fdd000236.shtml

Panagiotis G. Ipeirotis, Alexandros Ntoulas, Junghoo Cho,and Luis Gravano. Modeling [10]

.and managing changes in text databases. ACM Trans. Database Syst., 32(3):14, 2007

Adam Jatowt, Yukiko Kawai, and Katsumi Tanaka. Detecting age of page content. In [11]

.WIDM, pages 137{144, 2007

.Julien Masan_es. Web Archiving. Springer, New York, Inc., Secaucus, NJ, 2006 [12]

-Frank McCown and Michael L. Nelson. Evaluation of crawling policies for a web [13]

.repository crawler. In Hypertext, pages 157{168, 2006

Frank McCown, Joan A. Smith, and Michael L. Nelson. Lazy preservation: reconstructing [14]

.websites by crawling the crawlers. In WIDM, pages 67{74, 2006

G. Mohr, M. Kimpton, M. Stack, and I. Ranitovic. Introduction to heritrix, an archival [15]

.quality web crawler. In 4th International Web Archiving Workshop (IWAW'04), 2004

Marc Najork and Janet L. Wiener. Breadth-_rst search crawling yields high-quality [16]

.pages. In In Proc. 10th International World Wide Web Conference, pages 114{118, 2001

S_ergio Nunes, Cristina Ribeiro, and Gabriel David. Using neighbors to date web [17]

.documents. In WIDM, pages 129{136, 2007

Christopher Olston and Sandeep Pandey. Recrawl scheduling based on information [18]

longevity. In WWW '08: Proceeding of the 17th international conference on World Wide

.Web, pages 437{446. ACM, 2008

M. Spaniol, D. Denev, A. Mazeika, P. Senellart, and G. Weikum. Data Quality in Web [19]

.Archiving. In Proceedings of WICOW, Madrid, Spain, April 20, 2009, pages 19 { 26

.ACM Press, 2009

Qingzhao Tan, Ziming Zhuang, Prasenjit Mitra, and C. Lee Giles. E_ciently detecting [20]

.webpage updates u In ICWE, pages 285{300, 2007

چکیده

بایگانی وب پنهان سخت تر از بایگانی وب سطحی است. روش اصلی گردآوری محتوای وب بر پایه یافتن مسیر است هر صفحه در وهله اول باید به وسیله خزشگر پیدا شود تا بتوان آن را واکشی و بایگانی .کرد تاکنون روش مناسبی برای بایگانی وب پنهان پیش بینی نشده است. این امر نیازمند پیشرفتهایی برای حفاظت از وب پنهان از طریق راههای ساده و تکامل فنی وب است دو دلیل وجود دارد که بایگانی وب پنهان نباید مورد غفلت واقع شود نخست وب گستره وسیعی دارد و دارای منابع ارزشمندی است که بسیاری از مؤسسه های میراث فرهنگی به آن علاقه مندند. اینکه احتمالاً وب با معماریهایی از اطلاعات تکامل می یابد که در برابر شیوههای سنتی خزشگر مقاومت می.کنند این مقاله به بررسی ویژگیهای وب ،پنهان مسائل بایگانی وب ،پنهان مسیرهای بایگانی کردن و فناوریهای بایگانی وب پنهان می پردازد.

بایگانی وب پنهان

*(1)

:نوشته ژولین ماسانه(2) | :ترجمه افسانه تیموری خانی(3)

مقدمه

همـان طور که در فصـلهای قبلی ملاـحظه کردیم روش اصـلی گردآوری محتوای وب بر پایه یافتن مسـیر است با توجه به اینکه پیمان نامه HTTP قابلیت تهیه سـیاهه کامل را ،ندارد هر صـفحه در وهله اول باید به وسـیله خزشگر پیدا شود تا بتوان آن را واکشی و بایگانی کرد در فصل یک دیـدیم که میدانیم که خزشـگر محدودیت زمانی قابل توجهی را برای پردازه گردآوری کامل ارائه میکند اما لازم است که حداقل یک مسـیر برای بایگانی هر مدرک وجود داشته باشد که البته این امر همیشه بعید به نظر میرسد در واقع بخش عظیمی از وب به همین دلیل توسـط ابزارهای خودکار قابل دسترس نیست . این بخش برای اولین بار، در 1994، توسط جیل اچ الزوورث(4)، وب نامرئی نامیده شد (برگمن2001(5))؛ زیرا بخشـی از وب است که توسط موتورهای کاوش خزشـگر نمایه نمی شود. بعـدها، پیشـنهاد شد که این وب را وب عمیق نامگذاری کنند – در مقابل وب سطحی یا وب قابل نمایه سازی عمومی(پی آی دبلیو)(6) (لارنس(7)

---

1 – Archiving the Hidden Web: in Masanes, Julien (ed.), Web Archiving. Berlin Heidelberg New .York:Springer.pp.115-128

2 – Julien Masanè s بایگانی وب اروپا julien@iwaw.net

3 – دکترای کتابداری و اطلاع رسانی و کارشناس سازمان اسناد و کتابخانه ملی ایران

4 – Jill H. Ellsworth

5 – Bergman

6 – Publicly indexible web (PIW)

7 – Lawrence

و گیلز1999(1)) – زیرا خزشگرها به راحتی میتوانند به آن دسترسی داشته باشند.

ما در اینجا از این دو اصطلاح استفاده نمیکنیم زیرا سبب ابهام بیشتر می شود. نخستین اصطلاح یعنی وب «نامرئی میتواند به این فکر منتهی شود که مشکل اصلی نمایش یا ارائه صفحه هاست؛ در صورتی که مشکل اصلی دسترسی ابزارهای خودکار است اصطلاح دوم وب عمیق ممکن است با عمق منابع در ساختار سلسله مراتبی ابر متنی وبگاهها اشتباه گرفته شود. بنابراین ترجیح داده ایم برای تعیین بخشی از وب که خزشگرها قادر به رسیدن به آن نیستند از اصطلاح وب پنهان مخفی» استفاده کنیم. به یاد داشته باشیم که حدود آن از نظر فنی تعریف شده است و به هیچ نوع خاصی از محتوا و تجربیات مشترک انسانی در محدودیتهای ناوبری اشاره نمیکند و دیگر اینکه این تعریف کاملاً فنی است و مرز این دسته (گروه) ممکن است با تکامل فناوری تغییر یابد.

برای مثال، سایتهایی با توابع پیمایش رمزگذاری شده در فلش FLASH ساخته شده اند که قبل

از اینکه ماکرومدیا Macromedia نوعی SDK قادر به استخراج پیوندها از کد توابع را منتشر کند، به عنوان وب پنهان در نظر گرفته میشدند این سایتها به خزشگر امکان میدهند تا از طریق وبگاههای کدگذاری شده فلش راه خود را پیدا کنند بنابراین این نوع وبگاهها نمیتوانند بیش از این به عنوان وب پنهان در نظر گرفته شوند.

اگر چه بیشتر توجه این فصل و بقیه فصلها به محدودیتهای خزشگرها معطوف است. شایان ذکر است که خزشگرها تمایل دارند صفحه های را کشف کنند که به ندرت توسط انسان دیده میشوند. در واقع مطالعات انجام شده در این زمینه به وسیله بوف خواد(2) و ووی نوت(3) (2003)، از طریق استفاده از تراکنش های خدمات وب اینریا(4) 2002 نشان داد تعداد قابل توجهی از صفحه های را که خزشگرها پیدا میکنند کاربران از دست میدهند.

پیدا کردن حداقل یک مسیر به اسناد

صفحه های متصل به وبگاه ها با هم یک مسیر وب را برای دسترسی به اسناد تشکیل می دهند. هر صفحه میتواند به یک یا چندین صفحه پیوند داده شود و خزشگرها نخستین صفحه ای را که به آن میرسند میگیرند به این ترتیب صفحههای پیوند داده نشده از وبگاههای دیگر حتی اگر در همان سرور ساکن باشند، شامل آن نمیشوند این مسیر وب میتواند به هر سندی گسترش یابد، به طوریکه از طریق مسیرهای مجازی که با پرس وجو ساخته میشوند تولید یا قابل دسترس شوند به طور مثال صفحه های ،پویا که از طریق محتوای ذخیره شده در پایگاه داده ایجاد میشود تنها به صورت مجازی وجود دارند که به عنوان بخشی از یک مسیر وب در نظر گرفته میشوند نکته مهم این است که چگونه مسیر اشیا ساخته می شود و آیا خزشگرها میتوانند آن را پیدا کنند یا نه.

1- Giles

2- Boufkhad

3- Viennot

4- INRIA

در اینجـا میتوان دو مورد کلی را مشـخص :کرد نخست اینکه مجموعه هـایی از مسـیرهای با ارزش از پیش تعریف شـده و محـدود وجود دارد؛ که در واقع نمونه ای از ابر پیونـدها و منوهاست. خزشـگرها میتواننـد چنین مسـیرهایی را با اسـتفاده از زبانی که برای رمزگـذاری آنها استفاده شده استخراج، تفسیر،ودنبال کنند.

دومین مورد که برای خزشـگرها هم بسـیار مشـکل است نیـاز به تعامل صـریح کاربر دارد (منظور بیشتر از یکبار کلیک کردن است. برای خزشـگر تقیلد مشکل است مانند زمانی که کاربران مجبورند از طریق فرمت HTML وارد پرس وجو شوند تا بتوانند به اسناد خاصی مثل یـک ،تصویر یک مقاله و یا یک صـفحه پویا دسترسـی یابنـد این نوع وب پنهان – که وب پنهان ساختاری نامیـده می شود– شامل مقـدار زیـادی محتواست که از آن برای انتشـار در انبـاره اسـناد بزرگ وب اسـتفاده شـده است، که هم ساختار یافته است (پایگاه داده) و هم غیر ساختاریافته مثل مجموعه ای از تصاویر مقالات علمی موسـیقی و مانند آن (اسـتوری(1) و جنکه 1999 (2))راه مناسب برای انتشار این قبیـل مجموعه هـای بزرگ ،محتوا اسـتفاده از پیونـدهای اختصاصـی نیست؛ بلکه بهتر است از دسترسـی پویا از طریق دروازه های پایگاه دادهها اسـتفاده شود که شامل اطلاعات توصـیفی برای هر مورد نیز میباشـد. امروزه سایتهای زیادی از این نوع معماری اطلاعات شـکل، پایگاه داده و گردآوری اسـتفاده میکنند که بخشـی از وب پنهان محسوب می شوند. این نوع معماری را دروازه مستند می نامیم زیرا ورود به فضـای بزرگ اطلاعاتی را از طریق تعاملات (جست وجو) فراهم میکنـد (ماسانه(3) 2002). در برخی موارد رابـط جایگزین مرورگر به گونه ای ارائه شده است که خزشگر هنوز به محتوای چنین انبارههایی دسترسی دارد در این صورت آنها در نوع اول قرار میگیرند.

در اینجا انواع تعامل کاربر را با جزئیات بیشتر نقد میکنیم تا ببینیم چه مشکلی برای خزشگرها

ایجاد میشود.

انواع تعامل با کاربر

لدسچر(4) و گوپتا(5) (1999)، نوعی مدل منبع تعاملی پیشنهاد دادند این مدل شامل چهار نوع عنصر ورودی است:

– ابر پیوندها روشی کلاسیک برای تهیه ورودی کاربر که محدود به یک نفر است؛

– منوها برای انتخاب زیر مجموعه ای از مقدارها از مجموعه ای از پیش تعریف شده ؛

– فرمها پیوندهای پویا با ویژگیهای ورودی متعدد (چندگانه)؛ و

– نوع چهارم که میتوانـد تنها بر اساس این مـدل با تعامل صـریح کاربر حمایت شود و عناصـر بدون لفاف نامیده میشوند تصاویر نقشـهها تعامل گرافیکی کاربر بر اساس جاوا و مانند آن).

1- Storey

2- Jahnke

3- Masanè s

مثال آن در دنیای واقعی توزیع عناصـر ورودیهای مختلف در وبگاههای تصاویر غنی پزشکی است. (فرانکویچ(1) و پروکوش(2) را ببینید (2001) هر یک از این عناصـر ورودی موردی متفاوتی برای خزشگر است. دو مورد اول) فراپیوندها و منوها تنها مشکلات تفسیر را برای خزشگرها ایجاد میکنند و با هم چیزی را میسازند که ما آن را عناصر ورودی تعریف شده می نامیم.

عناصر ورودی ارزش باز (معمولاً فرمت HTML)که در مقابل مورد دوم هستند؛ فضای امکان به یک یا چند مسیر تعریف شده را کاهش میدهد آنها در زمینههای مختلف استفاده می شود و ما در مورد آنها بیشتر بحث خواهیم کرد. نوع چهارم، گرچه نسبت به انواع دیگر کمتر معمول است؛ دسترسـی به آن برای خزشگرها بسیار سختتر است. برای آگاهی از جزئیات نوع شناسی مشکلاتی که خزشگرها با آن مواجه شـده اند دو گزارش کنسرسـیوم بین المللی حفاظت اینترنت را ملاحظه کنید (بویکو(3) 2004؛ ماریل و همکاران 2004)در بخش زیر موارد اصلی و مباحثی در مورد مشکلات به وجود آمده برای خزشگرها را مطرح خواهیم کرد.

تعریف مقدار عناصر ورودی

پیوندهای ساده HTML یکی از رایج ترین این نوع است که مسـیر راحتی را برای جست وجو فراهم میکند. پیوندهای نسبی، ممکن است بعضـی از موارد تفسـیر را افزایش دهند آنها بر اساس قوانین شـرح داده شده RFC 1808 و RFC 2369 شبیه نشانیهای یونیکس هستند مکان نسبی، از راهنمای جاری شـروع میشود و با استفاده از اسلش (/) پایین میرود و از طریق راهنمای والد با استفاده از دو نقطه (...) بالا میرود یـک مسـیر نسـبی کـه بـا اسـلش شـروع میشـود بـه ایـن معنی است که راهنمـای ریشه میزبـان است مشـکلات در اثر اسـتفاده بد پدیدآروندگان صـفحه ها یا مدیریت سیستم محتوا به وجود می آید. به طور مثال شامل نقاط ،اضافی راهنماهای فوق عددی ،والد و یا دیگر ترکیبات عجیب و غریب می شود. در صورت ،وجود تگ ،اصـلی در بخش رأس یک صفحه HTML یک مکان پیشفرض جایگزین برای پیوندهای نسبی تهیه میکند که تمامی ،پیوندها به جای شـروع از راهنمای جاری، از آن شروع شوند. از پیوندهای جاوا اسکریپت به خوبی دیگر زبانهای برنامه نویسـی برای ایجاد پیوندهای خاص برای منوها و ناوبری طومارنمایی (پیمایش) تقویم ،پویا و مانند آن استفاده میشود از آنجاکه هر پیوند نتیجه ترکیبی از دسـتور متغیر یا ورودی تعامل کاربر است؛ در بعضی موارد تفسیر پیوندها بدون اجرای اسکریپت تقریباً غیر ممکن خواهد بود.

خزشگرها می توانند تفسیر مبتنی بر قواعد(4) و یا هر ترکیب ممکنی از مسیر یا نام فایل را برای پیدا کردن اسکریپت(5) داشته باشند در هر مورد موفقیت کامل تضمین شده .نیست راه حل جایگزین – که هنوز

1- Frankewitsch

2- Prokosch

3- Boyko

4- This is the case for instance of the open-source web copier HTTrack (see Chap. 3 (Roche 2006).

5- Heritrix, the open source archive-quality crawler developed jointly by the Internet Archive and the nordic libraries in the IIPC, implements this approach (Mohr et al. 2004)

آزمایش نشده – تفسیر با اجرای کدها به جای تجزیه است که به معنی اجرای مرورگرها شبیه سازی زمینه ها و تعامل کاربر است. بعضی از انواع ،پیوندها مثل پیوندهای فریم و پیوندهای تصویر ممکن است مشکلاتی برای تفسیر به وجود آورند؛ ولی به طور کلی میتوانند با موفقیت دنبال شوند.

عناصر ورودی مقدار باز

در همه موارد ،قبلی با وجود مشکلات ایجاد شده دامنه محدودی از مسیرهای ممکن توسط کد تعریف شده اند که باعث میشود تفاوت زیادی با نوع دوم به وجود آید که در آن میتوان یک مجموعه بی نهایت و تعریف نشده از مقادیر را به پیوندها اختصاص داد.

مکانیسم اصلی برای این مورد فرمت HTML است آنها کاربران را قادر میسازند که یک مقدار دلخواه را به سرور عبور دهند مقادیر ورودی به طور مثال میتوانند برای پرس و جوی اسناد مورد استفاده قرار گیرند که ناشی از تولید مجموعه ای از پیوندها به صفحه ها یا اسناد .است وارد کردن یک پرس وجو حاوی اطلاعات نویسنده و عنوان در سیاهه میتواند یک سیاهه انتشار همراه با پیوند به هر یک از آنها ایجاد کند این پیوندها از یک پایگاه داده ایجاد و در نتایج صفحه HTML جاسازی شده اند. اگر هیچ پیوند دیگری به این اسناد – به عنوان مثال از طریق سیاهه مرتب شده نشریات بر اساس حروف الفبا وجود نداشته باشد خزشگرها تنها از طریق مسیر مجازی ایجاد شده توسط رابط پرس وجو میتوانند به آنها دسترسی پیدا کنند.

این نوع معماری اطلاعات به دروازه مستند معروف است(مازانه 2002)که بسیار معمول است و باید از استفاده دیگر فرمها متمایز باشد (شکل 1) فرمها به طور عمده برای جمع آوری ورودی کاربر مانند ورود به سیستم ورود اطلاعات تماس و یا بازخورد ارسال نظرات جعبه جست وجوی عمومی و مانند آن مورد استفاده قرار میگیرند کوپ(1) و همکارانش در سال 2003 دریافتند که حدود 50 درصد از فرمهای HTML رابط جست و جو هستند و لاگ(2) و همکارانش در سال 2002 در نمونه خود یافتند که95 درصد از ،فرمها از جمله جعبه جست و جوی عمومی فرمهای ناخواسته بودند.

اما حتی اگر بسیاری از فرمهای موجود بر روی وب برای مقاصد دیگر استفاده شوند، آنهایی که باقی مانده اند نقطه ورود به فضای اطلاعات بزرگی هستند که وب پنهان نمایش میدهد. این دو مطالعه تلاش کرده اند تا آن را مشخص کنند.

ویژگیهای وب پنهان

اولین مطالعه در سال 2000 توسط برگمن برگمن (2001) با استفاده از تحلیل همپوشانی بین جفت موتورهای کاوش به منظور برآورد تعداد وبگاههای پنهان صورت گرفت. آنها مشخص کردند که طیف وسیعی بین 43000–96000 وبگاه پنهان براساس حضور فرم وجود دارد. متأسفانه، فیلترهای مورد

Cope –1

Lage –2

استفاده مستند نشده اند؛ بنابراین به سختی میتوان درباره ارزش این نتایج و مقایسه آنها با دیگران قضاوت کرد. آنها همچنین به تحلیل 60 موتور کاوش بزرگ پرداختند و اندازه آنها را 550 بیلیون صفحه برآورد کردند که 550 برابر بزرگتر از وب سطحی در آن زمان است.

فرض اصلی در پشت این مطالعه این است که هر مورد در پایگاه داده با یک صفحه ایجاد شده

مرتبط است و اندازه آن با اچ تی ام. ال. تخمین زده می شود.

این ،اصل شامل تمامی HTML و اطلاعات مربوط به کد(HTML) به اضافه محتوای متن

،استاندارد تصاویر تعبیه شده منحصر به فرد و اطلاعات استاندارد سرآیند HTTP(پیمان نامه انتقال ابرمتن) است. استفاده از این پیمان نامه استاندارد اجازه میدهد تا مقایسه دقیقی بین وب سطحی و عمیق صورت گیرد (برگمن 2001).

برای مثال ورودی پایگاه دادههای آب و هوایی ملی ایالات متحده (بزرگترین مثال در نمونه خود)

با یک صفحه 13 کیلوبایتی مطابقت میکند.

واقعیت این است که این بانک اطلاعاتی و پایگاه داده NASA EOSDIS تقریباً 80 درصد از کل نمونه ها را ارائه میدهد و نشان میدهد که این مطالعه نگاهی سو گرفته به وب پنهان نسبت به محتوای تکراری و غیر مستند دارد اگر چه در وهله اول به نظر میآید این مطالعه با هدف توجه به اهمیت و غنای این بخش از وب صورت گرفته است.

مطالعه اخیر توسط چانگ(1) و همکارانش (2004)، جزئیات بیشتر و ویژگیهای مستندی از وب پنهان را آشکار میکند. این مطالعه تمایز بین پایگاه داده ساختار یافته یعنی پایگاه داده مرتبط با مقادیر زوج کلیدی و محتوای بدون ساختار (متن، عکس ،شنیداری ،دیداری) را مشخص مینماید که در این کتاب محتوای دروازه مستند نامیده میشوند در مطالعه چانگ و همکارانش همانند بسیاری از مطالعات انجام شده بر روی وب ،پنهان توجه اصلی به نوع اول پایگاه داده ساختار یافته است که توسط پژوهشگران جامعه پایگاه داده که علاقه مند به یکپارچه سازی دادههای وب هستند – ساخته شده اند. اگر چه یافته های بسیاری مثل یافته ما بیشتر از منظر محتوا گرا مورد نظر هستند.

آنها مطالعات خرد و کلانی انجام داده اند مطالعه کلان بر روی یک میلیون نشانی تصادفی تولید IP صورت گرفت که برای پیدا کردن سرور HTTP مورد آزمایش قرار گرفتند 2260 وبگاه پیدا شد و مورد خزش قرار گرفت؛ 126 مورد وبگاه پنهان شناسایی شد که شامل 190 پایگاه داده است. به گزارش وب ،جهانی این به معنی 307000 سایت حاوی 102000 دروازه مستند و 348000 پایگاه دادههای ساختاریافته است. این نکته اهمیت وبگاههای پنهان را نشان میدهد. همچنین، چانگ و همکارانش توزیع پایگاه داده وب را در عمق شناسایی کردند که نشان میدهد 91/6 درصد از آنها در داخل عمق 3 یافت شدند. در مطالعه خرد، 441 منبع را با جزئیات بیشتر مورد بررسی قرار دادند. آنها در آغاز نشان دادند که در بسیاری از موارد مسیر ناوبری جایگزینی برای رسیدن به محتوا وجود دارد که در آن، محتوای نمونه برای موتورهای کاوش واقعا پنهان نیست. در واقع پنهان بودن وبگاه بستگی به دامنه دارد.

Chang -1

آنها طرح رابط پرس وجو و تعدادی از ویژگیهای دقیقتر را مورد مطالعه قرار دادند که کوچکترین اندازه طرح 1 بزرگترین 18 ، و حد وسط 6 است. همچنین به مطالعه واژگان طرح پرداختند و پنج ویژگی اصلی طرح عنوان کلید واژه ها»، «قیمت»، «ساخت» و «هنرمند» را نشان دادند.

گفته چانگ و همکارانش تشویق به پردازش خودکار میتواند باعث ایجاد نظم و قاعده شود ظاهراً بررسیهای ما پدیدهای دوگانه را نشان میدهد که با هم سیاهه ویژگیهای منحصر به فرد مرز

وب عمیق را مشخص میکنند:

نخست به عنوان یک چالش منابع در ونخطی عملاً نامحدود هستند؛ حتی برای یک حوزه خاص مورد علاقه منابع جایگزین بی شماری پدیده تکثیر منابع وجود دارد. بنابراین، یکپارچه سازی در مقیاس بزرگ چالشی واقعی است.

دوم به عنوان یک فرصت مطرح است با این حال زمانی که منابع در حال تکثیر هستند، در مجموع، پیچیدگی آنها برای نشان دادن زیر بنای ساختار تمایل به هماهنگی دارد به طور خاص ما این ساختار هماهنگ را در ویژگیهای واژگان و الگوهای پرس وجو در منابع وب مشاهده کردیم این نوع تجمیع واژگانی در موقعیت و اندازههای همگرا دسته بندی میشود (چانگ و دیگران 2004).

بایگانی وب پنهان سرویس گیرنده

همان طور که در بخش قبلی نشان داده شده است نقطه ورودی وب پنهان (فرم) قواعد مربوط را نشان میدهد که برای استخراج خودکار محتوا با لفافه مورد استفاده قرار می گیرد و گاهی عوامل وب پنهان نامیده می شوند (راگهاوان(1) و گارسیا-مولینا 2001(2)؛ لاگ(3) و دیگران 2002؛ برای مقدمات کلی درباره این عنوان هرست(4) ، 1998؛ و آدامز(5) ، 2001 را ملاحظه بفرمایید).

نقش این عوامل شناسایی فرمهای HTML آموزش پر کردن آنها شناسایی و واکشی نتایج محتواست این فرآیند میتواند برای ارائه یک رابط جست و جوی یکپارچه اجرا شود (برای مثال این عناوین را بررسی ،کنید ،فلورسکو(6) و دیگران 1998 مثال برای اجرای خدمات جست وجو بورگمن 200 و نمونه پیشرفتهای اخیر در این حوزه هی(7) و دیگران2005) معمولا ـ ردیابی در حالت طبیعی خزشگرها از طریق تجزیه و تحلیل صفحه های حاوی فرمهای HTML صورت میگیرد برای از بین بردن فرمهای نامطلوب از روش اکتشافی (فناوری در هوش مصنوعی) استفاده میشود (صفحه ورود به سایت و یا ارتباط با صفحه اطلاعات جعبه جست وجوی عمومی و مانند آن). سپس عوامل زمینه ای پرس وجو و برچسبها را استخراج میکنند و سعی میکنند آنها را با برچسبهای شناخته شده مقایسه

1- Raghavan

2- Garcia- Molina

3- Lage

4- Hearst

5- Adams

کنند و گاهی اوقات برای ارزیابی ،موضوع آنها را با واژگان شناخته شده بررسی میکنند(گراوانو(1) و دیگران 2003).

در نهایت فرمها به صورت خودکار پر شده و در نتیجه صفحه ها یا اسناد ذخیره میشوند. روش اکتشافی (فناوری در هوش مصنوعی) که توسط لاگ و همکارانش (2002) استفاده شده است فرمهای ناخواسته را فیلتر میکند و اجازه میدهد تا فرمها با کمترین عناصر از بین برده شوند و عناصر نوعHTML با هر رمز عبوری ساخته شوند کوپ و همکارانش (2003) درخت تصمیم گیری شفافی را برای کشف رابط جست و جو بر روی وب شکل (2) بر اساس تجزیه فرمهای HTML با توجه به ویژگیهای ذره ای (اتمی) مانند HTTP ،دامنه کنترل ،متن کنترل رمز عبور و نظیر آن پیشنهاد کردند که براساس مجموعه آموزش و یادگیری الگوریتم ساخته شده است.

پرکردن فرمها، به طور خودکار مستلزم درک زمینه است. لاگ و همکارانش (2002)، فرض کردند که برچسبها معمولاً در گوشه سمت چپ و یا بالای زمینه فرم قرار داده میشود. جانگ(2) و همکارانش (جانگ و همکارانش 2004 فرض کردند که نظم و قاعده ای یا الگوهای طراحی خاصی در میان فرم پرس و جوی وب وجود دارد که با هم نوعی زبان دیداری قابل تجزیه را تشکیل میدهند آنها نوعی ابزار تبدیل پرس و جوی فرم HTML به مجموعه ای از نشانه ها را به وجود آوردند که هر یک نشان دهنده یک عنصر دیدنی ذره ای در چارچوبی دو بعدی است. آنها اخذ الگوهایی مانند قرابت همجواری و رابطه معنایی میان اصطلاحات را پیشنهاد کردند. با استفاده از الگوریتم تجزیه آنها دقت و پوشش 80/0 را برای تشخیص فرمها به دست آوردند.

هنگامی که این کار انجام شد لازم است اصطلاحات به صورت خودکار، از طریق فرمها، جهت ایجاد پاسخ ارائه شوند مشکلات دیگری در هنگام تکمیل فرم به طور خودکار وجود دارند(نگاه کنید لیدل(3) و دیگران2002) ورود به فرمها میتواند محدود به فیلدهای متنی کلید رادیویی(4) جعبه بازبینی(5) سیاههها، و مانند آن و یا هر نوع فایل پیوست کدگذاری شده MIME باشد درخواست منطقی می تواند به اشکال مختلف با اطلاعات دولتی گرفته شده از روی سرور تقسیم گردد کوکیها فیلدهای پنهان مقادیر کدگذاری شده به URL پایه برخی فرمها قبل از ارسال فرم برای تغییر فیلد بر اسکریپتها تکیه میکنند (بازبینی گسترده(6)، دیگر اعتبارات ،فیلد و محاسبه خودکار بعضی از فیلدها).

اگر واژگان اصطلاحات ارائه شده درست تعریف شده باشند فرصت مناسبی برای واکشی محتوا به وجود میآید این مسئله زمانی مهم است که دامنه این واژگان محدود باشد (مانند کد پستی، تاریخ، و مانند آن) در واقع یکی از این زمینه ها کافی است به عنوان مثال دروازه مستند به متون فیلسوفان فرانسوی که یک ورودی تاریخ را ارائه میدهد میتواند با تمام تاریخها از سال 1100 - 2005 بدون هیچ

1- Gravano

2- Zhang

3- Liddle

4- radio buttons

5- check boxes

6- range checking

پیشفرضی در مورد نام نویسنده یا عنوان نوشته ها پرس وجو شود در 905 پرس وجو فرد می تواند مطمئن باشد که میتواند تمام متون را به دست آورد.

محدودیت این روش زمانی است که دامنه پرس و جو برای تمامی فیلدهای بیش از حد باز یا تعریف نشده به صورت نظام مند بررسی شود. در این موارد امکان استفاده از یک رویکرد دیگر امکان پذیر است که از یک پرس وجو برای استخراج اصطلاحات پرس و جوهای جدید سوء استفاده میکند که سپس ارسال و تکرار خواهد شد روشهای ارائه شده توسط کالان(1) وکانل(2)(2001) نمونه برداری پرس و جو محور نامیده میشود و با موفقیت استفاده شده است(آژیچتین(3) و همکارانش 2003 ، و (بربوزا(4) و فریره(5) 2004) ندولاس(6) و همکارانش (2005) پیشنهاد کرده اند که از الگوریتم تطابق برای انتخاب بهترین و مهمترین کلمات – کلماتی که به بسیاری از سندها مربوط اند – استفاده شود آنها نشان میدهند که تنها با 83 پرس وجو تقریباً 80 درصد از 14 میلیون اسناد ذخیره شده در پاپ مد(7) را میتوان بارگذاری کرد.

همکاری سرور خزشگر

آشکار کردن دروازه مستند خزشگر

زمانی که همکاری با پدید آورندگان وبگاه امکان پذیر است این امکان وجود دارد تا محتوای وبگاه پنهان برای خزشگر به منظور بایگانی شدن آشکار شود. در اینجا اغلب برای بایگانی کردن از روشهای ارائه شده برای موتورهای کاوش استفاده میشود آنها شامل ایجاد سیاهه کاملی از اسناد و یا فراهم آوردن امکان دسترسی به یک خدمت هستند که میتوانند به صورت خودکار توسط خزشگر مورد پرس وجو قرار گیرند آنها طیف وسیعی از صفحه های تولید شده هستند که اغلب برای کاربران انسانی پنهان اند و به همه اشیای وبگاه حتی پیمان نامه های پرس و جو اختصاص دارند پیمان نامه متعددی به جز چند مورد از جمله پیمان نامه ..ای.آی. از اواخر دهه 90 با موفقیتهای محدودی پیشنهاد شده اند. ما مهمترین آنها را در زیر مرور میکنیم.

1 – صفحه های پیوندهای پنهان

ساده ترین راه برای فعال کردن خزشگر به منظور دریافت محتوای ،پنهان ایجاد سیاهه پیوندهاست که به همه اسناد وبگاه پنهان اشاره دارد این امر به ویژه برای دروازه مستند مناسب است و میتواند به طور کامل از دید کاربران عادی با پنهان کردن این صفحه ها از ناوبری ،طبیعی پنهان بماند؛ به عنوان مثال با تعبیه پیوندهای پنهان در صفحه اصلی.

1- Callan

2- Connell

3- Agichtein

4- Barbosa

5- Freire

6- Ntoulas

اینکار برای مثال توسط کتابخانه ملی فرانسه و کتابخانه ملی استرالیا از طریق آشکار کردن مجموعه برای موتورهای کاوش انجام شد:

سیاهه جداگانه ای از URLها برای هر یک از مجموعه های دیجیتال کتابخانه یعنی عکسها، نقشه ها، صفحه های موسیقی نسخه های خطی کتابها و پایندها ساخته شدهاست. هر مجموعه شامل هزاران مورد است که در یک سری از صفحه های وب سیاهه شده اند هر کدام شامل 100 پیوند هستند که اقلام (موارد) مجموعه را تفکیک میکنند این صفحه با استفاده از دستور ربات anoindex, follow موتورهای کاوش را مستقیم هدایت میکند تا با دنبال کردن پیوند به محتوا برسد؛ اما صفحه ها را سیاهه نمی کند. سیاهه URL خود به صورت پویا تولید میشود و با محتوای جدید به سیاهه اقلام جدید دیجیتالی شده به صورت خودکار اضافه میگردد و توسط اینترنت قابل دسترس میشود (بوستن(1) 2005)

برای مؤثر بودن طرح پیوند پایدار باید در جای خود قرار گیرد و با تمام اسناد در پیوند باشد. براند من(2) و همکارانش (2000) محاسبه کردند که چنین مکانیزمی میتواند تا 80 درصد در پهنای باند تبادل خزشگر سرویس دهنده صرفه جویی کند. از سال 1990، طرحهای پیشنهادی متعددی برای استاندارد کردن این نوع مکانیسمهای رسمی ارائه شده اند یکی از آنها که در حال حاضر مورد استفاده قرار می گیرد استاندارد باز RSS است:

RSS مخفف خلاصه سایت RDF سایت مختصر غنی یا «پیوند واقعاً ساده» است. در ابتدا، تصور می شد که این توسعه RDF برای خدمات مای نت اسکیپ(3) است ،اما از آن امروزه در وب نوشتها یا سایتهای خبری برای ارائه سیاهه کوتاه از آخرین اخبار و / یا روز آمد سازی سایتها، به طور وگسترده استفاده میشود چنین استانداردی را میتوان برای تولید دورهای فایل RDF حاوی URL تاریخ آخرین تغییرات در تمام صفحه های سایت مورد استفاده قرار داد پس از آن، خزشگر می تواند برای اولین بار فایل را بررسی و از آن برای خزش در سایت استفاده کند یا آن را با سیاههای موجود خود مقایسه کند اگر صفحه های سایت قبلاً خزش شدهاند تنها صفحه های تغییر یافته را واکشی کند کاستیلو(4) (2004 ، ص 109) این نوع پیاده سازی را ارائه کرده است.

چنین مکانیسمی می تواند برای سیاهه و آشکار کردن هر نوع صفحه ای که به وبگاههای دیگر (و نه به صفحه های وب پنهان) پیوند داده شدهاند مورد استفاده قرار گیرد.

1 – 2 سطح پیمان نامه

گامی به جلو را میتوان با پیمان نامه های ارتباطی اختصاصی برداشت مانند پیمان نامه برداشت طرح ابر داده بایگانی آزاد(5) (OAL–MHP) که توسط لاگوز(6) و ون دی سامپل(7) در سال 2001 پیشنهاد شد و با استفاده

1– Boston

2– Brandman

3– MyNetscape

4– Castillo

5– Open Archive Initiative Metadata Harvesting Protocol

از ترکیب نحوی(1) XML به آشکارسازی ابر داده بر روی HTTP می پردازد. پیاده سازی در سطح پیمان نامه، باعث ارتباط واقعی از طریق درخواست و پاسخ سرور میشود و از این طریق احتمالات افزایش می- یابد به عنوان مثال امکان پرس و جو از سند به کمک تاریخ و نوع اسناد فراهم می شود. خزشگر می تواند به طور مستقیم با سرور OAI ارتباط برقرار کند تا سیاهه ای از اسناد مرتبط با ابر داده (فایل) را به دست آورد. اگر انباره سازگار OAI دسترسی نامحدود به اسناد را فراهم کند این امکان به وجود می آید که آنها را واکشی و با ابرداده خود در بایگانی ذخیره کند برخی خدمات دروازه واسطه نیز برای خزشگرهایی اجرا شده است که قادر به استفاده از پیماننامه OAI برای ایجاد صفحه های پیوند به تمامی اسناد از سرور OAI نیستند (لیو(2) و دیگران2002)

حتی اگر این نوع مکانیسم همکاری در جای خود استفاده شود باید آن را در بیشتر موارد به عنوان روش مکمل جمع آوری محتوا برای بایگانی سازمان در نظر گرفت همانطور که سایتهای تولید کننده معمولاً از آنها برای کاهش بار وارده بر سرور خود - با هدف قرار دادن خزشگر موتور کاوش برای بازدید صفحه های روزآمد شده - استفاده میکنند در حالی که نمایه سازی خزشگر بر روی سطح سایتها باقی میماند و نتایج را روز آمد میکند بایگانی خزشگر نیازمند یکپارچگی و تکامل است.

دروازه بایگانی مستند

عکس

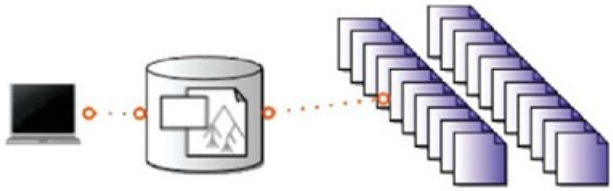
از ترکیب نحوی XML[1] به آشکارسازی ابرداده بر روی HTTP می‌پردازد. پیاده‌سازی در سطح پیمان‌نامه، باعث ارتباط واقعی از طریق درخواست و پاسخ سرورمی‌شود، و از این طریق احتمالات افزایش می‌یابد. به عنوان مثال، امکان پرس‌وجو از سند به کمک تاریخ و نوع اسناد فراهم می‌شود. خزشگر می‌تواند به‌طور مستقیم با سرور OAI ارتباط برقرار کند تا سیاهه‌ای از اسناد مرتبط با ابرداده (فایل) را به‌دست آورد. اگر انباره سازگار OAI دسترسی نامحدود به اسناد را فراهم کند این امکان به‌وجود می‌آید که آنها را واکشی و با ابرداده خود در بایگانی ذخیره کند. برخی خدمات دروازه واسطه نیز برای خزشگرهایی اجرا شده است که قادر به استفاده از پیماننامه OAI برای ایجاد صفحه‌های پیوند به تمامی اسناد از سرور OAI نیستند (لیو[2] و دیگران ۲۰۰۲).

حتی اگر این نوع مکانیسم همکاری در جای خود استفاده شود باید آن را در بیشتر موارد به عنوان روش مکملِ جمع‌آوری محتوا برای بایگانی سازمان در نظر گرفت، همانطور که سایت‌های تولیدکننده معمولاً از آنها برای کاهش بار وارده بر سرور خود ـ با هدف قرار دادن خزشگر موتور کاوش برای بازدید صفحه‌های روزآمد شده ـ استفاده می‌کنند. در حالی که نمایه‌سازی خزشگر بر روی سطح سایت‌ها باقی می‌ماند و نتایج را روز آمد می‌کند. بایگانی خزشگر نیازمند یکپارچگی و تکامل است.

دروازه بایگانی مستند

در برخی موارد، روش‌های قبلی را نمی‌توان به کار برد، زیرا به از دست‌دادن غنا و ساختار ابرداده منجر می‌شوند، به‌خصوص درمورد اسنادی که نمی‌توانند یک پیوند ساده را ترسیم کنند. به عنوان مثال، مجموعه‌ای از تصاویر علمی را تصور کنید: بایگانی تمام تصاویر بدون ابرداده می‌تواند بی‌فایده باشد. در این موارد، ابرداده مرتبط با تصاویر به اندازه خود تصاویر مهم هستند.

گزینه جایگزین برای جلوگیری از اجرای پیمان‌نامه جدید توسط سرور، استخراج مستقیم ابرداده از بانک اطلاعاتی و بایگانی آن همراه با اسناد در یک فرمت آزاد است (شکل ۳). کتابخانه ملی فرانسه در سال ۲۰۰۲ ، این روش را با موفقیت در چند وبگاه پنهان به‌کار برده است. این روش همکاری تولید کننده را می‌طلبد ودر مقایسه با پیاده‌سازی یک سرویس جدید خواهان کمتری دارد، زیرا نیاز به استخراج دارد. مشکل اصلی ناشی از ناهمگونی نظام‌های پایگاه داده، طرح پایگاه داده، و طرح نگاشت (ترسیم) اشیاست.



شکل۳

1. fv'lkXML
2. Liu

در برخی موارد روشهای قبلی را نمیتوان به کار ،برد زیرا به از دست دادن غنا و ساختار ابر داده منجر می شوند به خصوص در مورد اسنادی که نمیتواننـد یک پیونـد ساده را ترسـیم کنند. به عنوان مثال، مجموعه ای از تصاویر علمی را تصور کنید بایگانی تمام تصاویر بدون ابر داده میتواند بی فایده باشد. در این موارد ابر داده مرتبط با تصاویر به اندازه خود تصاویر مهم هستند.

گزینه جایگزین برای جلوگیری از اجرای پیمان نامه جدید توسط سـرور اسـتخراج مسـتقیم ابر داده از بانک اطلاعاتی و بایگانی آن همراه با اسـناد در یک فرمت آزاد است شکل (3) کتابخـانه ملی فرانسه در سال 2002 ، این روش را بـا موفقیت در چنـد وبگاه پنهان به کار برده است. این روش همکاری تولید کننده را می طلبد و در مقایسه با پیاده سازی یک سـرویس جدید خواهان کمتری دارد زیرا نیاز به استخراج

دارد. مشکل اصلی ناشی از ناهمگونی نظامهای پایگاه داده طرح پایگاه داده و طرح نگاشت (ترسیم) اشیاست.

ص: 101

---

fv'lkXML -1

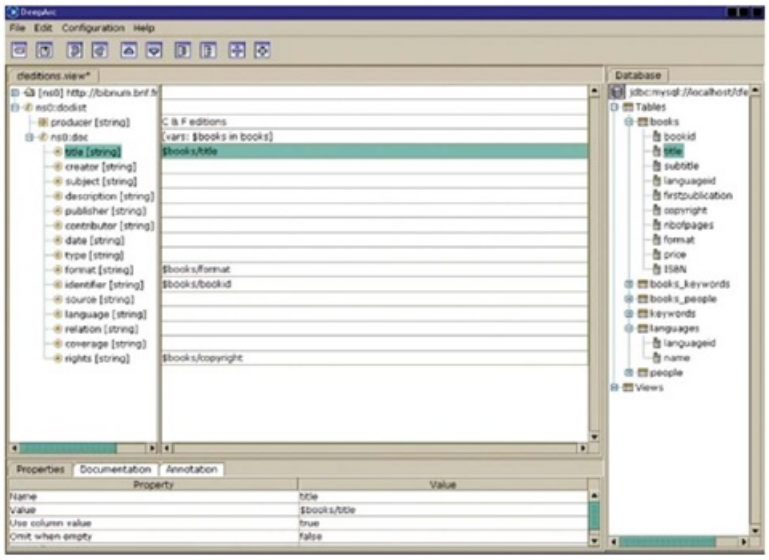
Liu -2

به منظور کاهش در دو مورد ،اول کتابخانه ملی فرانسه نوعی استخراج کننده پایگاه داده گرافیکی منبع آزاد را توسعه داده است که میتواند در چندین نظام پایگاه داده اجرا شود.(1) سپس، ابزار می تواند محتوای پایگاه داده را به اسناد XML مطابق با طرح انتخاب شده توسط بایگانی صادر کند. تولید کننده، که به خوبی طرح داخلی خود را میشناسد می تواند نقشه گرافیکی آن را با هدف طرح ارائه شده توسط بایگانی ترسیم کند (شکل4).

شکل 4

عکس

به منظور کاهش در دو مورد اول، کتابخانه ملی فرانسه نوعی استخراج کننده پایگاه داده گرافیکی منبع آزاد را توسعه داده است که می‌تواند در چندین نظام پایگاه داده اجرا شود[1]. سپس، ابزار می‌تواند محتوای پایگاه داده را به اسناد XML، مطابق با طرح انتخاب شده توسط بایگانی صادر کند. تولید کننده، که به خوبی طرح داخلی خود را می‌شناسد، می‌تواند نقشه گرافیکی آن را با هدف طرح ارائه شده توسط بایگانی ترسیم کند (شکل ٤).



شکل ٤

به‌عنوان مثال، اگر تولید کننده «AUT» را به عنوان نام فیلد در پایگاه داده نویسنده به‌کار برده باشد می‌تواند با کمک ویزارد این دو را به‌صورت گرافیکی رسم کند. ویزارد، همچنین، تولید کننده را قادر می‌سازد تا اطلاعات خاصی را که به‌دلیل حفظ حریم خصوصی نیاز به بایگانی شدن ندارند، فیلتر کند. تمام این مراحل را می‌توان در مدت زمان محدود انجام داد و به تولید خروجی نسخه XML ساختار پایگاه داده پرداخت که می‌تواند با اسناد شرح داده شده توسط این ابرداده صادر و حفظ شود.

مشکل اصلی این روش نامگذاری و پیوند با طرح و چگونگی ترجمه در محیط بایگانی است. شناسایی اشیا در پایگاه داده می‌تواند برای پیوند با داده در محیط بایگانی پیچیده و دشوار باشد. اگر از نوعی ساختار راهنما برای سازماندهی اشیا استفاده شده باشد و یا یک اسکریپت در محیط اصلی راه را

1. DeepArc, an opensource database extractor. http://bibnum.bnf.fr/downloads/deeparc/ (last visited May 2006).

به عنوان مثال، اگر تولید کننده AUT» را به عنوان نام فیلد در پایگاه داده نویسنده به کار برده باشد میتواند با کمک ویزارد این دو را به صورت گرافیکی رسم .کند ،ویزارد ،همچنین تولید کننده را قادر می سازد تا اطلاعات خاصی را که به دلیل حفظ حریم خصوصی نیاز به بایگانی شدن ندارند فیلتر کند. تمام این مراحل را میتوان در مدت زمان محدود انجام داد و به تولید خروجی نسخه XML ساختار پایگاه داده پرداخت که میتواند با اسناد شرح داده شده توسط این ابر داده صادر و حفظ شود.

مشکل اصلی این روش نامگذاری و پیوند با طرح و چگونگی ترجمه در محیط بایگانی است. شناسایی اشیا در پایگاه داده میتواند برای پیوند با داده در محیط بایگانی پیچیده و دشوار باشد اگر از نوعی ساختار راهنما برای سازماندهی اشیا استفاده شده باشد و یا یک اسکریپت

در محیط اصلی راه را

ص: 102

DeepArc, an opensource database extractor.http://bibnum.bnf.fr/downloads/deeparc/ (last visited May – 1 2006).

برای اشـیا ایجاد کند بایگانی باید طرح پیوند خود را بسازد که با ساختار و طرح نامگذاری سازگار است. از آنجا که مکانیسم پیوند اصلی را میتوان تعریف کرد روش دیگری برای به کارگیری وجود ندارد؛ روش مورد به مورد برای ایجاد یک بایگانی کارآمد امری ضروری است.

این بایگانی باید فرمت HTML خود را برای پرس و جوی XML ابر داده بایگانی شده ایجاد کند و به مجموعه ای از اشیا پیوند داده شود.

توجه داشـته باشید که ابر داده را میتوان در بایگانی سیستمهای سنتی مدیریت پایگاه داده رابطه ای به راحتی تزریق .کرد نکته مهم این است که نسخه XML از ابر داده اصلی در پایگاه داده به منظور اطمینان از اینکه در آینده قابل خواندن خواهد بود باقی میماند و محافظت خواهد شد.

نتیجه گیری

همانطور که ملاحظه کردیم بایگانی وب پنهان سـختتر از بایگانی وب سـطحی است با اینکه برخی از روشها به موفقیت رسیده اند تا زمان نوشـتن این مطلب هیچ یک از آنها را نمی توان به عنوان روش مناسب در نظر گرفت این امر نیازمند پیشرفتهایی برای حفاظت از وب پنهان از طریق راههای ساده است و البته به تکامل فنی وب نیز بسـتگی دارد اما حداقل دو دلیل وجود دارد که این بخش از وب مورد غفلت واقع نشود نخست اینکه وب گستره وسـیعی دارد و دارای منابع ارزشـمندی است که بسـیاری از مؤسـسه های میراث فرهنگی به آن علاقه مندند دوم ،اینکه احتمالاً وب با معماریهایی از اطلاعات تکامل می یابد که در برابر شـیوههای سـنتی خزشگر مقاومت می.کنند نخستین عامل ضـد تعـادلی فشاری است که موتور کاوش در سایت قرار میدهـد برای نمایه شـدن بایـد خزش شونـد. اما اگر هماهنگیهای مسـتقیم و به روزرسانی دو طرفه و وجود داشته باشد ممکن است تغییر کند که دلیل خوبی برای ادامه کار و دیدگاهی محافظت گرا می باشد.

منابع

Adams, K. C. (2001). The Web as Database: New Extraction Technologies and Content

Management. Online, March

Agichtein, E., Ipeirotis, P. G., Gravano, L. (2003). Modeling Query-Based Access to Text

Databases

Barbosa, L. Freire, J. (2004). Siphoning Hidden-Web Data through Keyword-Based

Interfaces. Paper presented at the SBBD

Bergman, M. I. K. (2001). The Deep Web: Surfacing Hidden Value. The Journal of Electronic

(Publishing, 7(1

Boston, T. (2005). Exposing the deep web to increase access to library collections. Paper

,presented at the AusWeb05. The Twelfth Australasian World Wide Web Conference

Queensland, Australia

Boufkhad, Y. Viennot, L. (2003). The Observable Web. RR Boyko, A. (2004). Test Bed

Taxonomy. IIPC Reports, 16

Brandman, O., Cho, J., Garcia-Molina, H., Shivakumar, N. (2000). Crawler- Friendly Web

Servers. SIGMETRICS Performance Evaluation Review, 28(2), 9-14

Callan, J. Connell, M. (2001). Query-based sampling of text databases. ACM Transactions

on Information Systems 19(2), 97-130

Castillo, C. (2004). Effective Web Crawling. University of Chile

Chang, K. C.-C., He, B., Li, C., Patel, M., Zhang, Z. (2004). Structured databases on the

web: observations and implications. SIGMOD Records, 33(3), 61-70

Cope, J., Craswell, N., Hawking, D. (2003). Automated discovery of search interfaces

on the web. Paper presented at the Proceedings of the Fourteenth Australasian Database

Conference on Database Technologies 2003

Florescu, D., Levy, A., Mendelzon, A. (1998). Database techniques for the World-Wide

Web: A survey. SIGMOD Records, 27, 59-74

.Frankewitsch, T. Prokosch, U. (2001). Navigation in medical Internet image databases

Medical Informatics and the Internet in Medicine, 26(1), 1-15 5 Archiving the Hidden

Web 129

Gravano, L., Ipeirotis, P. G., Sahami, M. (2003). QProber: A System for Automatic

(Classification of Hidden-Web Databases. ACM Transactions on Information Systems, 21(1

He, H., Meng, W., Yu, C., Wu, Z. (2005). WISE-Integrator: a system for extracting and integrating complex web search interfaces of the deep web. Trondheim, Norway

Hearst, M. (1998). Information Integration. IEEE Intelligent Systems, 13(5), 12-24

/HTTrack. http://www.httrack.com

Lage, J. P., Silva, A. S. D., Golgher, P. B., Laender, A. H. F. (2002). Collecting hidden Web pages for data extraction. Paper presented at the Proceedings of the fourth international workshop on Web information and data management low-barrier interoperability framework. Roanoke, Virginia, United States

,Lawrence, S. Giles, C. L. (1999). Accessibility of Information on the Web. Nature, 400 107-109

Liddle, W. S., Yau, S. H., Embley, D. W. (2002). On the Automatic Extraction of Data from

,.the Hidden Web. Springer, Berlin Heidelberg New York Liu, X., Maly, K., Zubair, M

Nelson, M. (2002). DP9 – an OAI gateway service for Web crawlers. Paper presented at the Second ACM/IEEE Joint Conference on Digital Libraries mation Mediation. Paper presented at the Intl. Workshop on the World–Wide Web and Conceptual Modeling WWWCM'99), Paris)

Marill, J., Boyko, A., Ashenfelder, M. (2004). Web Harvesting Survey, 10

Masanè s, J. (2002). Archiving the deep web. Paper presented at the 2nd International Workshop on Web Archives (IWAW'02), Roma, Italy

Mohr, G., Kimpton, M., Stack, M., Ranitovic, I. (2004). Introduction to Heritrix, an archival quality web crawler. Paper presented at the 4th International Web Archiving Workshop (IWAW'04), Bath, UK

Ntoulas, A., Zerfos, P., Cho, J. (2005). Downloading textual hidden web content through keyword queries. Denver, CO, USA sented at the Proceedings of the 27th International Conference on Very Large Data Bases

.Roche, X. (2006). Copying web sites. In J. Masanè s (Ed.), Web Archiving Springer, Berlin Heidelberg New York integration of data and its representation. Paper ,presented at the 1st International Workshop on Web Site Evolution (WSE'99), Atlanta USA

–Zhang, Z., He, B., Chang, K. C.–C. (2004). Understanding Web query interfaces: Best effort parsing with hidden syntax. Paper presented at the Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data

,Lagoze, C. Van de Sompel, H. (2001). The open archives initiative: building a Ludäscher

.B. Gupta, A. (1999). Modeling Interactive Web Sources for Infor–Raghavan, S

,Garcia–Molina, H. (2001). Crawling the Hidden Web. Paper pre– Storey, M.–A. Jahnke

J. H. (1999). Web site evolution – Towards a flexible

چکیده

پژوهش حاضر پژوهشی مفهومی است که با هدف تبیین تاثیر بستر نحوی بر میانکنش پذیری استانداردهای فراداده ای انجام شد جامعه پژوهش عبارت بودند از استانداردهای فراداده ای مارک 21 در بستر زبان نشانه گذاری گسترش پذیر مارک ایکس ام ال استاندارد انتقال و کدگذاری فراداده ها (متس)"، "طرح فراداده ای توصیف شیء (مودس) طرح فراداده ای توصیف مستند مدس)"، طرح فراداده ای هسته دوبلین (دی) سی ام آی فراداده برای نگهداری اشیای

آی،دیجیتالی (پریمیس)، فراداده فنی برای اشیای دیجیتالی متنی تکست ام. دی. فراداده برای مدیریت تصاویر در بستر ایکس ام ال (میکس)؛ و گردآوری داده ها با روش کتابخانه ای صورت گرفته است. در بخش نخست انواع و ابزارهای میانکنش پذیری استانداردهای فراداده ای توصیف گردید در بخش دیگر با مبنا قرار دادن استاندارد انتقال و کدگذاری فرادادهها (متس) به عنوان استاندارد هسته، نحوه تعامل استانداردهای فراداده ای مورد مطالعه با یکدیگر و با استاندارد متس با رویکرد تحلیلی-سیستمی مورد بررسی قرار گرفته و الگوهایی متناسب با هر یک ترسیم شده اند تحلیل صورت گرفته بیانگر آن است که استفاده از بستر نحوی مناسب در میانکنش پذیری استانداردهای فراداده ای نقش به سزایی ایفا میکند و به یکپارچه سازی درونی و برونی نظامهای اطلاعاتی می انجامد.

کلیدواژه ها بستر ،نحوی استانداردهای فراداده ای میانکنش پذیری یکپارچه سازی نظامهای اطلاعاتی

بررسی تاثیر بستر نحوی بر میانکنش پذیری استانداردهای فراداده ای گامی در راستای یکپارچه سازی نظامهای اطلاعاتی

سید مهدی طاهری(1)

مقدمه

مفهوم میانکنش پـذیری بر قابلیت تعامل و کار متقابل میان چنـد نظام اطلاعاتی با هدف تبادل داده ها و خدمات دلالت دارد اجرای فرایند میانکنش پذیری در راسـتای یکپارچه سازی درونی و برونی نظام اطلاعاتی با اجزای درونی خود و دیگر نظامهای اطلاعاتی صورت میگیرد و منجر به ارزش افزوده برای نظامهای موجود در فراینـد میشود این تعامل در دو سـطح نحوی و معنایی رخ میدهد در سـطح نحوی، تبادل داده ها بر اساس قالبهای مشترک و یا اسـتفاده از پروتکلهای ارتباطی، و در سـطح معنایی تفسیر دادههای مبادله شده به صورت معنادار به منظور تولید نتایج مفید همخوان با نیازها و سطح شناختی کاربران مد نظر است از آنجا که صفات و ویژگیهای هر شیء محتوایی (ورودی) در قـالب اسـتانداردها و طرحهای فراداده ای به صورت معنادار توصـیف شـده (پردازش)، و در قالب محصولی جدیـد به نام پیشـینه های فراداده ای بـازنمون می گردنـد فراداده نیز یـک نظام اطلاعاتی به شـمار میآیـد. بنابراین هماننـد دیگر نظامهای اطلاعاتی نیاز به تعامل میان نظامهای فراداده ای برای

1- دکترای علم اطلاعات و دانش شناسی و استادیار پژوهشگاه علوم و فرهنگ اسلامی taherismster@gmail.com ،

نیل به اهداف فرایند میانکنش پذیری بدیهی مینماید، و به "میانکنش پذیری فراداده ای(1)" که نوعی میانکنش پذیری معنایی ،است، شهرت یافته است.

به عبارت دیگر با توجه به حجم فراوان اشیای محتوایی منتشر شده در هر یک از حوزه های دانش ،بشری و تنوع خدماتی که به کمک پیشرفتهای حوزه فناوریهای اطلاعاتی و ارتباطی برای ارائه این اشیاء امکان پذیر گردیده است و نیز پشتیبانی هر یک از استانداردهای فراداده ای از کارکرد هایی ،خاص بهره مندی از طیفی از استانداردهای فراداده ای برای مدیریت اشیای محتوایی و خدمات ارائه شده در نظامهای اطلاعاتی مورد نیاز است و میانکنش پذیری این استانداردها به منظور یکپارچه سازی اجزا و فرایندهای نظام اطلاعاتی ضروری است در سالهای اخیر تلاشهای نظری و کاربردی گسترده ای برای انجام و تسهیل فرایند میانکنش پذیری نظامهای اطلاعاتی به ویژه نظامهای فراداده ای صورت گرفته است. برگزاری همایشهای متعدد با دامنه بین المللی برای تقویت ادبیات این موضوع(2)، و طراحی پروفایلهای کاربردی(3)، عناصر ارتباطی(4)، جداول یا گذرگاههای تطبیقی(5) و نظیر آنها برای اجرای فرایند میان کنش پذیری نشان دهنده اهمیت این موضوع است.

از سوی دیگر فراداده برای بازنمون خود نیاز به بستر نحوی(6) دارد ،یعنی ماشین – خوان و ماشین – فهم شدن فراداده منوط به استفاده از بستر نحوی مناسب است. با توجه به این مهم مسئله ای که در اینجا مطرح میگردد آن است که آیا بستر نحوی بر میانکنش پذیری استانداردهای فراداده ای موثر است؟ آیا تعامل میان استانداردهای فراداده ای در محیط این بستر صورت میگیرد؟ انتخاب بسترهای نحوی گوناگون فرایند میانکنش پذیری را تغییر خواهد داد؟ و در پایان این بستر میتواند زمینه را برای مدیریت بهینه فراداده ها و به پیروی از آن اشیای محتوایی به عنوان هدف اصلی یکپارچگی نظام اطلاعاتی فراهم نماید؟

روش شناسی

این پژوهش یک پژوهش مفهومی است که به تبیین تاثیر بستر نحوی بر میانکنش پذیری استانداردهای فراداده ای می پردازد. جامعه پژوهش را استانداردهای فراداده ای قالب" فرادادهای مارک 21 در بستر زبان نشانه گذاری گسترش پذیر (مارک ایکس ام ال)(7) استاندارد انتقال و کدگذاری فرادادهها متس(8)

1- Metadata interoperability

2- از جمله همایش- Metadata harmonization: Bridging languages of description: http://www.digcur-education.org/eng/Events/ Metadata-Harmonization-Bridging-Languages-of-Description

3- Application profiles

4- Linking Devices

5- Crosswalks or Mapping table

6- بستر نحوی عبارت است از مجموعه ای از ،قواعد دستورالعملها و نشانهها برای اعمال ساختاری خاص بر روی محتوای متنی به منظور ذخیره سازی فهم و انجام پردازشهای خاص توسط ماشین (رایانه)

7- (Machine-readable Cataloguing in XML (MARCXML

(Metadata Encoding and Transmission Standard (METS –8

،طرح فرادادهای توصیف شیء (مودس)(1)، طرح فرادادهای توصیف مستند (مدس)(2)، طرح فراداده ای هسته دوبلین (دی سی ام آی)(3) فراداده برای نگهداری اشیای دیجیتالی (پریمیس)(4)، فراداده فنی برای اشیای دیجیتالی متنی (تکست .ام دی)(5) و فراداده برای مدیریت تصاویر در بستر ایکس ام ال (میکس)(6) تشکیل میدهند. در بخش نخست ،پژوهش انواع ابزارهای مورد استفاده برای فرایند میانکنش پذیری با تاکید بر نوع میانکنش پذیری که میان استانداردهای فراداده ای ایجاد مینمایند، توصیف میشوند تبیین تاثیر بستر نحوی بر میانکنش پذیری استانداردهای فراداده ای بخش بعدی و اصلی پژوهش است در این بخش استاندارد" انتقال و کدگذاری فرادادهها متس به دلیل قابلیت مدیریت فراداده ها و امکان جاسازی دیگر استانداردهای فراداده ای درون آن به عنوان استاندارد هسته مد نظر قرار گرفته، و تعامل دیگر استانداردها با یکدیگر و با این استاندارد بر اساس بخشهای هفتگانه آن با رویکرد تحلیلی بررسی میشود همچنین عناصر ارتباطی مورد استفاده برای برقراری تعامل میان استانداردها بر پایه توصیه های ارائه شده از سوی استانداردهای مورد مطالعه تعیین می.شوند برای گردآوری دادهها از روش کتابخانهای اسنادی) استفاده شده است و الگوهای تعاملی ارائه شده بر مبنای رویکرد تحلیلی سیستمی طراحی گردیده اند.

انواع و ابزارهای میانکنش پذیری استانداردهای فراداده ای

همان طور که پیشتر اشاره شد میانکنش پذیری میان نظامهای اطلاعاتی موجب یکپارچه سازی درونی و برونی آنها شده ارزشهای افزوده فراوانی برای این نظامها به ارمغان می آورد. میانکنش پذیری فراداده ای عبارت است از توانایی ،نظامها ،خدمات و سازمانها در تعامل با یکدیگر، تبادل دادهها، از داده های مبادله شده بدون نیاز به تلاشی خاص بر روی نظام مبدأ این فرایند در سه سطح انجام می شود نخست سطح فرانماها که در آن سطح عناصر فراداده ای مد نظر قرار می گیرند، و از محیط فنی شبکه ای سخت افزاری و نرم افزاری مستقل هستند محصول این سطح از فرایند، مجموعه ای از عناصر استخراج شده، گذرگاههای تطبیقی پروفایلهای کاربردی ثبت های فراداده ای(7) هستند. سطح دیگر، به میانکنش پذیری پیشینههای فراداده ای اختصاص دارد در این سطح یکپارچه سازی پیشینههای فراداده ای از طریق همخوانی عناصر از بعد معناشناختی صورت میگیرد پیشینههای تبدیل شده و تولید پیشینههای جدید با ترکیب ارزشهای عناصر پیشینههای موجود خروجی سطح پیشینهها به شمار می آیند و در سطح دو دیگر که به سطح مخازن اطلاعاتی موسوم است رشته های ارزشهای برخی عناصر خاص با گردآوری فرادادهها از نظامهای مختلف و یکپارچه نمودن آنها همایند میشوند این سطح امکان جستجوی یکپارچه

1- Metadata Object Description Schema (MODS)

2- Metadata Authority Description Schema (MADS)

3- Dublin Core Metadata Initiative (DCMI)

4- PREservation Metadata: Implementation Strategy (PREMIS)

5- Technical Metadata for Text (TextMD)

6- Metadata for Images in XML (MIX)

7- Metadata registries

میان چند نظام اطلاعاتی را فراهم مینماید (معارف(1) و یحیی(2)، 2009؛ هیروید(3)، 2011).

برای اجرای فرایند میانکنش پذیری فراداده ای ابزارهای گوناگونی طراحی شده است. پروفایلهای کاربردی عناصر ،ارتباطی گذرگاهها یا جداول تطبیقی و بستر ،نحوی ابزارهای فرایند میانکنش پذیری فراداده ای محسوب میشوند. پروفایلهای کاربردی مجموعه عناصر فراداده ای (استخراج شده از یک یا چند استاندارد فراداده ای) خط مشی،ها تجربیات برتر و رهنمودهایی که به منظور کاربردهای خاص (محلی) تعریف شده است ،یا اعلان ضوابطی که یک سازمان یک منبع اطلاعاتی، یک برنامه کاربردی، یا جامعه استفاده کنندگان در به کارگیری فراداده هایشان استفاده می کنند (طرح فرداده ای هسته دوبلین 2013)و میانکنش پذیری سطح فرانماها را پشتیبانی میکنند عناصر ارتباطی به صفات یا خصایص اشیای محتوایی مانند موضوع پدیدآورنده ،ناشر و مانند آن گفته می شود که ارتباط میان چند شیء محتوایی را برقرار مینمایند و به میانکنش پذیری سطح پیشینه ها و نیز مخازن سازمانی می انجامد جداول یا گذرگاههای تطبیقی به جداولی اطلاق میشود که عناصر معادل در بیش از یک استاندارد فراداده ای نشان میدهند، و همانند پروفایلهای کاربردی موجب میانکنش پذیری در سطح فرانماها میشوند.

و اما بستر ،نحوی میانکنش پذیری استانداردهای فراداده ای در سطح فرانماها انجام میشود. هر استاندارد فراداده ای دارای فرانمایی ویژه است که میزان سازگاری پیشینههای تولید شده بر مبنای آن استاندارد را اعتبار سنجی مینماید استانداردهای فراداده ای مجموعه ای از عناصر مرتبط و ساختارمند از لحاظ معناشناختی هستند که برای پشتیبانی از کارکردهایی خاص و متناسب با نیازهای جامعه کاربران خود طراحی شدهاند (فتاحی و طاهری،1388) این استانداردها برای پیاده سازی پیشینههای مبتنی بر خود یک یا چند قالب ذخیره سازی و نمایش دادهها را به عنوان بستر نحوی بر میگزینند. طیف گسترده ای از قالبهای ذخیره سازی وجود دارد که برخی مبتنی بر پایگاه داده ها(4) و برخی مبتنی بر فایل(5) هستند.

مهمترین این قالبها ،زبانهای نشانه گذاری(اس .جی. ام. ال اچ تی ام ال.، و ایکس. ام. ال.)، قالب مدارک قابل انتقال پی دی اف با استفاده از چارچوب توصیف منبع(6)، قالب متن تکست)، و قالب بومی نظامهای مدیریت پایگاههای داده ای دی .بی.ام .اس میباشند طاهری، 1391). هر یک از این قالبها قابلیتهای خاصی برای ذخیره سازی و نمایش دادهها دارند و بر پایه مقاصد خاصی تولید شده اند بنابراین انتخاب آنها از سوی استانداردهای فراداده ای میبایست سازگار با کارکردهای خاص آنها باشد. افزون بر این، به دلیل آن که نظامهای اطلاعاتی برای مدیریت محتوا و خدمات خود از چند استاندارد فراداده ای به طور همزمان استفاده میکنند تعامل میان این استانداردها در جهت نیل به اهداف نظام ضروری است. از این رو این ویژگی نیز در انتخاب بستر نحوی حائز اهمیت است.

1- Maarof

2- Yahya

3- Hirwade

4- Database-based format

5- File-based format

6- Resource Description Framework (RDF)

زبـان نشـانه گـذاری فرامتن (اچ تی ام ال) قـالبی برای توصـیف ساختار صـفحات وب به منظور نمایش آنهاست از مهمترین قابلیتهای این ،زبان امکان استفاده از فناوری فراپیوند، و ذخیره داده های چند رسانه ای است اما در طراحی این قالب انتقال دادهها مورد اقبال نبوده است به همین دلیل تعداد برچسـبها و فرابر چسـبهای آن محدود و از پیش تعریف شده هسـتند و نمیتوان آنها را گسترش داد. توصـیف دادههای ذخیره شده در این قالب به ویژگیهای نرم افزاری نظام اطلاعاتی که از اچ تی ام ال بهره میبرد وابسـته است. این ویژگی تعامل استانداردهای فراداده ای که پیاده سازی پیشـینههای خود در بستر این زبان را توصـیه میکنند محدود مینماید (کنسرسیوم وب جهانی،2012). کنسرسیوم قالب مـدارک قابل انتقال (پی دی اف) برای بازنمون اشـیای محتوایی به صورت مسـتقل از سـخت افزار نرم افزار و نظام عامل طراحی شده است. هنگامی که حفظ ویژگیهای صـفحه آرایی یک شـیء دیجیتالی ذخیره شـده در قالب الکترونیکی دیگر و یا یک شـیء آنالوگ مطرح باشـد، از قالب پی دی. .اف. اسـتفاده میشود(ویکی پدیا2013) بنـابراین یکی از بهترین قالبهـا برای تهیه نسـخه چاپی از شـیء دیجیتالی .است اگر چه این قالب به دلیل حفظ ویژگیهای صـفحه آرایی مستقل از پلت فرم میباشد اما دادههای ذخیره شده در آن به صورت معنادار توصـیف نمی شوند و صـرفاً تصویری از شیء تبدیل شده به قالب پی دی. اف. قلمداد میشوند به عبارت دیگر در انتقال دادهها از نظامی به نظام دیگر ساختار دادههای ذخیره شـده در آن چنـدان قابلیت پردازشـی نـدارد و هـدف اصـلی آن هماننـد قالب اچ تی ام ال نمایش داده هاست از آنجا که در پیشـینههای فراداده توصیف معنادار عناصر و روابط میان آنها از اهمیت فراوانی برخوردار است این قالب چندان مورد توجه بافت فراداده ای واقع نشد.

قالب متن ، (تکست) برای ذخیره سازی دادهها بدون اسـتفاده از نشانه ها یا اعمال ساختاری خاص طراحی شده است. دادهها ذخیره شده در این قالب به دلیل عدم وجود نشانه های اضافی در آن حجم بسـیاری کمی را اشـغال می.کنند در برخی از مواقع با افزودن نشانههایی به دادههای ذخیره شده در این ،قالب میتوان پردازشهای خاصی بر روی آن اعمال نمود. مهمترین ضعف این قالب در فرایند میانکنش پذیری ساختارمند نبودن و عدم توصیف دادههای ذخیره شده در آن است.

قـالب بومی نظامهای مـدیریت پایگاههای داده ای (دی بی ام اس) هر نظام مـدیریت پایگاه داده ای از یک قالب محلی و بومی متناسب با ویژگیهـا و قابلیتهـای فنی خود سود میبرد. این قالب بر اساس اهـداف کارکردها و همچنین لحاظ منحصـر به فرد بودن نظام طراحی شـده است. دادههـای ذخیره شـده در قـالب بومی یک دی بی.ام اس در یک نظام مـدیریت پایگاه داده ای دیگر قابل پردازش نیسـتند و حتماً بر روی آنهـا فراینـد تبـدیل به قـالب جدیـد صورت گیرد. این ویژگی، بـا توجه به این که نظامهـای اطلاعـاتی از پلت فرمی خاص و در نتیجه نظامهای مدیریت پایگاه دادهای متفاوت استفاده میکنند باعث عدم استفاده از قالبهای بومی در فرایند تعامل میان نظام شده است.

زبـان نشـانه گـذاری گسترش پـذیر (ایکس ام ال) یـک قـالب سـاده مبتنی بر متن است که به عنوان اسـتانداردی بین المللی برای بـازنمون دادههـای ساختارمنـد نظیر اشـیای محتوایی و تبـادل و اشتراک دادههـا گسترش یافته است (بری و دیگران 2008)دادههـایی که در قالب ایکس ام ال نشانه گذاری می شوند.

به داده هایی ساختارمند ،تبدیل و اشیای محتوایی خود-توصیف(1) بوجود میآورند. این ویژگی موجب استقلال اشیای محتوایی مبتنی بر ایکس ام ال از هر پلت فرمی ،شـده تبادل آنها را میان نظامهای ناهمگن(2) ممکن، و بنابراین میانکنش پذیری نظامهای اطلاعاتی را باعث می گردد. ایکس. ام. ال. همانند اچ تی ام ال یک مجموعه ثابت از برچسـبها نیست با استفاده از این استاندارد، کاربران میتوانند برچسبهای مورد نیاز خود را ،تعریف و در محیطهای اطلاعاتی دیگر اسـتفاده .کننـد قابلیتهای منحصـر به فرد این زبان گرایش طراحان اسـتانداردهای فراداده ای به اسـتفاده از این زبـان به عنوان بسـتر نحوی پیشـینههای فراداده ای را در پی داشـته است افزون بر آن که پیـاده سـازی برخی از استانداردها مانند مارک 21 که تا پیش از این در قالب زبانهای نشانه گذاری ممکن نبود در قالب زبانهای نشانه گذاری با بکارگیری .ایکس ام ال مهیا شـده است(طاهری، 1387ب؛ کین(3)، 2000؛ گیجی(4) و کلی(5)، 2006)، ساختارمنـد بودن و خود توصـیف بودن آن میانکنش پذیری نظامها و استانداردهای فراداده ای را تسهیل نموده است (طاهری، 1391)

در ادامه مقاله و در قالب طراحی الگوهایی ،چند تاثیر بستر نحوی بر فرایند میانکنش پذیری اسـتانداردهای فراداده ای که امکان اسـتفاده از طیفی از استانداردها در یک نظام اطلاعاتی به صورت همزمان را توجیه میکند با رویکرد تحلیلی سیستمی مورد بررسی قرار میگیرد.

تبیین تاثیر بستر نحوی بر میانکنش پذیری استانداردهای فراداده ای

در این بخش از ،مقاله با استفاده از استاندارد انتقال و کدگذاری فرادادهها (متس) به عنوان استاندارد هسته تعامل دیگر استانداردهای فراداده ای با این اسـتاندارد و در هنگام لزوم تعامل دیگر استانداردها با یکدیگر، با ارائه چند الگو نشان داده میشود دلیل انتخاب استاندارد متس به عنوان استاندارد ،هسته کارکرد اصلی آن یعنی مدیریت فرادادههاست (طاهری 1387الف) متس همانند بسته ای عمل مینماید که میتواند دیگر اسـتانداردهای فراداده ای با کارکردهای گوناگون را در بر گرفته به مدیریت یکپارچه اشیای محتوایی بپردازد.(6) استاندارد متس دارای هفت بخش است هر یک از این بخشـها دارای کارکردی خاصـی هسـتند. برخی برای جاسازی طرحهای فراداده ای و برخی برای مدیریت محتوا طراحی شـده انـد. در عین تمامی این بخشـها قابلیت تعامل با یکـدیگر دارند و میانکنش پذیری این بخشـها نیز بر اهمیت متس می افزاید.

1- Self-description

2- Heterogeneous or Disparate systems

3- Qin

4- Gigee

5- Kelly

6- برای اطلاعـات بیشتر در مورد کارکردها و دیگر ویژگیهای اسـتانداردهای فراداده ای به این دو منبع مراجعه کنیـد سـید مهـدی طاهری 1387 طراحی یـک کتابخـانه دیجیتالی اسـتاندارد در مجموعه مقالات نخستین همایش کتابخانه های دیجیتالی به کوشـش شـرکت پارس آذرخش :تهران سبزان سـید رحمت الله فتاحی سـید مهدی طاهری 1388 فهرستنویسـی رایانهای ،مفاهیم شیوهها و کاربرد نرم افزارهای رایانه ای در سازماندهی اطلاعات با همکاری فرشته ناقدی احمدی تهران کتابدار

این بخشـها به ترتیب عبارتند از بخش سرپیشینه(1)، بخش فراداده های توصـیفی بخش فراداده های مـدیریتی، بخش مربوط به فایلها بخش نقشـههای ساختاری بخش پیوندهای ساختاری و بخش رفتارهای شـیء (طاهری 1387 الف؛ دفتر اسـتانداردهای مارک و توسعه ،شبکه C2013.) تعامل هر یک از اسـتانداردها توسط عناصر ارتباطی، و در قالب بخشهای هفتگانه متس صورت می گیرد. هر پیشینه فراداده ای به دو روش با پیشـینه متس تعامل ایجاد میکند نخست روش درونی که پیشینه یاد شده درون پیشینه متس به دو صورت دادههای کدگذاری شده با .ایکس ام ال (توسط برچسب) و دادههای مبتنی بر کدهای دودویی یا متن خام توسط برچسب درج (جاسازی)(2) میشود و دیگر تهیه پیوند توسط یو. آر. آی. یک پیشـینه یا دیگر شناسـگرها (پی یو آر ال ای. آر. کی، و دی. أ. آی.)) از درون عنصر مرتبط متس به پیشینه فراداده ای مبتنی بر اسـتاندارد فراداده ای دیگر لازم به ذکر است امکان درج پیشـینههای بیش از اسـتاندارد در هر بخش از پیشینههای متس وجود دارد. پیشـینههای تولید شده بر مبنای هر اسـتاندارد فراداده ای دارای یک عنصـر ریشه هسـتند این عنصر نقش عنصر ارتباطی را برای ارتباط با پیشینههای متس ایفا می.کند در ادامه شیوه تعامل هر یک از استانداردها و عناصر ارتباطی آنها مورد تحلیل قرار میگیرد.

قالب فرادادهای مارک 21 در بستر زبان نشانه گذاری گسترش پذیر (مارک ایکس ام ال.)

این قالب توسط دفتر استانداردهای مارک و توسعه شبکه(3) کتابخانه کنگره آمریکا به منظور پیاده سازی داده های مارک در بستر ایکس ام ال طراحی شـده است. انعطاف پـذیری و گسترش پـذیری این چارچوب امکان پاسـخگویی به نیازهای گوناگون و خاص کاربران را میسـر ساخته است (دفتر استانداردهای مارک و توسعه شبکه ،a2013). وجود عناصـر متعدد باعث شده قالب مارک از چند کارکرد به صورت موثر پشتیبانی .نماید دو کارکرد مدیریت و توصـیف کارکردهای اصلی قالب مارک هستند. در ذیل نحوه میانکنش پذیری قالب مارک 21 در بستر زبان نشانه گذاری گسترش پذیر بر اساس دو کارکرد مدیریتی و توصیفی با استاندارد متس ترسیم شده است.

به عنوان فراداده توصیفی

عنصر ریشه (record)پیشینه مبتنی بر مارک 21 برای پشتیبانی از کارکرد توصـیفی در بخش فراداده های توصـیفی متس با برچسب بر پایه روش درونی در برچسب ، و پیوند به پیشـینه مارک بر پایه روش برونی در عنصر جاسازی می.شود چنان چه روش درونی مـد نظر ،باشد، دادههای کدگذاری شده در قالب .ایکس ام ال در برچسب ، و داده های در قالب دودویی یا متن خام در برچسب binData قرار میگیرند دیگر استانداردهای فراداده ای با کارکرد توصیفی نیز به همین صورت با استاندارد متس تعامل برقرار میکنند.

1- METS Header

2- Embedding

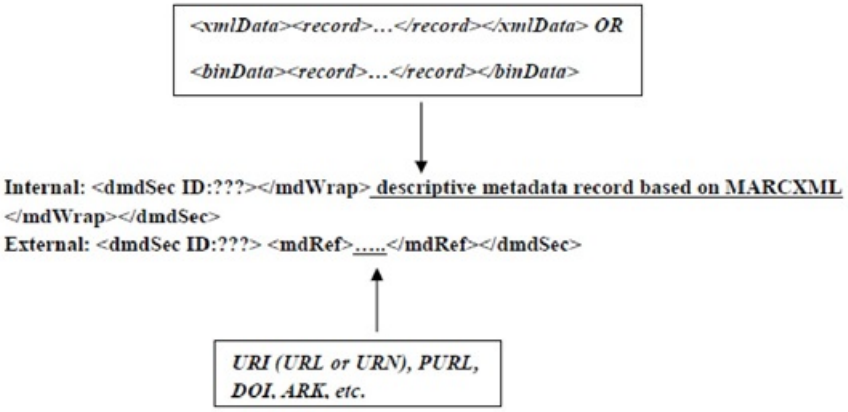
3- Network Development and MARC Standards Office

الگوی .1 شیوه درج پیشینه مارک 21 با کارکرد توصیفی در بخش فرادادههای توصیفی پیشینه متس

الگوی 1 نشـان میدهـد عنصـر ریشه یـک پیشـینه مارک 21 برای پشتیبـانی از کـارکرد توصـیفی قـادر اسـت بـه ترتیـب در عناصـر ??? یا یا <mdWrap:

به عنوان فراداده مدیریتی

عکس



الگوی ۱. شیوه درج پیشینه مارک ۲۱ با کارکرد توصیفی در بخش فراداده‌های توصیفی پیشینه متس

الگوی ۱ نشان می‌دهد، عنصر ریشه یک پیشینه مارک ۲۱ برای پشتیبانی از کارکرد توصیفی قادر است به ترتیب در عناصر <dmdSec ID:???>، </mdWrap> یا <mdRdf>، <xmlData> یا <binData>، و<record> جاسازی شود.

به عنوان فراداده مدیریتی

عنصر ریشه پیشینه مارک ۲۱ با کارکرد مدیریتی در بخش فراداده‌های مدیریتی استاندارد متس (با برچسب <amdSec>)، بر پایه روش درونی در برچسب <mdWrap>، و پیوند به پیشینه مارک بر پایه روش برونی در عنصر <mdRef> درج می‌گردد. پیشینه‌های فراداده‌ای مبتنی بر استانداردهای با کارکرد مدیریتی در عناصر <techMD> (برای استانداردهای فراداده‌ای با کارکرد مدیریت فنی)، <rightsMD> (برای استانداردهای فراداده‌ای با کارکرد مدیریت حقوق معنوی)، <sourceMD> (برای استانداردهای فراداده‌ای با کارکرد مدیریتی و توصیفی مربوط به اشیای آنالوگ)، و <digiprovMD> (برای استانداردهای فراداده‌ای با کارکرد مدیریت اشیای با منشاء دیجیتالی) استاندارد متس جاسازی می‌شوند. دیگر استانداردهای فراداده‌ای با کاکرد مدیریتی با توجه به کارکرد فرعی خاص خود همانند پیشینه‌های مارک ۲۱ در ایکس. ام. ال. با استاندارد متس تعامل پیدا می‌کنند.

عنصر ریشه پیشینه مارک 21 با کارکرد مدیریتی در بخش فرادادههای مدیریتی استاندارد متس (ب) برچسب amdSec) بر پایه روش درونی در برچسب mdWrap> و پیوند به پیشینه مارک بر پایه روش برونی در عنصر درج می.گردد پیشینه های فراداده ای مبتنی بر استانداردهای با کارکرد مدیریتی در عناصر techMD برای استانداردهای فراداده ای با کارکرد مدیریت فنی rights برای استانداردهای فراداده ای با کارکرد مدیریت حقوق معنوی sourceMD> (برای استانداردهای فراداده ای با کارکرد مدیریتی و توصیفی مربوط به اشیای آنالوگ و برای) استانداردهای فرادادهای با کارکرد مدیریت اشیای با منشاء (دیجیتالی استاندارد متس جاسازی می شوند. دیگر استانداردهای فراداده ای با کاکرد مدیریتی با توجه به کارکرد فرعی خاص خود همانند پیشینههای مارک 21 در ایکس ام ال با استاندارد

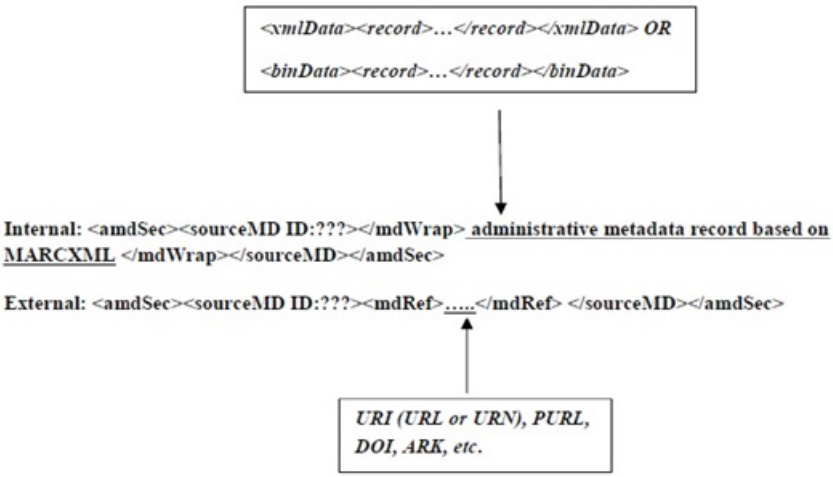
متس تعامل پیدا می کنند.

ص: 114

الگوی 2 شیوه درج پیشینه مارک 21 با کارکرد مدیریتی در بخش فراداده های مدیریتی پیشینه متس

چنـان چه در الگـوی 2 مشاهـده میشـود پیشـینه مبتنی بر مـارک 21 برای ایفـای کـارکرد مـدیریتی میتوانـد به ترتیب درون برچسـبهای ???:KamdSec ID/mdWrap> یا ، یا

عکس



الگوی ۲. شیوه درج پیشینه مارک ۲۱ با کارکرد مدیریتی در بخش فراداده‌های مدیریتی پیشینه متس

چنان چه در الگوی ۲ مشاهده می‌شود، پیشینه مبتنی بر مارک ۲۱ برای ایفای کارکرد مدیریتی، می‌تواند به ترتیب درون برچسب‌های <amdSec ID:???>، </mdWrap> یا <mdRdf>، <xmlData> یا <binData>، و <record> درج شود. **طرح فرادادهای توصیف شیء (مودس)**

این طرح برای مجموعه عناصر کتابشناختی که با اهداف گوناگون، به خصوص کاربردهای کتابخانه‌ای استفاده می شوند، در بستر زبان نشانه‌گذاری گسترش‌پذیر تهیه شده است. مودس امکان انتقال داده های کتابشناختی گزیده از پیشینه‌های موجود مارک و ایجاد پیشینه‌های توصیفی برای اشیای محتوایی جدید را فراهم می‌آورد. طرح فراداده‌ای مودس مجموعه‌ای از عناصر مورد نیاز برای توصیف اشیای دیجیتالی که از فیلدهای مارک ۲۱ استخراج شده است را در بر می‌گیرد. نام‌های برچسب عناصر مودس بر خلاف مارک ۲۱ در قالب ایکس. ام. ال.، مبتنی بر واژگان زبان طبیعی هستند (مک کالم[1]، ۲۰۰٤؛ دفتر استانداردهای مارک و توسعه شبکه، e۲۰۱۳). کارکرد اصلی مودس، کارکرد توصیفی است و پیشینه‌های آن در بخش فراداده‌های توصیفی متس درج می‌شود.

1 . McCallum

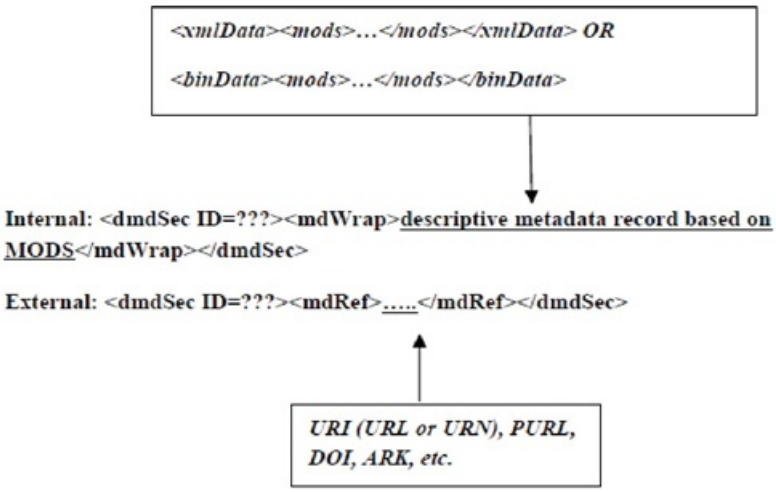
این طرح برای مجموعه عناصر کتابشناختی که با اهداف گوناگون به خصوص کاربردهای کتابخانه- ای استفاده میشوند در بستر زبان نشانه گذاری گسترش پذیر تهیه شده .است مودس امکان انتقال داده های کتابشناختی گزیده از پیشینههای موجود مارک و ایجاد پیشینههای توصیفی برای اشیای محتوایی جدید را فراهم می آورد طرح فراداده ای مودس مجموعه ای از عناصر مورد نیاز برای توصیف اشیای دیجیتالی که از فیلدهای مارک 21 استخراج شده است را در بر میگیرد نامهای برچسب عناصر مودس بر خلاف مارک 21 در قالب ایکس ام ال مبتنی بر واژگان زبان طبیعی هستند (مک کالم(1)، 2004 دفتر استانداردهای مارک و توسعه شبکه e2013). کارکرد اصلی مودس کارکرد توصیفی است و پیشینه های آن در بخش فرادادههای توصیفی متس درج میشود.

McCallum -1

الگوی 3. شیوه درج پیشینه مودس در بخش فرادادههای توصیفی پیشینه متس

طرح فرادادهای هسته دوبلین (دی. سی. ام. آی.)

عکس



الگوی ۳. شیوه درج پیشینه مودس در بخش فرادادههای توصیفی پیشینه متس

طرح فرادادهای هسته دوبلین (دی. سی. ام. آی.)

طرحی بین‌المللی و میان رشته‌ای که مجموعه عناصری ساده و کارآمد برای توصیف طیف گسترده‌ای از اشیای محتوایی شبکه‌ای ارائه می‌دهد. این طرح نخستین تلاش جدی در حوزه طراحی استانداردهای فراداده‌ای پس از تعمیم شبکه جهانی وب محسوب می‌شود. کارکرد اصلی طرح هسته دوبلین نیز همانند طرح مودس، کارکرد توصیفی است. قالب ایکس. ام. ال. یکی از بسترهای نحوی هسته دوبلین است، و امکان پیاده‌سازی پیشینه‌های این طرح در قالب‌هایی دیگر نیز وجود دارد (جانستون[1] و پاول[2]، ۲۰۰۶؛ فتاحی و طاهری، ۱۳۸۸).

1. Juhnston
2. Powell

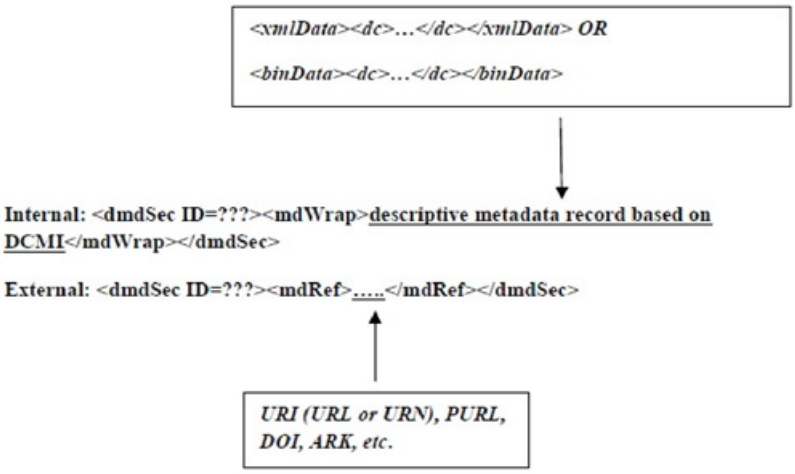
طرحی بین المللی و میان رشته ای که مجموعه عناصری ساده و کارآمد برای توصیف طیف گسترده ای از اشیای محتوایی شبکه ای ارائه

میدهد این طرح نخستین تلاش جدی در حوزه طراحی استانداردهای فراداده ای پس از تعمیم شبکه جهانی وب محسوب میشود کارکرد اصلی طرح هسته دوبلین نیز همانند طرح مودس کارکرد توصیفی است قالب .ایکس ام ال یکی از بسترهای نحوی هسته دوبلین است، و امکان پیاده سازی پیشینههای این طرح در قالبهایی دیگر نیز وجود دارد (جانستون(1) و پاول(2)، 2006؛ فتاحی و طاهری،1388).

1- Juhnston

2- Powell

عکس



الگوی ٤. شیوه درج پیشینه هسته دوبلین در بخش فرادادههای توصیفی پیشینه متس

طرح فرادادهای توصیف مستند (مَدس)

مدس شامل مجموعه عناصری برای توصیف دادههای مستند مربوط به اشخاص حقیقی و حقوقی، رویدادهای مهم، شناسههای موضوعی و جغرافیایی، و نظیر آن است، و به عنوان مکمل طرح فرادادهای توصیف شیء طراحی شده است (دفتر استانداردهای مارک و توسعه شبکه، b۲۰۱۳). با این وجود میتواند برای مستندسازی ارزشهای عناصر دیگر طرحهای فرادادهای با کارکرد مدیریتی و توصیفی مانند هسته دوبلین نیز استفاده شود. طرح مدس به صورت مستقیم در پیشینههای متس درج نمیشود و یا پیوند نمییابد، بلکه به صورت غیر مستقیم و پیوند با طرحهای فرادادهای با کارکرد مدیریتی یا توصیفی با متس تعامل بر قرار میکند.

<?xml version="1.0" encoding="UTF-8"?><mads xmlns=http://www.loc.gov/mads/ xmlns:mods="http://www.loc.gov/mods/v3" xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.loc.gov/mads/mads.xsd">
<authority><name><namePart>Smith, John</namePart><namePart type="date">1995-</namePart></name></authority><variant type ="other"><name><namePart>Smith, J</namePart></name></variant>
<variant type="other"><name><namePart>Smith, John J</namePart>

الگوی4.شیوه درج پیشینه هسته دوبلین در بخش فرادادههای توصیفی پیشینه متس

طرح فرادادهای توصیف مستند (مدس)

مدس شامل مجموعه عناصری برای توصیف دادههای مستند مربوط به اشخاص حقیقی حقوقی رویدادهای مهم شناسه های موضوعی و

جغرافیایی و نظیر آن است و به عنوان مکمل طرح فراداده ای توصیف شیء طراحی شده است دفتر استانداردهای مارک و توسعه شبکه b2013 با این وجود میتواند برای مستندسازی ارزشهای عناصر دیگر طرحهای فراداده ای با کارکرد مدیریتی و توصیفی مانند هسته دوبلین نیز استفاده شود. طرح مدس به صورت مستقیم در پیشینههای متس درج نمی شود و یا پیوند نمی یابد، بلکه به صورت غیر مستقیم و پیوند با طرحهای فراداده ای با کارکرد مدیریتی یا توصیفی با متس تعامل بر قرار میکند.

"xmlns:mods="http://www.loc.gov/mods/v3" xmlns:xlink="http://www.w3.org/1999/xlink

//:xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance xsi:schemaLocation="http

<"www.loc.gov/mads/mads.xsd

,Smith

-type="date " > 1995

John

type

other">Smith, J"=

Smith, John J

ص: 117

عکس

</name></variant><notetype="history">BiographicalnoteaboutJohnSmith.</note><affiliation><organization>Lawrence Livermore Laboratory</organization><dateValid>1987</dateValid></affiliation></mads>

نمونه ۱. نمونه‌ای از یک پیشینه مدس مربوط به یک شخص حقیقی

Internal: <dmdSec ID=???><mdWrap><mods><name type="personal"> <namePart type="termsOfAddress">Dr.</namePart> <namePart>*Smith, John*</namePart> </name></mods></mdWrap></dmdSec>

پشینیه مبتنی بر استاندارد مَدس

Internal: <dmdSec ID=???><mdWrap><mods><name type="personal"> <namePart type="termsOfAddress">Dr.</namePart> <namePart>*Smith, John*</namePart> </name></mods></mdWrap></dmdSec>

پشینیه مبتنی بر استاندارد مَدس

الگوی ۵. شیوه تعامل غیر مستقیم پیشینه‌های مدس با پیشینه متس به وسیله پیشینه مودس

پیشینه مبتنی بر استاندارد مدس در پایگاه مستند با استفاده از شناسگر پیشینه[1] با عنصر (فیلد) ارتباطی پیشینه مودس در پایگاه کتابشناختی که فقط ارزش‌های کد شده می‌پذیرد، پیوند می‌یابد، و فرایند کنترل مستندات را پشتیبانی می‌کند. بنابراین میان پیشینه‌های مدس و مودس به صورت مستقیم، و میان پیشینه‌های مدس و متس ارتباط غیر مستقیم بر قرار می‌شود، و بدین گونه کارکرد کنترل مستندات در نظام اطلاعاتی وجود خواهد داشت.

فراداده برای نگهداری اشیای دیجیتالی (پریمیس)

مجموعه عناصر مبتنی بر ایکس. ام. ال. که با هدف ثبت فراداده‌های مربوط به نگهداری شیء دیجیتالی در کتابخانه یا دیگر مجموعه‌های دیجیتالی گسترش یافته است. بنابراین کارکرد اصلی استاندارد پریمیس، نگهداری اشیای دیجیتالی است (هابینگ[2]، ۲۰۰۸). پیشینه‌های مبتنی بر این استاندارد بر اساس نوع

1. RecordID
2. Habing

.BiographicalnoteaboutJohn Smith

iliation> 1987

<dateValid

نمونه .1 نمونه ای از یک پیشینه مدس مربوط به یک شخص حقیقی

:Internal

<???=ID

Dr: Smith, John

<namePart

پشینیه مبتنی بر استاندارد مدس

:Internal

type="termsOfAddress">Dr.Smith, John

پشینیه مبتنی بر استاندارد مدس

الگوی 5 شیوه تعامل غیر مستقیم پیشینههای مدس با پیشینه متس به وسیله پیشینه مودس

پیشینه مبتنی بر استاندارد مدس در پایگاه مستند با استفاده از شناسگر پیشینه(1) با عنصر (فیلد) ارتباطی پیشینه مودس در پایگاه کتابشناختی که فقط ارزشهای کد شده می پذیرد، پیوند می یابد، و فرایند کنترل مستندات را پشتیبانی می.کند بنابراین میان پیشینههای مدس و مودس به صورت مستقیم و میان پیشینه های مدس و متس ارتباط غیر مستقیم بر قرار میشود و بدین گونه کارکرد کنترل مستندات در نظام اطلاعاتی وجود خواهد داشت.

فراداده برای نگهداری اشیای دیجیتالی (پریمیس)

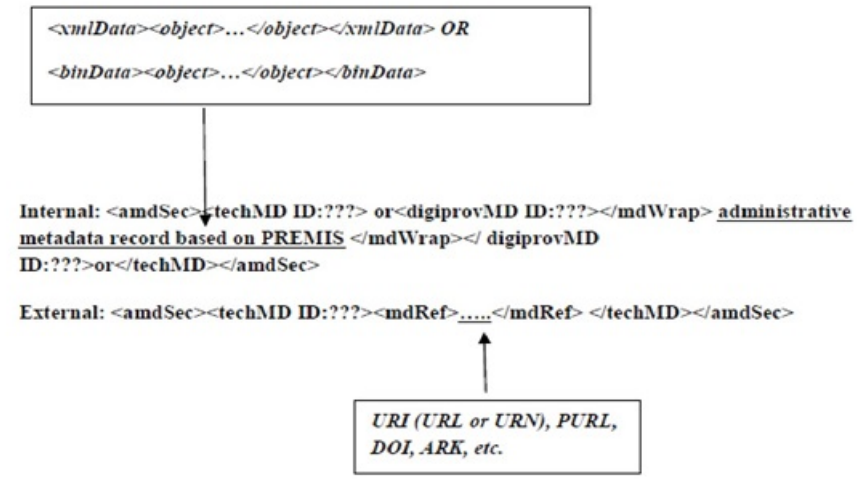
مجموعه عناصر مبتنی بر ایکس ام ال که با هدف ثبت فرادادههای مربوط به نگهداری شیء دیجیتالی در کتابخانه یا دیگر مجموعه های دیجیتالی گسترش یافته است بنابراین کارکرد اصلی استاندارد پریمیس نگهداری اشیای دیجیتالی است (هابینگ(2)، 2008)پیشینههای مبتنی بر این استاندارد بر اساس نوع

RecordID -1

Habing -2

عکس

موجودیتی که در بر می‌گیرند، باید در بخش فرادادههای مدیریتی متس و در عناصر <techMD> و <digiprovMD> درج گردند.



الگوی ٦. شیوه درج پیشینه پریمیس با کارکرد مدیریتی در بخش فرادادههای مدیریتی پیشینه متس

شکل ۱. نمونه‌ای از پیشینه‌های مبتنی بر پریمیس جاسازی شده در پیشینه متس

موجودیتی که در بر میگیرند باید در بخش فرادادههای مدیریتی متس و در عناصر درج گردند.

الگوی 6. شیوه درج پیشینه پریمیس با کارکرد مدیریتی در بخش فراداده های مدیریتی پیشینه متس

<Spremis:object

<Spremis:objectIdentifierType>uris/premis:objectIdentifierType

<Spremis:objectIdentifierValue>info:n la/nla.pic-vn 3579101-c premis:obje ctIdentifierValue

<premis:objectIdentifier

unknown premis: preservationLevel> file

<Spremis:objectCharacteristics

?Spremis:format

?Spremis:formatDesignation

image/tiff

6.0

<premis:format

<premis:objectCharacteristics

<premis:object

<Spremis:event

<Spremis:eventIdentifier

internal

<premis:evend dentifierValue 28903-1

<Spremis:evend dentifier

creation

Spremis:eventDateTime>2005-11-03T12:15:59

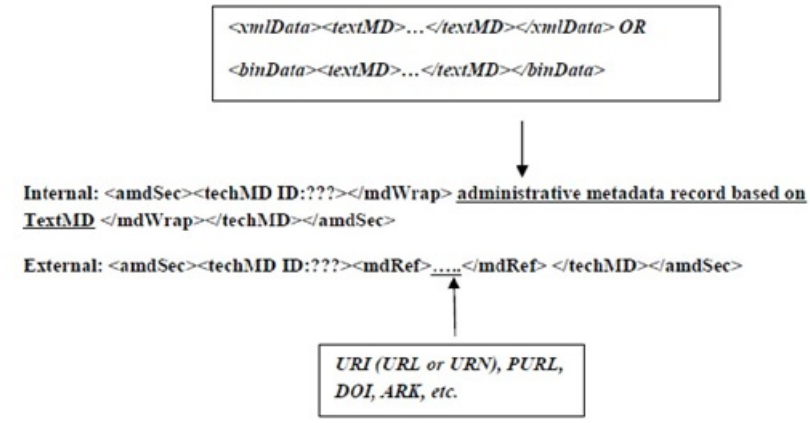
<premis:event

شکل .1 نمونه ای از پیشینههای مبتنی بر پریمیس جاسازی شده در پیشینه متس

عکس

فراداده فنی برای اشیای دیجیتالی متنی (تکست اِم. دی.)

استانداردی فراداده‌ای و مبتنی بر زبان نشانه‌گذاری گسترش‌پذیر که عناصری برای توصیف جنبه‌های فنی اشیای دیجیتالی متنی ارائه می‌دهد. پیشینه‌های مبتنی بر این استاندارد می‌توانند مستقیماً در بخش فراداده‌های مدیریتی متس، و یا به طور غیر مستقیم در عنصر <additionalTechnicalCharacteristics> مربوط به موجودیت شیء[1] استاندارد پریمیس درج گردند (دفتر استانداردهای مارک و توسعه شبکه، ۲۰۱۳f).



الگوی ۷. شیوه درج پیشینه تکست ام. دی. در بخش فراداده‌های مدیریتی پیشینه متس

فراداده برای مدیریت تصاویر در بستر ایکس. ام. ال. (میکس)

میکس محصول دفتر استانداردهای مارک و توسعه شبکه، با همکاری کمیته استانداردهای فراداده‌ای فنی برای تصاویر ثابت وابسته به «سازمان استانداردهای اطلاعات ملی (NISO)» است که به منظور مدیریت تصاویر دیجیتالی توسعه یافته است. کارکرد اصلی این طرح مدیریت فنی تصاویر ثابت دیجیتالی است (دفتر استانداردهای مارک و توسعه شبکه، ۲۰۱۳d)، و پیشینه‌های آن در بخش فراداده‌های مدیریتی متس در عنصر <techMD> قرار می‌گیرند.

1. Object

فراداده فنی برای اشیای دیجیتالی متنی (تکست ام. دی.)

استانداردی فراداده ای و مبتنی بر زبان نشانه گذاری گسترش پذیر که عناصری برای توصیف جنبه های فنی اشیای دیجیتالی متنی ارائه میدهد پیشینههای مبتنی بر این استاندارد میتوانند مستقیماً در بخش فراداده های مدیریتی متس، و یا به طور غیر مستقیم در عنصر additionalTechnicalCharacteristics مربوط به موجودیت شیء(1) استاندارد پریمیس درج گردند(دفتر استانداردهای مارک و

توسعه شبکه، f2013).

الگوی 7 شیوه درج پیشینه تکست .ام دی در بخش فرادادههای مدیریتی پیشینه متس

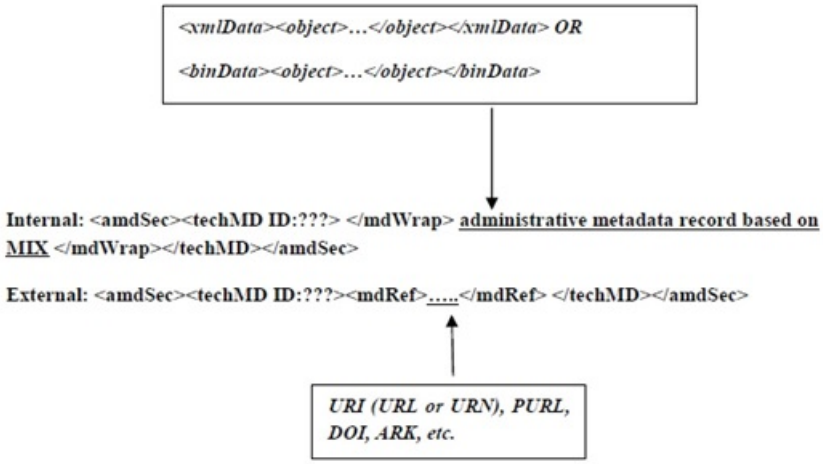
فراداده برای مدیریت تصاویر در بستر ایکس ام ال (میکس)

. میکس محصول دفتر استانداردهای مارک و توسعه شبکه با همکاری کمیته استانداردهای فراداده ای فنی برای تصاویر ثابت وابسته به سازمان استانداردهای اطلاعات ملی (NISO) است که به منظور مدیریت تصاویر دیجیتالی توسعه یافته است. کارکرد اصلی این طرح مدیریت فنی تصاویر ثابت دیجیتالی است (دفتر استانداردهای مارک و توسعه شبکه d2013) و پیشینههای آن در بخش فراداده های مدیریتی متس در عنصر techMD قرار میگیرند.

ص: 120

Object -1

عکس



الگوی ۸ شیوه درج پیشینه میکس در بخش فراداده‌های مدیریتی پیشینه متس

<metsHdr CREATEDATE="2013-01-05T14:00:00" RECORDSTATUS= "Complete"><agent ROLE="CREATOR" TYPE="INDIVIDUAL"><name>Sayyed Mahdi Taheri</name></agent></metsHdr>

<dmdSec ID:???></mdWrap> descriptive metadata record based on MARCXML </mdWrap></dmdSec><dmdSec ID=???><mdWrap>descriptive metadata record based on MODSExamples/mods99042030.xml</mdWrap></dmdSec><dmdSec ID=???><mdWrap>descriptive metadata record based on DCMIExamples/mods99042030.xml</mdWrap></dmdSec><amdSec><sourceMD ID:???></mdWrap> administrative metadata record based on MARCXML </mdWrap></sourceMD></amdSec><amdSec><techMD ID:???> or<digiprovMD ID:???></mdWrap> administrative metadata record based on PREMIS </mdWrap></ digiprovMD ID:???>or</techMD></amdSec><amdSec><techMD ID:???></mdWrap> administrative metadata record based on TextMD </mdWrap></techMD></amdSec><amdSec><techMD ID:???> </mdWrap> administrative metadata record based on MIX </mdWrap></techMD></amdSec><fileSec><fileGrp ID="VERS1"><file ID="FILE001" MIMETYPE="application/xml" SIZE="257537" CREATED="2013-01-05"><FLocat

الگوی 8 شیوه درج پیشینه میکس در بخش فرادادههای مدیریتی پیشینه متس

Sayyed Mahdi Taheri

<name

descriptive metadata record based on MARCXML_

ID=???>descriptive

record

based on MODSExamples/mods99042030.xml

ID=???>descriptive

metadata record based on

/DCMIExamples

mods99042030.xml

administrative metadata record based on MARCXML

amdSec> or

administrative metadata record based on PREMIS or

techMD> administrative metadata

record based on TextMD

ID:???> administrative metadata record based on MIX

<techMD

"ID="FILE001

<"MIMETYPE="application/xml" SIZE="257537" CREATED="2013-01-05

LOCTYPE="URL">http://dlib.nyu.edu/tamwag/beame.xml

<fileGrp

<"Introduction" ORDER="1

"BEGIN="INTVWBG

<div

to

<"2

"END="INTVWND

"FILEID="FILE001

</BETYPE="IDREF

<"1

"LABEL="Page TYPE="page" ID="P1

</"FILEID="HTMLF1

"ID="DISS1.1

"STRUCTID="S1.1" BTYPE="uva-bdef:stdImage" CREATED="2002-05-25T08:32:00

LABEL="UVA Std Image Disseminator" GROUPID="DISS1" ADMID="AUDR

<"EC1

-NEW AND IMPROVED Image Mechanism" LOCTYPE="URN" xlink:href="uva

</ "bmech:BETTER-imageMech

الگوی .9 برآیند الگوهای ارائه شده یک پیشینه کامل متس که تمامی استانداردهای مورد مطالعه را در بر میگیرد

همان طور که در الگوهای طراحی شده مشاهده میشود بستر نحوی استانداردهای فراداده ای مورد امکان میانکنش پذیری میان آنها را

فراهم آورده است. هر یک از این استانداردها میتوانند با

بررسی یکدیگر و با استاندارد هسته متس ارتباط بر قرار کنند و با پشتیبانی از کارکردهای گوناگون، یکپارچگی نظام های اطلاعاتی را میسر سازنـد به علاـوه میتوان بیش از یـک اسـتاندارد فراداده ای با کارکرد یکسان درون پیشـینههای متس جاسازی نمود هنگامی که بیش از یک اسـتاندارد و یا گزیده ای از عناصـر هر اسـتاندارد مود نیاز است و نیز میتوان پروفایل کاربردی یک مرکز یا محیط اطلاعاتی خاص را درون پیشـینه های متس بسـته بندی .نمود اسـتانداردهای مورد بررسـی تنها بخشـی مهمترین و پرکاربردترین از اسـتانداردهای فراداده ای بودند. بـدیهی است بـا اسـتفاده از بستر نحوی ،منـاسب میانکنش پـذیری دیگر اسـتانداردها نیز میسـر خواهـد بود این تاثیر بستر نحوی بر فرایند میانکنش پذیری فراداده ای را نشان می دهد.

نتیجه گیری

ضرورت اسـتفاده از چنـد اسـتاندارد فراداده ای در یـک نظام اطلاعاتی به منظور پشتیبانی از کارکردهای گوناگون مورد نیاز آن نظام و اقبال ویژه نظامهای اطلاعاتی به مقوله یکپارچگی و ارزشـهای افزوده مرتبط با آن بیانگر اهمیت فراینـد میانکنش پذیری فراداده ای است طراحی پروفایلهای کاربردی متناسب

ص: 122

با نیازهای مراکز یا محیطهای اطلاعاتی خاص نیز این اهمیت را دو چنـدان نموده است. بستر نحوی یکی از ارکان اصـلی فرایند میانکنش پـذیری میان اسـتانداردها و نظامهای فراداده ای است اسـتانداردهای فراداده ای با انتخاب بستر نحوی مناسب سـطح تعامل پذیری خود با دیگر اسـتانداردهای فراداده ای را افزایش می دهنـد، و بـدین گـونه علاـقه نظامهـای اطلاعـاتی بـه انتخـاب آنهـا را بر می انگیزاننـد گرایش اسـتانداردهای فراداده ای به انتخاب زبان نشانه گـذاری گسترش پـذیر به دلیل قابلیتهای منحصـر به فرد آن از جمله خود توصـیف بودن که فرایند میانکنش پذیری را در هر دو سـطح نحوی و معنایی تسهیل می نماید، به عنوان قالب اصلی و یا یکی از قالبهای پیاده سازی پیشینهها در همین راستا بوده است (حریری و دیگران ، 1391) بستر نحوی زمینه را برای پشتیبانی کارکردهای مورد نظر اسـتانداردهای فراداده ای و ارتباط پیشینه های فراداده ای با یکدیگر را فراهم نموده افزودن بر بهبود یکپارچگی درونی نظامهای اطلاعاتی، میانکنش پذیری آنها با دیگر نظامهـای اطلاعـاتی از جمله موتورهای کاوش وب به عنوان پرکاربردترین ابزار جسـتجو و بازیابی اطلاعات در شـبکه وب (کین، 2008؛ طاهری و حریری،2012)یـا به عبـارت دیگر یکپارچگی برونی یا آنها را نیز افزایش میدهـد امکان دسترسـی یکپارچه به اشـیای محتوایی منتشـر شـده در نظامهـای اطلاعـاتی مختلف از طریق ابزارهـایی چون دروازههای اطلاعاتی و درگاهها از پیامـدهای نیک قابلیتهای بستر نحوی میباشـند یکی دیگر از ارزشـهای افزوده ای که بستر نحوی مناسب تولیـد خواهد نمود تولید دانش بر اساس ارتباط میان پیشـینههای فراداده ای است که جلوه ای دیگر از یکپارچگی درونی و برونی نظامهای اطلاعاتی به شمار می آیند.

منابع

حریری نجلا سید مهدی طاهری سید رحمت الله فتاحی 1391. میانکنش پذیری نظامهای فراداده ای و موتورهای کاوش :وب تحولات و رویکردهـای جـاری پژوهشـنامه کتابـداری و اطلاع رسانی، 2 (2) طاهری، سـید مهـدی 1387 الف. طراحی یـک کتابخـانه دیجیتالی استاندارد. در مجموعه مقالات نخستین همایش کتابخانه های دیجیتالی به کوشش شرکت پارس آذرخش تهران سبزان

طاهری سید مهدی 1387ب. مقایسه کارایی طرح فرادادهای هسته دوبلین و قالب فراداده مارک 21 در سازماندهی منابع اطلاعاتی شبکه جهانی وب فصلنامه کتابداری و اطلاع رسانی، 43(پاییز (1387)

طاهری سید مهدی 1391 بررسی امکان نمایه سازی و پیدانمایی نامهای برچسب عناصر فراداده ای هسته ،دوبلین مارک ،21 و طرح فراداده ای توصیف شیء توسط موتورهای کاوش عمومی و ارائه الگوی مناسب رساله دکترا گروه کتابداری و اطلاع رسانی دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران

فتاحی، سید رحمت الله طاهری سید مهدی (1388) فهرستنویسی رایانهای مفاهیم، شیوهها و کاربرد نرم افزارهای رایانهای در سازماندهی اطلاعات با همکاری فرشته ناقدی احمدی تهران کتابدار

.Bray, Tim, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, François Yergeau. 2008

Extensible Markup Language (XML) 1.0 (Fifth Edition): W3C Recommendation 26

/November 2008. Retrievd 7 Jun. 2013 from: http://www.w3.org/TR/xml

//:Dublin Core Metadata Initiative 2013. Glossary. Retrieved 14 Jun. 2013 from: http

dublincore.org/documents/2003/08/26/usageguide/glossary.shtml

//:Gigee, Grant, Kelly 2006. MARC and MARCXML. Retrieved 5 Nov. 2011 from: http

threegee.files.wordpress.com/2006/05/marcxml.pdf

Gill, Toney 2008. Metadata and the web: Introduction to metadata. Retrieved 5 Jun. 2013

/from: http://www.getty.edu/research/publications/electronic_publications/intrometadata

metadata.pdf

Habing, Tom 2007. METS, MODS and PREMIS, Oh My!: Integrating Digital Library

.Standards for Interoperability and Preservation. Retrieved 5 Jun. 2013 from: http://www

.loc.gov/standards/mods/presentations/habing-ala07

:(Hirwade, Mangala Anil .2011. A study of metadata standards. Library Hi Tech News, 28(7

.18-25

Juhnston, pete, Andy Powell.2006. Expressing Dublin Core Metadata Using XML. Retrieved

.Jun. 2013 from: http://dublincore.org/documents/dc-xml 5

Maarof, M.H.B.S., Y. Yahya 2009. Digital libraries interoperability issues. Electrical

.Engineering and Informatics. ICEEI '09. International Conference on. Retrieved 5 Jun

from: http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?tp=arnumber=5254718isnu 2013

.mber=5254684

McCallum, H. Sally .2004. An Introduction to the Metadata Object Description Schema

-MODS). Retrieved 5 Jun. 2013 from: http://dldc.lib.uchicago.edu/talks/2004/lita2004)

greenstone/CS_Greenstone3_Olson.pdf

Network Development and MARC Standards Office (NDMSO) 2013a. MARC 21 XML .Schema. Retrieved 5 Jun. 2013 from: http://www.loc.gov/standards/marcxml

Network Development and MARC Standards Office (NDMSO) .2013b. Metadata Authority /Description Schema (MADS). Retrieved 5 Jun. 2013 from: http://www.loc.gov standards/mads

Network Development and MARC Standards Office (NDMSO) .2013c. Metadata Encoding /Ttransmission Standard (METS). Retrieved 5 Jun. 2013 from: http://www.loc.gov .standards/mets

Network Development and MARC Standards Office (NDMSO) .2013d. Metadata for Images .in XML Standard (MIX). Retrieved 5 Jun. 2013 from: http://www.loc.gov/standards/mix

Network Development and MARC Standards Office (NDMSO) 2013e. Metadata Object /Description Schema (MODS). Retrieved 5 Jun. 2013 from: http://www.loc.gov .standards/mods

.Network Development and MARC Standards Office 2013f. Technical Metadata for Text Retrieved 5 Jun. 2013 from: http://www.loc.gov/standards/textMD

.Network Development and MARC Standards Office 2013g. Understanding PREMIS .Retrieved 5 Jun. 2013 from: http://www.loc.gov/standards/premis/understanding-premis .pdf

Qin, Jian. 2000. Representation and organization of information in the web space: From .MARC to XML. Retrieved 5 Jun. 2013 from: http://inform.nu/Articles/Vol3/v3n2p83-88 .pdf

Taheri, S. M., Nadjla Hariri.2012. A comparative study on the indexing and ranking of the content objects including the MARCXML and Dublin Core's Metadata elements by .(general search engines. Electronic Library. 30(4

//:Word Wild Web Consortium.2013. HTML CSS. Retrieved 05 Jun. 2013 from: http www.w3.org/standards/webdesign/htmlcss

.Wikipedia 2013. Portable document fromat. Retrieved 05 Jun. 2013 from: http://en.wikipedia org/wiki/Portable Document Format

چکیده

آرشیو سازی وب به طور خودکار توسط خزشگرهای وب انجام میگیرد. این خزشگرهای صفحات را به صورت ادواری بازبینی و آرشیوها را با عکسهای جدید و تازه، روزآمد میکنند. مقاله حاضر به موضوع آرشیوسازی صفحات وب به طور کارآمد و بهسازی کیفیت آن اشاره دارد رویکرد پیشنهادی در این مقاله سه مفهوم را با هم تلفیق میکند بخش بندی صفحه ،دیداری شناسایی تغییر و اهمیت بلاکهای صفحه های وب برای تشخیص بهتر تغییرات مهم میان نسخهها چالش اصلی در این مقاله بهبود کیفیت آرشیو .است هدف ما این است که راهبردهای خزش را طوری تنظیم کنیم که خزش وبگاههای آینده تا آنجا که ممکن است به طور جامع و منسجم انجام شود بنابراین در مقاله حاضر نوعی راهبرد خزش مبتنی بر الگو و دو سنجه کیفیت جامعیت و انسجام برای ارزیابی راهبرد خزش پیشنهاد شده است.

بهینه سازی کیفیت آرشیوهای وب

نوشته: میریام بن سعد(1)

ترجمه: مهشید برجیان(2) – ساناز باغستانی(3)

مقدمه

با توجه به اهمیت روزافزون شبکه جهانی وب آرشیوسازی اطلاعات آن به منظور حفظ منابع اطلاعاتی مفید بسیار حائز اهمیت است. به همین دلیل حفظ وبگاههای مفید در نظر بسیاری از مؤسسه ها و سازمانهای آرشیوی ملی در سراسر دنیا به مسئله مهمی تبدیل شده است. اغلب اوقات، آرشیو سازی وب، به طور خودکار توسط خزشگرهای وب انجام میگیرد این خزشگرها، صفحه ها را به صورت ادواری بازبینی و آرشیوها را با عکسهای جدید و تازه روزآمد می.کنند دریافت تمام نسخه ها از کل وبگاه ها و حفظ کیفیت آرشیوها کار بی اهمیت و پیش پا افتاده ای نیست. در حقیقت، حفظ یک آرشیو کامل از کل وبگاه شامل تمامی نسخه های کل صفحه ها غیر ممکن است؛ زیرا وبگاهها دائماً در حال تکامل و گسترش هستند و امکانات و منابع(4) تخصیص یافته نیز معمولاً محدودند (مثل پهنای نوار، فضای ذخیره سازی و قوانین اخلاقی سایت). همچنین خزش یک وبگاه بزرگ ممکن است ساعتها و حتی روزها طول بکشد. این ،امر باعث افزایش احتمال تغییرات صفحه در حین خزش میشود و درنهایت

1- Myriam Ben Saad

2- کارشناس ارشد کتابداری و اطلاع رسانی سازمان اسناد و کتابخانه ملی ایران

3- کارشناس ارشد کتابداری و اطلاع رسانی

4- resources

عدم انسجام میان صفحه های آرشیو شده را به دنبال دارد به عنوان مثال فرض میکنیم که خزشگر در حال بارگذاری دو صفحه از وبگاه France است. اولین صفحه حاوی پیوند فرامتنی موسوم بهobama است که ما را به خزش صفحه دومی ارجاع میدهد که درباره رئیس جمهور آمریکا صحبت می کند. در حین خزش صفحهٔ ،اول محتوای صفحهٔ دوم به روز میشود و اکنون در مورد رئیس جمهور فرانسه یعنی سارکوزی صحبت می.کند ،بنابراین نسخه ذخیره شده صفحه اول به obama اشاره دارد با که نسخه صفحه ای که عمل خزش در آن انجام شده است و در مورد سارکوزی صحبت می کند، متناقض(1) میشود. مسئله تناقض بسیار رایج است و معمولا حین آرشیوسازی وب شناسایی میشود. بنابراین، سیستمهای آرشیو وب باید دو مسئله را در نظر بگیرند برای داشتن آرشیو جامع، عمل خزش را چگونه بهبود بخشند و چگونه کیفیت آن را با به حداقل رساندن عدم انسجام میان صفحات جمع آوری شده، حفظ کنند؟

یکی از روشهای ایده آل برای آرشیو سازی وب این است که تمام صفحه های سایت را در یک زمان و با هر تغییر خزش کنیم و یا از تغییر محتوای صفحات در طول عمل خزش جلوگیری نماییم، البته، این مورد به دلیل محدودیتهای امکانات و منابع عملاً غیر ممکن است حصول اطمینان 100 درصدی از جامعیت و نیز انسجام در آرشیو عملاً غیر ممکن است. با وجود این هدف ما این است که راهبرد خزش خود را طوری تنظیم کنیم که خزشهای وبگاه در آینده بتوانند تا آن جایی که امکان دارد آرشیوها را جامع سازد و تا حد امکان به آن هماهنگی و انسجام ببخشند. در این مقاله این دو چالش را برای بهبود کیفیت آرشیو بیان میکنیم.

یکی از ایده هایی که برای بهبود جامعیت آرشیو مطرح شده این است که مهمترین نسخه ها از صفحه ها را طوری بارگذاری کنیم که اطلاعات مفید از دست رفته به حداقل ممکن برسد. نسخه مهم، نسخه ای است که بعد از آرشیو آخرین نسخه واجد تغییر مهمی باشد. تغییرات بی اهمیت صفحه مثل تبلیغات را میتوان نادیده گرفت.

پژوهشهای بسیاری هستند که بر تخمین فراوانی تغییرات صفحه های وب برای بهبود خزشگرها

تأکید می کنند. اما هیچ روش مؤثری وجود ندارد که بتواند مشخص کند تغییرات مهم میان نسخه ها دقیقاً چه زمانی و هر چند وقت یک بار رخ میدهد تا امروز رویکردهایی که فراوانی تغییرات را برآورد کرده اند فقط میزان تغییرات شناسایی شده را مورد توجه قرار داده اند ولی اهمیت تغییرات رخ داده را در نظر نگرفته اند. اگر بتوانیم فراوانی تغییرات مهم را به طور دقیق تری پیش بینی کنیم، آنگاه می توانیم عمل خزش وب را بهبود بخشیم.

برای تخمین تعداد روز آمدسازیها باید تغییرات میان نسخههای بازیابی شده اسناد شناسایی شوند. بسیاری از الگوریتمهای موجود [21 ، 14 ، 28] به منظور شناسایی تغییرات اسناد نیمه ساختار یافته 14 (فرمت XML و فرمت HTML)طراحی شدهاند. با وجود این هیچ روشی وجود ندارد که بتواند تغییرات مرتبط نامرتبط را از اطلاعات مفید زائد اطلاعاتی که موجب اختلال میشوند شناسایی کند و آنها را

1- incoherence

از یکدیگر متمایز سازد.

هدف ما ارائه رویکردی است برای:1) شناسایی تغییرات مهم میان نسخه ها و به کارگیری آنها،2)بهینه سازی خزش ،وب 3)بهبود کیفیت نسخههای صفحه های آرشیو شده 4) نمایه سازی /ذخیره سازی مؤثر نسخههای صفحات رویکرد ما در یکی از مجموعه های(1) سازمان INA (مؤسسه شنیداری – دیداری ملی (فرانسه) به کار برده شده است. یکی از مأموریتهای INA ایجاد مجموعه ای(2) قانونی است که صفحههای وب رادیویی و تلویزیونی و نیز صفحه های مربوط به آنها را نگهداری .کند یکی از الزامات مهم این ،پروژه حفظ جنبه های دیداری صفحه هاست. بنابراین ما میخواهیم برای تعیین اهمیت بخشهای صفحه های وب با توجه به جایگاه نسبی آنها در صفحه، از روش آنالیز صفحه تصویری استفاده کنیم.

پژوهشهای پیشین [9،7] نشان میدهند که یک صفحه را میتوان به چند بخش یا چند بلاک تقسیم کرد این بلاکها غالباً در صفحه از اهمیت متفاوتی برخوردارند. در حقیقت بخشهای مختلف موجود در یک صفحه وب با توجه به مکان اندازه ،بخش و محتوا از وزنهای اهمیت متفاوتی برخوردار هستند معمولاً مهم ترین اطلاعات در مرکز صفحه تبلیغات در بالای صفحه(3) یا در سمت چپ، و بخش حق مؤلف در قسمت پایین صفحه(4) قرار دارند. هنگامی که یک صفحه بخش بندی می شود، برای هر بلاک باید یک اهمیت نسبی تعیین شود با استفاده از یک الگوریتم [26] و روش یادگیری ماشینی تحت نظارت میتوان به طور خودکار این فرآیند را انجام داد ،سپس میتوانیم میزان اهمیت تغییرات میان دو نسخه یک صفحه را براساس موارد زیر محاسبه کنیم:1) اهمیت نسبی بلاکها و 2)اهمیت نسبی عملیات (درج، حذف، روزآمدسازی و مانند آن) رخ داده در بلاکها که با مقایسهٔ این دو نسخه شناسایی شده اند. از این رو در پژوهش حاضر در صددیم که این مفاهیم را با هم ترکیب کنیم تا مسائل مربوط به شناسایی آنالیز تغییرات مهم صفحات وب را بیان کنیم سپس نتایج حاصل از این تجزیه و تحلیل را می توانیم در آرشیو سازی مؤثر صفحات وب به کار ببریم و موجب بهبود کیفیت وبگاههای آرشیو شده شویم. سؤالهای اصلی و درخور توجه این پژوهش که در آینده را بیان خواهیم کرد نیز مورد بحث قرار گرفته اند.

مسئله پژوهش

طراحی سیستمهای آرشیو وب چالشهای بسیاری را به همراه دارد؛ 1) بهینه سازی خزش وب، به منظور بهبود کیفیت آرشیو 2)نمایه سازی ذخیره سازی مناسب و 3)پرسش مؤثر در لحظه جست وجو با در نظر گرفتن مسئله بارگذاری نسخه های صفحه ای بی اهمیت بی فایده به طور قابل توجهی میتوان اثر بخشی این سه نکته را بهبود بخشید .بنابراین خزشگرهای وب باید به دقت مشخص کنند که کدام

1- repository

2- header

3- reposite

4- footer

صفحه و بـا چه الویتی بازسازی(1) / آرشـیو شود. آنها، همچنین، باید مشـخص کنند که برای به روز نگه داشـتن آرشـیو وب باید هر چند وقت یکبار صـفحه ها را مورد بازبینی قرار داد. در حقیقت، حفظ یک آرشـیو کامل از کل وب و یا حتی بخشـی از آن که شامل تمام نسخه ها از کل صـفحه ها می شود، کاری غیر ممکن است؛ زیرا وب دائماً در حال تکامل و گسترش است و امکانات و ابزارهای تخصیص یافته نیز معمولاًـ محـدود هسـتند .بنـابراین این مسـئله را میتوان به این صورت مطرح کرد: چگونه عمل خزش وب را بهبود بخشـیم تا مهمترین نسـخهها را بارگذاری کنیم طوری که اطلاعات مفید از دست رفته به حداقل ممکن برسـند؟ یک نسـخهٔ مهم، نسـخه ای است که تغییرات مهمی بعد از آخرین نسخه آرشیو شده در آن صورت گرفته باشد البته این مسئله باید بدون هیچ کمکی از سوی مدیران وبگاه ها حل شود. بنـابراین، به روشـی مؤثر و مفیـد نیـاز داریم تـا به واسـطه آن بـدانیم که تغییرات مهم در چه زمانی و هر چنـد وقت یکبار بین نسـخه ها در وبگاهها رخ میدهنـد تاکنون، رویکردهای موجودی که فراوانی خزنده ها را برآورد میکنند اهمیت تغییرات میان نسـخه ها را در نظر نگرفته اند در حقیقت اغلب اتفاق می افتد که خزشـگرها صـفحه هایی را بارگذاری میکنند که دارای اطلاعات بی اهمیت هسـتند (به عنوان مثال تبلیغاتی که به روز میشوند) برای برآورد تعداد مناسب خزنده ها، تغییرات میان نسـخه های بازیابی شده باید شناسایی شوند و مورد تجزیه و تحلیل قرار گیرند با وجود اینکه برای شـناسایی تغییرات میان اسـناد، الگوریتمهای مختلفی طراحی شده اند هیچ روشـی وجود ندارد که بتواند تغییرات مهم بی اهمیت را از اطلاعات مفید بی فایده شناسایی کند و آنها را از هم متمایز سازد.

در این مقاله، ما برخی چالشهای مهم را بیان میکنیم:1) سیستمهای آرشیوسازی چگونه میتوانند تغییرات مفید میان نسخههای آرشیو شده را شـناسایی کننـد و چگونه میتوانند اهمیت آنها را تعیین کنند؟ 2)با توجه به امکانات و ابزار محدود و تعداد زیاد اسـنادی که باید آرشـیو شونـد خزشـگرها چگونه می تواننـد ضـروری ترین / مهمترین نسـخه بازسازی شـده را انتخاب کنند؟ 3)برای بهبود کیفیت (جـامعیت و انسجام) وبگاههای آرشیو شده چگونه میتوان از نتایج آنالیز اهمیت تغییرات بهره برد؟

پژوهشهای مرتبط

در ادامه به مطالعات مرتبط با این تحقیق میپردازیم. موضوع های اصـلی عبارت انـد از: آرشـیوسازی وب تجزیه و تحلیل دیداری صـفحه ،وب شناسایی تغییرات و خزش وب.

آرشـیوسازی وب مؤسسـههای آرشیوی متعددی (کتابخانه های ،ملی آرشیوهای داده های تاریخی، و مانند آن) در سراسر دنیا، برای حفظ میراث وب کشور خود چندین پروژه را راه اندازی کرده اند. برخی مطالعات بر تعیین محدوده وب برای انتخاب صـفحات برای آرشیو شدن تأکیـد دارنـد. پژوهشـهای دیگری نیز بر روی مـدل سـازی و ارزیابی فراوانی تغییرات وب کار کرده انـد. آنها برای بهبود بازسازی آرشـیو، تخمین زنهای فراوانی تغییرات و سیاسـتهای بازسازی گوناگونی را ارائه میدهنـد. برخی محققان نیز مسائل مربوط به فرمت اطلاعات ذخیره سازی نمایه سازی شده را با ارائه سیستم ذخیره سازی خودشان

---

refresh -1

بیـان میکننـد مطالعات دیگر بر کنترل و نمایش تغییرات تأکیـد میکننـد این مطالعات، برای پرسـش(1) و ذخیره سـازی مؤثر آرشـیو ،وب الگوریتم شـناسایی تغییر و یـا فرمت دلتـا ارائه میدهنـد. کارهـای اخیر نیز مسـئله انسـجام و کیفیت آرشـیو را توسط راهبردی خزش بیان میکنند.

جالب اینجاست که این روشـها و رویکردها اهمیت تغییرات صـفحه ها برای آرشیوسازی وب به طور مؤثر را در نظر نمی گیرند؛ در حالی که محور اصلی روش ما همین اهمیت است.

تجزیه و تحلیل دیـداری صـفحه وب برای تجزیه و تحلیل نمایش دیـداری صـفحات وب از چنـدین روش اسـتفاده شـده است. بیشتر این روشـها، ساختار منطقی صـفحه را با آنالیز اسناد ارائه شده یا آنالیز کد اسناد، کشف می کنند. گو(2) و همکارانش [20]، نوعی الگوریتم بالا به پایین ارائه دادند که الگوریتم ساختار محتوای وب را مبتنی بر اطلاعات صـفحه آرایی شناسایی میکند. کوواشنیک(3) و همکارانش [17]، برای شـناسایی بخشـهای رایج یک (صـفحه بالای ،صـفحه پایین صـفحه مرکز صـفحه) فرآیندهای مکاشـفه ای مبتنی بر اطلاعات دیداری را تعیین کردند. کای(4) و همکارانش [9] الگوریتم VIPS را مطرح کردند. این الگوریتم براساس اطلاعات دیداری بازیابی شده توسط مرورگر صـفحه وب را به چندین بلاک معنایی تقسـیم میکند کوسولشی(5) و همکارانش [15] رویکردی را مطرح کردنـد که میزان تشابه بلاکها را در صـفحه های وب با استفاده از اطلاعات موقعیتی عناصر درخت DOM محاسبه میکند. به نظر میرسد که روش VIPS در مقایسه با روشهای موجود مناسبترین روش برای رویکرد ماست زیرا دانگی(6) مناسبی برای بخش بندی صفحه ایجاد میکند منظور) از دانگی، اندازه قطعه های حافظه در سیسـتم مجازی است این ،روش سلسله مراتبی از بلاکهای معنایی صفحه را ایجاد می.کند. این سلسله مراتب چگونگی درک کـاربر از ساختـار صـفحه آرایی ،وب مبتنی بر درک دیـداری وی را بهتر شبیه سازی می کنـد. از این رو برای ایجاد ساختار دیداری اسناد از VIPS استفاده شد.

شناسـایی تغییرات برای شـناسایی تغییرات میان دو نسـخه یک سـند نیمه ساختاری (XML و HTML)، چنـدین الگوریتم طراحی شده است. این الگوریتمها حداقل مجموعه ای از عملیات تغییر (درج، حذف، و) را پیدا می کنند که یک درخت دادهها را به درخت دادههای دیگر تبـدیل میکند. این عملیات تغییر غالباً در یک متن دلتا و یا یک فایل دلتا گردآوری میشوند. طراحی الگوریتمهای مختلف به اهداف و الزامات(7) آنهـا پیچیـدگی ،زمانی عملیاتی که قرار است به کار برده شونـد کیفیت دلتا و ماننـد آن) بسـتگی دارد. کوبنا(8) و همکارانش [14] برای بهبود مـدیریت حافظه و مدیریت زمان الگوریتم XyDiff را مطرح کردند. الگوریتم XyDiff عملیات انتقال(9) را پشتیبانی میکند و پیچیدگی زمانی ((on log(n را به دست می آورد.

1- query

2- Giu

3- Kovacevic

4- Cai

5- Cosulshi

6- granularity

7- requirement

Cobena –8

move –9

این الگوریتم، با وجود عملکرد بالایی که دارد همیشه نمیتواند نتیجهٔ بهینه و مطلوبی را تضمین کند (منظور از نتیجه بهینه و مطلوب حداقل ویرایش برای متن است). وانگ(1) و همکارانش[28]، الگوریتم XyDiff را مطرح کردند این الگوریتم میتواند تفاوتهای مطلوب میان دو درخت سامان نیافته XML در معادله زمانی درجه دوم(O(n(2)) را شناسایی کند؛ ولی هیچ انتقالی را پشتیبانی نمیکند الگوریتم Delta 21]] این اسناد XML را برای درختان سامان یافته و سامان نیافته با حمایت از عملیات اصلی میتواند مقایسه ادغام و هماهنگ میکند؛ اما انتقال را شناسایی نمیکند الگوریتمهای دیگری همانند الگوریتم DTD – Diff[22] و غیره نیز وجود دارند پس از مطالعه و بررسی این الگوریتمها تصمیم گرفتیم که برای رویکرد آرشیو سازی وب خود از روشهای موجود استفاده نکنیم زیرا هدف آنها کلی است. از آنجا که الزامات خاص متفاوتی در ارتباط با ساختار صفحه آرایی دیداری اسناد وجود دارد، ترجیح میدهیم که از الگوریتم ویژه و متناسب با کار خود استفاده کنیم (الگوریتم Vi DIFF).این الگوریتم امکان ارزیابی بهتر پیچیدگی و جامعیت مجموعه عملیات شناسایی شده را فراهم می کند.

خزش وب

تعدادی از پژوهشهای موجود مسئله بهینه سازی خزش وب را از طریق ایجاد راهبردیهای زمان بندی [23 و 10 و یا از طریق برآورد فراوانی ،تغییرات بیان میکنند [27 و 16] مطالعات اخیر [13 و 12]، مسئله کیفیت و انسجام آرشیو را با ارائه راهبردی خزش وبگاه، بیان می کنند. اما، از آنجا که راهبردهای آنها مبتنی بر فرآیند پوآسون است برای صفحاتی که زود به زود تغییر میکنند مفید نیستند (مانند صفحات رادیویی و تلویزیونی همچنین تا آنجا که ما میدانیم؛ مطالعات خزش ،موجود اهمیت تغییرات رخ داده در نسخههای تجزیه و تحلیل شده را در نظر نمی گیرند. اگر بتوانیم فراوانی تغییرات مهم را به طور دقیق تری پیش بینی کنیم شاید بتوانیم از نمایه سازی و ذخیره اطلاعات بی اهمیت و غیر ضروری جلوگیری کنیم و اثر بخشی و کیفیت آرشیو وب را بهبود بخشیم.

2.رویکرد آرشیوسازی وب

رویکرد آرشیوسازی وب ما، آنالیز آرشیوسازی ساختار دیداری اسناد و تعیین ارزشهای(3) اهمیت برای بلاکهای صفحه های وب با توجه به جایگاه نسبی آنها در صفحه است. به عبارت دیگر، نسخه های یک صفحه مطابق با نمایش دیداری شان بازسازی میشوند شناسایی تغییرات در چنین نسخه های صفحه ای بازسازی شده اطلاعات مناسبی را برای درک دینامیک وبگاهها ارائه میدهد همچنین امکان تشخیص تغییرات مرتبط نامرتبط را از اطلاعات مفید بی فایده فراهم می آورد ،بنابراین، روش مطرح شده، سه مفهوم زیر را با هم ترکیب میکند آنالیز دیداری صفحه(بخش بندی)، شناسایی تغییرات، و اهمیت بلاکهای صفحه های وب به منظور بهینه سازی خزش .وب این مفاهیم جدید نیستند؛ اما تا آنجا که ما

Wang –1

importance values –2

Poisson Process –3

میـدانیم هرگز این مفاهیم را برای آرشـیوسازی وب با هم ترکیب نکرده اند. معماری آرشـیو وب به طور مفصل تری در [6] شرح داده شده است.

بخش بندی صفحات دیداری

ما مـدل بخش بندی دیداری موجود VIPS [9] را برای ایجاد ساختار دیـداری صـفحه های وب گسترش دادیم از مدل VIPS برای بخش بنـدی صـفحه وب به بلاکهای معنایی تو در تو مبتنی بر گره های مناسب در درخت HTML DOM ،صـفحه استفاده شد. این مدل جدا کننده های افقی و عمودی را در صـفحه وب شناسایی می.کند همچنین این مدل براساس جداکننده ها درخت معنایی صفحه وبی را ایجاد میکند که به چندین بلاک تقسـیم بندی شده است. اساس کار کل صفحه است. هر بلاک به عنوان یک گره در درخت نشان داده می شود. با اسـتفاده از اسـتخراج پیوندها، تصاویر، و متن برای هر بلوک، الگوریتم VIPS را برای تکمیل درخت معنایی کل صـفحه گسترش دادیم الگوریتم VIPS توسعه یافته ما نوعی سند Vi-XML به عنوان خروجی تولید می.کند این سند ساختار سلسله مراتبی کامل صفحه وب را توصیف می نماید.

شناسایی تغییرات

برای شناسایی تغییرات میان دو صفحه وب براساس بعد دیداری الگوریتم Vi Diff را مطرح کردیم. این الگوریتم دو نوع تغییر را شناسایی میکنـد تغییرات ساختاری و تغییرات محتوایی. تغییرات ساختاری (درج، حـذف و جابه جایی)، معمولاً ساختار سـند XML را در سطح بلاکها تغییر میدهند؛ در حالی که تغییرات محتوایی ،درج، حذف روزآمدسازی و انتقال محتوای متنی را در سـطح پیوندها، تصاویر، متنها تغییر میدهنـد تمامی عملیات تغییر شـناسایی شده در یک فایل Vi-Delta توصـیف میشونـد. اگر فرض کنیم که هیچ تغییری در ساختار وجود ندارد میزان پیچیدگی Vi-Delta لگاریتم خطی ((onlog(n است که در آن همان تعـداد کلی گره هاست. اگر تغییرات ساختاری وجود داشـته باشند، در بدترین حالت حالتی که تمام ساختار تغییر (کند که پیچیدگی به صورت معادله درجه دوم o(n است؛ ولی ارزش دارد که ببینیم n همیشه اندازه اش کوچک باقی میماند.

اهمیت تغییرات

با توجه به Vi-Delta ایجاد شده توسط Vi Diff تابعی [4] را ارائه میدهیم که اهمیت تغییرات شناسایی شده را ارزیابی میکند. این تابع به سه پارامتر اصلی بستگی دارد:

اهمیت بلوک روزآمـد شـده معمولاً مهمترین اطلاعات در مرکز صـفحه وب و تبلیغات در قسـمت بالای صـفحه و ماننـد آن قرار میگیرند بنـابراین میتوان اهمیت بلاکهـا را با توجه به جایگاه نسبی آنها در صـفحه تعیین کرد به عنوان مثال میتوان از روش سانگ(1) و همکارانش [26] برای به دست آوردن آن

Song -1

استفاده کرد. آنها از الگوریتمهای یادگیری ماشینی تحت نظارت مبتنی بر ویژگیهای محتوایی و فضایی استخراج شده بلاکها استفاده میکنند تا میزان اهمیت هر بلاک به طور خودکار تعیین شود. همچنین، میتوانیم پارامترهای دیگری را برای ارزشیابی اهمیت هر بلاک با توجه به تاریخچه تغییرات آن در نظر بگیریم. به عنوان مثال، می توانیم فرض کنیم که هر چه یک بلاک بیشتر تغییر کند، میزان اهمیت آن کمتر .است در حال حاضر به دنبال بهترین تکنیک برای تخمین میزان اهمیت بلاکها هستیم.

اهمیت عملکرد اهمیت عملکردها به نوع عملیات (انتقال درج و مانند آن) و عنصر تغییر یافته پیوند، تصویر، و مانند آن بستگی دارد؛ مانند عملیات درج و حذف که میتوانند مهمتر از عملیات انتقال محسوب شوند. همچنین درج یک تصویر میتواند مهمتر از درج یک پیوند و یا متن باشد. ما می خواهیم برای انتخاب بهترین پارامترها برای هر کدام از انواع عملیات به مطالعه روشهای یادگیری ماشینی بپردازیم.

میزان تغییر هر بلوک میزان تغییر عملیاتی (حذف ، درج و مانند آن) که برای هر عنصر (پیوند، تصویر و متن) در هر بلاک ایجاد میشود از VI-Delta ایجاد شده استنتاج میشود. این میزان درصد تغییر عملیات مشخص شده در هر بلاک (این بلاکها خود به تعدادی عنصر تقسیم شده اند) را نشان میدهد. عملیات پیشنهادی با توضیحات دقیق تر در [4] توصیف شده است.

آزمایشها

با استفاده از الگوریتم VIPS گسترش داده شده، آزمایشهای بخش بندی دیداری بر روی صفحه های وب HTML انجام شدند ما کارآیی های بخش بندی دیداری در طول زمان و اندازه خروجی را اندازه گیری کردیم همچنین آزمایشهایی برای آنالیز کارآیی الگوریتم Vi-DIFF پیشنهادی در طول زمان و اندازه برونداد، انجام شدند. آزمونها نشان میدهند که مدت زمان امیدوار کننده است. مدت زمان کلی رضایت بخش است زیرا این زمان امکان پردازش بیش از 100 صفحه (اندازه گیری فعلی در هر لحظه (ثانیه) و در هر پردازشگر شرایط پروژه کارتک CARTEC را فراهم می.کند به هر حال، زمان بخش بندی بسیار بیشتر از زمان مقایسه است برای بهینه سازی بیشتر ،سیستم باید بر کاهش زمان بخش بندی و یا جلوگیری از بخش بندی تمامی نسخه های صفحه تمرکز کنیم.

کیفیت آرشیو وب

یکی از اهداف ما بهره وری از اهمیت تغییرات شناسایی شده برای بهسازی کیفیت آرشیو است رویکرد ایده آل آرشیوسازی ،وب خزش تمامی صفحه های وبگاه به طور همزمان در هر تغییر و یا جلوگیری از تغییر محتویات صفحه در طول خزش است البته باتوجه به تعداد زیاد صفحه های هر سایت و محدویتهای امکانات و منابع این کار به طور عملی غیر ممکن است بنابراین، نمی توانیم آرشیو جامعی داشته باشیم آرشیوی که تمامی نسخه های تمامی صفحه های سایت را در بردارد. همچنین اطمینان از انسجام کل آرشیو صفحه های گردآوری شده وضعیت واقعی یک سایت را در یک لحظه از زمان

عکس

منعکس می‌کنند) غیرممکن است. با وجود این، قصد داریم که راهبردهای خزش را طوری تنظیم کنیم که خزش وبگاه در آینده تا آنجا که ممکن است جامع و منسجم باشد. تعدادی از راهبردهای خزش بر اساس مدل فرآیندپواسون هستند. این مدل، زمانی مناسب است که تقسیم‌بندی زمانی بیش از یک ماه باشد. به هر حال، کار ما بر روی صفحه های وبی است که بارها تغییر می‌کنند (چندین بار در روز) مانند صفحه های وب تلویزیون، رادیو، و صفحه های خبری. تقسیم‌بندی زمانی برای این صفحات کمتر از یک ماه است. بنابراین، مدرک مهمی وجود دارد که مدل Poission برای این مورد اعتبار ندارد. [۷،۹،۲۵] با بازنگری تغییرات صفحه مشاهده کردیم که روزآمد سازی صفحه به ساعت روز و روز هفته بستگی دارد. همچنین، مشاهده کردیم که تغییرات در طول روز به میزان قابل توجهی بیشتر از شب و در روزهای کاری بیشتر از آخر هفته هستند.

بنابراین، با بازنگری تغییرات صفحه از یک خزش ناپیوسته[1] الگوهایی را کشف می‌کنیم. یک الگو، رفتار تغییرات صفحه و اهمیت آنها را در طول زمان به‌عنوان مثال، در طول یک روز طراحی می‌نماید. این مسئله، به روز هفته بستگی دارد. الگو باید پی‌درپی روزآمد شود تا همیشه بتواند تغییرات جاری صفحات وب را منعکس کند. صفحات یک سایت با الگوی مشابه می‌توانند برای داشتن یک الگوی مشترک گروه‌بندی شوند. بر اساس این الگوها، راهبرد این خزشگرهای وبی تنظیم می‌شوند و به‌طور مؤثری در صفحههای وب می‌خزند و کیفیت آرشیو را بالا می‌برند.

تعریف. یک الگو از صفحه P_i با طول فاصله l دنباله

$$Patt(P_i)=\{(W_1,T_1);(W_2,T_2); \ldots; (W_k,T_k)\}$$

است به طوری‌که W_k میانگین اهمیت تغییرات در زمان T_k است. مجموع زمانها T_j $\Sigma_{j=1}^{k}$ برابر است با l.

ما l را برابر یک روز، به‌عنوان طول الگو، در مدل مورد نظرمان انتخاب کردیم.

Page Changes Pattern

Periods T	Workdays ω_k	Saturday ω_k	Sunday ω_k
[0:00-6:00]	0,2	0,1	0,2
[6:00-12:00]	0,4		
[12:00-18:00]	0,6	0,4	0,35
[18:00-24:00]	0,1	0,2	0,13

شکل ۱. نمونه الگو

مثال. همانطور که در شکل ۱ نشان داده شده است، دنباله زیر الگوی دوره‌ای صفحه P_1 برای روزهای هفته است.

$$Patt(P_1)=\{(0.2,[0h-6h]);(0.4,[6h-12h]);(0.6,[12h-18h]);(0.1,[18h-24h])\}$$

1. off-line

منعکس میکنند غیر ممکن است با وجود این قصد داریم که راهبردهای خزش را طوری تنظیم کنیم که خزش وبگاه در آینده تا آنجا که ممکن است جامع و منسجم باشد تعدادی از راهبردهای خزش بر اساس مدل فرآیند پواسون هستند این مدل زمانی مناسب است که تقسیم بندی زمانی بیش از یک ماه باشد به هر حال کار ما بر روی صفحه های وبی است که بارها تغییر میکنند چندین بار در روز مانند صفحه های وب ،تلویزیون رادیو و صفحه های خبری تقسیم بندی زمانی برای این صفحات کمتر از یک ماه است بنابراین مدرک مهمی وجود دارد که مدل Poission برای این مورد اعتبار ندارد. [79،25] با بازنگری تغییرات صفحه مشاهده کردیم که روزآمد سازی صفحه

به ساعت روز و روز هفته بستگی دارد. همچنین، مشاهده کردیم که تغییرات در طول روز به میزان قابل توجهی بیشتر از شب و در روزهای کاری بیشتر از آخر هفته هستند.

،بنابراین با بازنگری تغییرات صفحه از یک خزش ناپیوسته(1) الگوهایی را کشف میکنیم یک الگو را رفتار تغییرات صفحه و اهمیت آنها را در طول زمـان به عنوان مثال در طول یک روز طراحی مینمایـد. این مسـئله به روز هفته بسـتگی دارد. الگو بایـد پی در پی روزآمـد شود تا همیشه بتوانـد تغییرات جاری صـفحات وب را منعکس کند صـفحات یک سایت با الگوی مشابه میتوانند برای داشـتن یک الگوی مشترک گروه بندی شوند. بر اساس این الگوها راهبرد این خزشـگرهای و بی تنظیم میشوند و به طور مؤثری در صفحههای وب می خزند و کیفیت آرشیو را بالا میبرند.

تعریف یک الگو از صفحه . با طول فاصله 1 دنباله

{(Patt (P={(W,,T,);(W,,T); …; (W,T)

است به طوری که میانگین اهمیت تغییرات در زمان T است. مجموع زمانها W T:

است با 1.

ما 1 را برابر یک ،روز به عنوان طول الگو در مدل مورد نظرمان انتخاب کردیم.

شکل .1 نمونه الگو

مثال همانطور که در شکل 1 نشان داده شده است دنباله زیر الگوی دوره ای صفحه برای روزهای هفته است.

{([Patt(P)-([h-th]); (., [h-h]); (., [h-Ah]); (.,[Ah-th

---

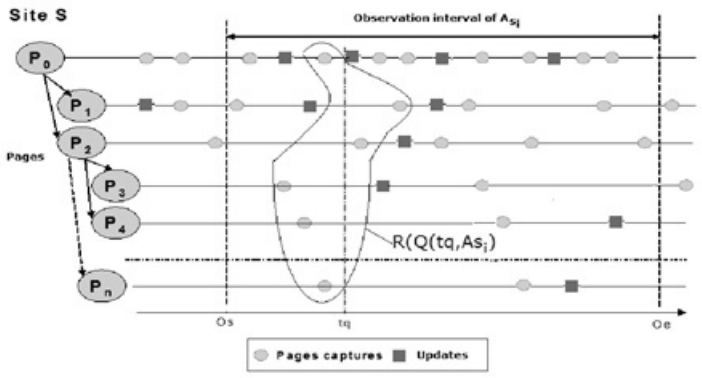
off-line -1

عکس

مقدارهای عددی ۰/۲، ۰/٤، ۰/٤، ۰/۱ به‌ترتیب میانگین اهمیت تغییرات برای هر بازه زمانی
T_1=[۰h-٦h],...,T_i=[۱۸h-۲٤h]
هستند. همچنین، می‌توان همانطوری که در شکل یک نشان داده شد، می‌توان به‌طور جداگانه برای روزهای شنبه و یکشنبه (آخر هفته) تعریف کرد.

مدل آرشیو وب

در مدل آرشیو وب مورد نظر، تمامی صفحه‌های وبگاه، بارها به‌صورت جداگانه ذخیره شده‌اند. خزشگر به‌طور معمول، بر روی تقسیم‌بندی صفحه‌ها کار می‌کند، اما بر روی تقسیم‌بندی یک وبگاه کاری انجام نمی‌دهد. خزشگر، مهم‌ترین صفحه‌های بازسازی شده را انتخاب می‌کند. به‌عنوان مثال، یک صفحه در هر ثانیه مورد خزش قرار می‌گیرد. صفحه‌هایی که بارها تغییرات قابل ملاحظه‌ای را در بردارند، بیشتر بازبینی می‌شوند. آرشیو وبی ما، در As_i مجموعه‌ای از نسخه‌های صفحه جداگانه‌ای تعریف می‌شود، که از یک سایت S_i بارگذاری شده‌اند. همانطور که در شکل ۲ نشان داده شده است، یک بازه مشاهده برای تعیین شروع(o_s) و پایان(o_e) مشاهده آرشیو تعریف می‌شود.



شکل ۲. مدل آرشیو وب

تصویربرداری از صفحات ¡

به روزرسانی ۱

آرشیو، توسط پرسش استفاده‌کننده‌ای $Q(t_q, A_S)$ که آخرین نسخه‌های در دسترس صفحه‌های وبگاه سایت S را در زمان ارسال پرسش t_q جست‌وجو می‌کند، قابل دسترس است.

توجه: فرض می‌کنیم که دنباله $\{S_1, S_2, ..., S_n\}$ = فهرست سایت‌هایی است که باید مورد خزش قرار گیرند. هر سایت از n صفحه $\{P_1, ..., P_n\}$ تشکیل شده است. هر صفحه P_i الگوی

مقدارهای عددی 0/2 ، 0/4 ، ،0/4 ، 0/1 به ترتیب میانگین اهمیت تغییرات برای هر بازه زمانی

هستند. همچنین میتوان همانطوری که در شکل یک نشان داده شد میتوان به طور جداگانه برای

روزهای شنبه و یکشنبه (آخر هفته) تعریف کرد.

مدل آرشیو وب

در مـدل آرشـیو وب مورد نظر تمامی صـفحه های وبگاه بارها به صورت جـداگانه ذخیره شده.اند خزشـگر به طور معمول بر روی تقسـیم بندی صفحه ها کار میکند اما بر روی تقسیم بندی یک وبگاه کاری انجام نمی دهد خزشگر مهمترین صفحه های بازسازی شده را انتخاب میکنـد به عنوان مثال، یک صـفحه در هر ثانیه مورد خزش قرار میگیرد صـفحه هایی که بارها تغییرات قابل ملاحظه ای را در بردارند بیشتر بازبینی می شوند آرشیو وبی ما در AS مجموعه ای از نسخه های صفحه جداگانه ای تعریف می شود، که از یک سایت S: بارگذاری شده اند همانطور که در شکل 2 نشان داده شده است یک بازه مشاهده برای تعیین شروع (o) و پایان (o) مشاهده آرشیو تعریف می شود.

شکل 2 مدل آرشیو وب

تصویر برداری از صفحات :

به روزرسانی 1

آرشیو، توسط پرسش استفاده کننده ای که آخرین نسخه های در دسترس صفحه های وبگاه

سایت S را در زمان ارسال پرسش t جست و جو میکند قابل دسترس است.

تـوجه فرض میکنیـم کـه دنبـاله { s,s, s,,..} = فهرسـت سایتهـایی اسـت کـه بایـد مـورد خزش قرار گیرنـد هر سـایت از 1 صفحه {p...,,,p}تشکیل شده .است. هر صفحه : الگوی P

{ (WT) ..... (WT);(W,T,); Patt(P) = را دارد. W، میانگین اهمیت تغییرات در زمان است. ما به نسخه صفحه P که در زمان t توسط v ذخیره شده است توجه میکنیم فرض میکنیم (P) کپی واقعی صفحه است. که بدون مکث تغییری را در زمان t دنبال میکند.

تعریف آرشیو AS توسط مجموعه ای از نسخه های صفحه ای (P) تعریف می شود. که از سایت Sدر طول بازه مشاهده [0] ذخیره شده اند؛ به طوری که in≥1 و jk≥1 .

تعریف RQt (AS)، نتیجه پرسش Q(t As) استفاده کننده تعریف می شود. همانطور که در شکل 2 نشان داده شد RQt (AS) مجموعه ای از n نسخه صفحه های ذخیره شدهای V(P) را نشان میدهد که نزدیکترین [صفحه] به زمان داده شده t هستند.

R‹Q(t,A_j)) = { V(P) e A¬VP)e At – tq

ما قصد داریم که راهبرد خزش را طوری تنظیم نماید که کیفیت نتایج بازیابی شده Qt (AS) را

در هر زمان پرسش t افزایش دهد .

کیفیت سنجه ها

در اینجا، به تعریف دو سنجه جامعیت(1) و انسجام میپردازیم که برای ارزشیابی کیفیت آرشیو به کار می روند. جامعیت جامعیت توانایی آرشیو را برای در بر گرفتن تمامی نسخه های کل سایتها را در مقایسه با تعداد کل نسخههایی که میتوان در یک خزش ایده آل ذخیره کرد، اندازه می گیرد.

توجه آرشیو زمانی دارای جامعیت است که تمامی کپیهای واقعی V(P) تمامی صفحه هایی را که

میتوان در یک خزش ایده آل ذخیره کرد، در برگیرد. بنابراین، هیچ نسخه ای از دست نرفته است.

انسجام انسجام سنجهای است که درجه مجموعه ای از نسخه های صفحه ها را برای انعکاس

وضعیت واقعی وبگاه بدون حضور اطلاعات متناقض می سنجد.

توجه: مجموعه ای از صفحات آرشیو شده زمانی دارای انسجام هستند که وضعیت واقعی را حداقل در یک نقطه از زمان منعکس کنند.

دو سنجه وزن دار و بدون وزن برای هر دو سنجه جامعیت و انسجام تعریف کرده ایم. سنجه های وزن دار اهمیت صفحه و تغییرات مرتبط را مورد توجه قرار میدهند؛ در حالی که سنجه های بدون وزن به نسبت تغییرات توجه دارند این سنجه ها به دلیل محدودیت جا در این مقاله به طور کامل توضیح داده نشده اند.

---

completeness -1

راهبرد خزش الگو محور

با توجه به محدودیت امکانات و ابزارهای قابل دسترس برای ذخیره صفحه ها هدف راهبردی خزش ما تعیین صفحه ها به گونه ای است که کیفیت (جامعیت و انسجام) آرشیو به حداکثر برسد. هر خزشگر میتواند کل M صفحه را در هر بازه زمانی T بارگذاری کند بر اساس الگوها، صفحه ها، براساس ارجحیت و یا ضرورت تعیین میشوند. به هر صفحه، یک ارزش ضرورت (UP متناسب با اهمیت تغییرات مورد نظر اختصاص داده میشود که توسط الگو در بازه زمانی T مشخص میگردد در هر بازه زمانی T M صفحه با بالاترین ارجحیت ذخیره میشوند ضرورت ،صفحه با گذشت زمان تغییر می کند. این ضرورت به زمان آخرین بازسازی و میانگین اهمیت تغییراتی بستگی دارد که توسط الگو مشخص میشود:

Pالگوی زیر را دارد:

t زمان جاری است.(te T)

W میانگین اهمیت تغییراتی است که توسط الگوی صفحه در بازه زمانی T است.

t آخرین بازسازی ، آخرین زمان بازسازی صفحه است.

a ضریب نرمال سازی است. از آنجا که واحدهای اندازه گیری W و زمان t متفاوت هستند، برای کاهش میزان تأثیر ضریب a معرفی شده است.

سپس،M صفحه منتخب در یک ترتیب نزولی و براساس میانگین اهمیت تغییر W مرتب و بارگذاری شدند. ریسک روز آمدسازی در طول خزش انسجام) ممکن است توسط ذخیره با اهمیت ترین صفحه ها در حال تغییر در آغاز هر بازه زمانی کاهش داده شود در حقیقت احتمال اینکه یک صفحه نامنسجم باشد، به جایگاه نسبی آن هنگام خزش سایت نیز بستگی دارد. بعد از آن هر صفحه ذخیره شده در زمان (t) با وضعیت قبلی آن صفحه در زمان (1) برای شناسایی تغییرات مقایسه شد. شاید اهمیت تغییراتی که بین دو نسخه شناسایی شدند، به بهره برداری از الگوی روز آمدسازی مربوط شود. به عنوان مثال اگر هیچ تغییری در بازه زمانی [] شناسایی نشود، میانگین اهمیت تغییر توسط الگو در زمان که به بازه [1] متعلق است، می تواند دوباره حساب شود در مقابل اگر تغییری بین دو نسخه ایجاد شود الگوی صفحه نمی توانسته مستقیماً روزآمد شود؛ زیرا ما دقیقاً نمی دانیم که در کدام زمان T تغییر حاصل شده است. به همین دلیل ممکن است جالب باشد که گاهی ابزاری را به نظارت بر تغییرات صفحه و روزآمدسازی الگو اختصاص دهیم.

آزمایش

ما به طور تجربی کارآیی رویکرد خزش پیشنهادی خود را از طریق مقایسه آن با راهبردهای مربوط ارزیابی کردیم آزمایشها بر روی دادههای ترکیبی برای ارزیابی عملکرد راهبردهای مختلف در شرایط آزمایشگاهی کنترل شده انجام شدند میزان امکانات و ابزارهای اختصاص داده شده، نسبت و اهمیت

تغییرات و مانند آن) به ویژه با استفاده از شبیه سازی میزان جامعیت و انسجام به دست آمده (که در بخش 2,3 توصیف شدند) از راهبردهای زیر را با هم مقایسه کردیم:

فراوانی در این ،راهبرد خزشگر، با توجه به فراوانی تغییرات ،صفحات آنها را برای آرشیو کردن انتخاب میکند فراوانی تغییرات هر صفحه بر اساس برآورد کننده Cho [12] ارزیابی شدند. سپس، صفحات منتخب بر اساس زمان آخرین بازسازی در یک ترتیب نزولی مرتب شدند و فقط M صفحه اول بارگذاری شدند. در مواردی که تعداد صفحات انتخابی از M کمتر ،بود خزش در بقیه صفحات غیر منتخب براساس آخرین زمان بازسازی انجام شد.

انسجام پیشرفته. این راهبرد با رویکرد پیشنهادی منبع [27] برای گسترش انسجام آرشیو مطابقت دارد. در این راهبرد خزشگر بر تقسیم بندی سایت کار میکند خزشگر مکرراً کل سایت را بارگذاری میکند در حالی که محدودیت صفحه ای را که در طول زمان ذخیره میشود در نظر دارد صفحه ها بر اساس احتمال تغییراتشان به ترتیب نزولی مرتب میشوند برای هر صفحه احتمال عدم انسجام در طول مدت خزش محاسبه میشود ،ابتدا صفحه ها با ریسک کمتر بارگذاری می شوند. سپس خزشگر کار ذخیره را در صفحه های جا افتاده ادامه میدهد.

شارک(1) این راهبرد با رویکرد منبع [16] پیشنهاد شده است که با گسترش میزان هشیاری آرشیو مطابقت دارد. همانند رویکرد انسجام ،پیشرفته خزشگر بر روی تقسیم بندی سایت کار میکند و مکرراً کل سایت را بارگذاری می.کند در این راهبرد صفحهها با توجه به میانگین نسبت تغییری (8) که توسط مدل Poission محاسبه میشود با ترتیب صعودی مرتب میشوند سپس صفحه هایی ذخیره می شوند که بیشترین تغییر را دارند تا حد امکان به وسط بازه مشاهده نزدیک هستند.

الگوها این راهبرد مبتنی بر الگوهایی است که رفتار تغییر را برای هر صفحه توصیف می کنند راهبرد حاضر، نخستین راهبرد پیشنهادی ماست که در آن الگوی هر صفحه به جای میانگین اهمیت تغییرات به نسبت تغییر – همانند Imp – Pattern – بستگی .دارد ،بنابراین این راهبرد اهمیت تغییرات بین صفحه ها را در نظر نمی گیرد.

الگوهای ایمپ(2). این دومین راهبرد الگو محور پیشنهادی است که در بخش 3/3 توضیح داده شد.صفحه ها براساس رفتار اهمیت تغییرشان که توسط الگوها تعریف میشوند ، بارگذاری میشوند.

تمامی این راهبردها به محدویتهای امکانات و ابزارها نیز بستگی دارند حداکثر M صفحه در هر زمان T ذخیره میشود تمامی آزمایشهای انجام شده در شرایط یکسان انجام گرفته اند. پیشرفت راهبرد الگوهای ایمپ را با در نظر گرفتن وزن انسجام و جامعیت آرشیو ارزیابی کردیم نتایج نشان دادند که راهبرد ما میزان جامعیت را در مقایسه با الگو 4 درصد، در مقایسه با شارک و انسجام پیشرفته 6 درصد و در مقایسه با فراوانی 25 درصد گسترش میدهد همچنین این راهبردها میزان انسجام را حدود 3 درصد در مقایسه با الگو شارک و انسجام پیشرفته و

1- SHARK

2- Imp- Pattern

40 درصد در مقایسه با فراوانی گسترش میدهد.

نتیجه گیری و پژوهشهای آینده

مقاله حاضـر به موضوع آرشیوسازی صفحات وب به طور کارآمد و به سازی کیفیت آن اشاره دارد. رویکرد پیشنهادی ما سه مفهوم را با هم تلفیق میکند بخش بندی صـفحه دیداری شناسایی تغییر، و اهمیت بلاکهای صفحه های وب برای تشخیص بهتر تغییرات مهم میان نسخه هـا رویکردهـای دیگر آرشـیو وب، فقـط براساس فراوانی تغییرات هسـتند. نخستین قـدم رویکرد ما ایجاد ساختار دیـداری سـند بر اساس بلاکهای معنایی خاصی است که الگوریتم VIPS را بسط میدهند [9] قدم دوم، شناسایی تغییرات بین آخرین نسخه صفحه آرشیو شده و نسـخه ماقبل آن است الگوریتم Vi-Diff که برای این بخش از کـار طراحی شـده از روشـهای کلی موجود برای ساختار دیـداری اسـناد مناسـبتر .است سـپس به موضوع ارزیابی اهمیت تغییرات پرداختیم برای اینکار راهبرد زمان بندی خزشـگر را طراحی و با اسـتفاده از آنالیز اهمیت تغییرات به بهینه سازی خزش وب پرداخته شـد. آزمایشـهای اولیه بر روی بخش بندیها و فازهای مختلف نشان دادند که زمان اجرا امیـدوار کننـده است. به هر حال زمان بخش بنـدی بسـیار بالاتر از زمان مقایسه است برای بهینه سازی بهتر ،سیسـتم باید بر روی کاهش مدت زمان بخش بندی توجه میکردیم. چالش اصـلی در این مقاله بهبود کیفیت آرشـیو .است. هدف ما این است که راهبردهای خزش را طوری تنظیم کنیم که خزش وبگاههای آینـده تا آنجا که ممکن است به طور جامع و منسـجم انجام شود. بدین منظور، نوعی راهبرد خزش مبتنی بر الگو را پیشـنهاد کردیم. یـک الگو رفتـار تغییرات صـفحه و اهمیت آنها را با گـذشت زمان در طول روز طراحی میکنـد دو سـنجه کیفیت جامعیت و انسـجام نیز برای ارزیابی راهبرد پیشـنهادی تعریف شـد سودمندی راهبرد الگو محور خود را با مقایسه آن با راهبردهای مرتبط نشان دادیم در شـرایط شبیه سازی شـده کامل جامعیت و انسـجام کلی هر راهبرد در آرشـیوسازی مقایسه شـد نتایج نشان دادند که راهبرد ما بهتر از راهبردهای موجود، جامعیت و انسـجام را بهبود میبخشد. با وجود ،این علاقه مندیم که جامعیت و انسجام به دست آمده را بیشتر گسترش دهیم. کار دیگری که در حال انجام است این است که ما راهبرد خود را بر روی صـفحه های وب واقعی بررسـی میکنیم. ما میخواهیم که الگوهایی را از صـفحه های وب رادیو و تلویزیون کشف کنیم و بعد از این الگوها در تنظیم خزشگرهای وب استفاده کنیم و کیفیت آرشیو را بالا ببریم.

منابع

[1] .The Web archive bibliography, http://www.ifs.tuwien.ac.at/aola/links/webarchiving.html

[2] S. Abiteboul, G. Cobena, J. Masanes, and G. Sedrati. A First Experience in Archiving the French Web. In ECDL '02: Proceedings of the 6th European Con- ference on Research .and Advanced Technology for Digital Libraries, 2002

[3] H. Artail and K. Fawaz. A fast HTML web page change detection approach based ,.on hashing and reducing the number of similarity computations. Data Knowl. Eng .2008 ,326–337:(2)66

[4] .M. Ben Saad and S. Gançarski. Using visual pages analysis for optimizing web archiving .In EDBT/ICDT PhD Workshops, Lausanne, Switzerland, 2010

[5] M. Ben Saad, S. Gançarski, and Z. Pehlivan. Archiving web pages based on visual (analysis and diff. In 25è me Journé es des Bases de Donné es Avancé es (BDA .Demonstration, Poster), Namur, Belgium, 2009)

[6] M. Ben Saad, S. Gançarski, and Z. Pehlivan. A novel web archiving approach based on ,visual pages analysis. In the 9th International Web Archiving Workshop (IWAW), Corfu .Greece, 2009

[7] B. Brewington and G. Cybenko. How dynamic is the web? In In World Wide Web .conference (WWW'2000), pages 257–276, 2000

[8] D. J. C. Lampos, M. Eirinaki and M. Vazirgiannis. Archiving the greek web. In 4th .International Web Archiving Workshop (IWAW04), Bath, UK, 2004

[9] D. Cai, S. Yu, J.–R.Wen, andW.–Y.Ma. VIPS: a Vision–based Page Segmentation

.Algorithm. Technical report, Microsoft Research, 2003

.C. Castillo and B. Sp. Scheduling algorithms for web crawling, 2004 [10]

W. Cathro. Development of a digital services architecture at the national library of [11]

.Australia. EduCause, 2003

J. Cho and H. Garcia–Molina. The Evolution of the Web and Implications for an [12]

Incremental Crawler. In VLDB '00: Proceedings of the 26th International Conference on

.Very Large Data Bases, 2000

J. Cho and H. Garcia–Molina. Estimating frequency of change. ACM Trans. Interet [13]

.Technol., 3(3), 2003

G. Cobena, S. Abiteboul, and A. Marian. Detecting changes in XML documents. In [14]
.ICDE '02: Proceedings of 18th International Conference on Data Engineering, 2002
C. N. Cosulschi M. and G. M. Classification and comparison of information structures [15]
.from a web page. In The Annals of the University of Craiova, 2004
D. Denev, A. Mazeika, M. Spaniol, and G. Weikum. Share: framework for ualityconscious [16]
.web archiving. Proc. VLDB Endow., 2(1):586-597, 2009
M. K. Evi, M. Diligenti, M. Gori, M. Maggini, and V. Milutinovi. Recognition of [17]
Common Areas in a Web Page Using Visual Information: a possible application in a page
classification. In the proceedings of 2002 IEEE International Conference on Data Mining
.ICDM'02, 2002
D. Gomes, A. L. Santos, and M. J. Silva. Managing duplicates in a web archive. In SAC [18]
.Proceedings of the 2006 ACM symposium on Applied computing, 2006 :06'
D. Gruhl, R. Guha, D. Liben-nowell, and A. Tomkins. Information diffusion through [19]
.blogspace. In In WWW '04, pages 491-501. ACM Press, 2004
X.-D. Gu, J. Chen, W.-Y. Ma, and G.-L. Chen. Visual Based Content Understanding [20]
towards Web Adaptation. In Second International Conference on Adaptive Hypermedia
.and Adaptive Web-based Systems (AH2002), 2002
R. La-Fontaine. A Delta Format for XML: Identifying Changes in XML Files and [21]
.Representing the Changes in XML. In XML Europe, 2001
E. Leonardi, T. T. Hoai, S. S. Bhowmick, and S. Madria. DTD-Diff: A change detection [22]
.algorithm for DTDs. Data Knowl. Eng., 61(2), 2007

C. Olston and S. Pandey. Recrawl scheduling based on information longevity. In WWW [23]

-Proceeding of the 17th international conference on World Wide Web, pages 437 :08'

.New York, NY, USA, 2008. ACM ,446

.Z. Pehlivan, M. Ben Saad, and S. Gançarski. Vi-diff: Understanding web pages changes [24]

In 21st International Conference on Database and Expert Systems Ap- plications

.DEXA'10), Bilbao, Spain, 2010)

.K. C. Sia, J. Cho, and H.-K. Cho. Efficient monitoring algorithm for fast news alerts [25]

.IEEE Transactions on Knowledge and Data Engineering, 19:950-961, 2007

R. Song, H. Liu, J.-R. Wen, and W.-Y. Ma. Learning block importance models for web [26]

pages. In WWW '04: Proceedings of the 13th international conference on World Wide

.Web, 2004

M. Spaniol, D. Denev, A. Mazeika, G. Weikum, and P. Senellart. Data quality in web [27]
archiving. In WICOW '09: Proceedings of the 3rd workshop on Information credibility
.on the web, pages 19–26, New York, NY, USA, 2009. ACM
Y. Wang, D. DeWitt, and J.–Y. Cai. X–Diff: an effective change detection algorithm for [28]
XML documents. In ICDE '03: Proceedings of 19th International Conference on Data
.Engineering, March 2003

چکیده

رشد فزایندۀ وب جهانگستر چالشـهایی را در حوزۀ حفاظت مؤثر داده های وب مطرح ساخته است. ابزارهایی که امروزه آرشیو سازان وب به کار میبرنـد، در وب به صورت کورکورانه خزش و بدین طریق بدون توجه به نوع صـفحه هایی که در دسترس قرار دارند (که به راهبردهای غیر بهینه(1) خزش منتهی میشود) و بدون توجه به ساختاری که محتوای صفحه هایی در خود دارند که منجر به گردآوری منابعی در سطح صفحه می شـود که محتوای آن را به دشواری می توان مورد بهره برداری قرار داد اقـدام به گردآوری میکننـد این نوشـته، معطوف به خزش و آرشـیوسازی برنـامه های کاربردی وبی به ویژه وب اجتماعی است که برای عموم قابل دستیابی.انـد برنامه کاربردی وب به هر نوع برنامه ای گفته میشود که از اسـتانداردهای ،وب نظیر HTML و HTTP برای انتشار اطلاعات در وب اسـتفاده کننـد و توسط مرورگر(2) های وبی قابل دسترسی.اند تالارهای گفت و گوی وبی شبکه های اجتماعی، خدمات مکانهای جغرافیایی(3) و مانند آن از آن جمله اند. ادعای ما این است که بهترین راهبرد برای خزش برنامه های کاربردی آن است که خزشـگر وب را طوری طراحی کنیم که نوع برنامه کاربردی و بی را که در حـال خزش آن است بشـناسد و بتوانـد فهرست یوآرال هایی را که در صف خزش قرار دارنـد پالایش نموده اطلاعاتی دربارۀ ساختار محتوای خزش شـده به آرشـیو ارائه دهـد. برای این ،کار ویژگیهایی به یک خزشـگر مختص آرشـیو وب میافزاییم که عبارت است از: توان تشخیص این که یک صفحه متعلق به چه نوع برنامه کاربردی وبی است و کاربرد روش شناسی خزش و برداشت متناسب با آن.

مقوله ها و توصیفگرهای موضوعی

کلیدواژه ها برنامه کاربردی ،وب ،آرشیووب ،خزش برداشت داده XPath

---

1- Suboptimal

2- Browser

3- Geolocation services

خزش هوشمند در برنامه های کاربردی وب

*خزش هوشمند در برنامه های کاربردی وب(1)

محمد فهیم زیر نظر پیر سنلار(2) | ترجمه فرزانه شادان پور(3)

1.مسئله

از زمان معرفی وب 2 وب اجتماعی(4) منبع مهمی برای برداشت(5) محتوا شده است. میلیونها کاربر از وب اجتماعی به عنوان وسیله ای برای انتشار اطلاعات بحث دربارهٔ موضوعات ،سیاسی به اشتراک گذاشتن محتوای ویدئویی ارسال نظرات مدیریت وب نوشت و بیان نظرات(6) شخصی خود دربارهٔ مباحث روز استفاده می کنند فقط کاربران عادی از وب اجتماعی استفاده نمی کنند، بلکه هر روز توجه رهبران سیاسی بیش از پیش به این پدیده جلب میشود. امروزه در امریکا و انگلستان پاسخ به پرسشهای پارلمانی با استفاده از توئیتر امری عادی شده است. به تازگی در 6 جولای ، 2011 باراک اوباما نام خود را به عنوان اولین رئیس جمهوری که از توئیتر به عنوان ابزاری برای ارتباط جمعی استفاده کرده است، ثبت نمود [13]. ،بنابراین وب اجتماعی به صورت بخشی از مبارزات سیاسی و هدایتگر برنامه آینده سیاسی درآمده است.

1- Intelligent crawling of Web applications for Web archiving

2- Muhammad Faheem, supervised by Pierre Sellenart

3- مربی عضو هیئت علمی سازمان سناد و کتابخانه ملی ایران

4- Social Web

5- Extraction

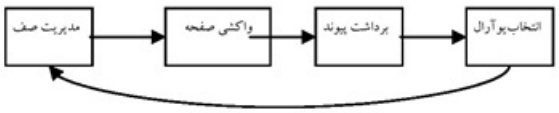
6- Comments

عکس

این مسئله، لزوم حفظ داده‌های اجتماعی را تشدید می‌کند.

ولی آرشیو کردن داده‌ها از وب اجتماعی به روشی هوشمند، هنوز چالشی برای اهل فنّ است. هدف ما پرداختن به این چالش با معرفی رویکردی انطباقی است که عمدتاً بر شناخت انواع گوناگون برنامه‌های کاربردی مورد استفاده در وب اجتماعی مبتنی است. برنامه کاربردی وب، به هر برنامه مبتنی بر HTTP گفته می‌شود که از وب جهانگستر برای انتشار اطلاعات استفاده می‌کند، در این نوشتار به‌ویژه بر وجوه اجتماعی وب تأکید خواهیم کرد زیرا، همچنان که به‌عنوان نمونه در تالارهای گفت‌وگوی وب[1]، وب‌نوشت یا توئیتر مشاهده می‌شود؛ تکیه زیادی بر محتوای تولیدشده توسط کاربران، تعاملات اجتماعی، و شبکه می‌شود.

برای آگاهی از اینکه چگونه توانمندی ابزارهای فعلی آرشیوسازی وب در حد کارکرد آرشیو کردن وب اجتماعی نیست، معماری ساده شدهٔ یک خزشگر وب سنتی (مثل هریتریکس[2] [۲۳]) را که در تصویر ۱ نشان داده شده است، در نظر بگیرید. یک خزشگر وب (که آن را عنکبوت یا روبات نیز می‌نامند) نوعی برنامه رایانه است که وب را به‌گونه‌ای روشمند وارسی می‌کند و مدارک مورد نظر را می‌یابد. خزشگرهای سنتی از حیث مفهومی به شیوه‌ای بسیار ساده در وب خزش می‌کنند. آنها کار خود را از یک فهرست یوآرال هسته[3] آغاز می‌کند که در صفی[4] نگه داری می شوند، (یوآرال ممکن است صفحه نخست یک ویگاه باشد). سپس، صفحه‌های وب یکی پس از دیگری از صف انتخاب و واکشی[5] می‌شوند. پیوندها [یا همان یوآرال‌ها]ی موجود در محتوای این صفحه‌ها واکشی شده استخراج و برداشت می‌شوند. اگر این پیوندها در دامنهٔ کار آرشیو قرار بگیرند، یو آر الهای تازه برداشت شده به صف اضافه می‌شوند. این فرآیند پس از زمان مشخصی، یا هنگامی که دیگر یوآرال مناسب دیگری در صف نباشد متوقف می‌شود.



تصویر ۱- زنجیرهٔ فرایند سنتی خزش در یک خزشگر وب

در رویکرد بالا، با چالش‌های خزش برنامه‌های کاربردی وب مواجه نخواهیم بود. ماهیت برنامه کاربردی وب که مورد خزش واقع شده است، یا محتوای مورد نظر، در تصمیم‌گیری‌های راهبردی

1. Web forums
2. Heritrix
3. Seed URLs (Uniform Resource Locator)
4. Queue
5. Fetch

این مسئله لزوم حفظ دادههای اجتماعی را تشدید میکند

ولی آرشیو کردن دادهها از وب اجتماعی به روشی هوشمند، هنوز چالشی برای اهل فن است. هدف ما پرداختن به این چالش با معرفی رویکردی انطباقی است که عمدتاً بر شناخت انواع گوناگون برنامه های کاربردی مورد استفاده در وب اجتماعی مبتنی است برنامه کاربردی ،وب، به هر برنامه مبتنی HTTP گفته میشود که از وب جهانگستر برای انتشار اطلاعات استفاده میکند، در این نوشتار به ویژه بر وجوه

اجتماعی وب تأکید خواهیم کرد زیرا همچنان که به عنوان نمونه در تالارهای گفت وگوی وب(1)، وب نوشت یا توئیتر مشاهده میشود تکیه زیادی بر محتوای تولید شده توسط کاربران تعاملات ،اجتماعی و شبکه می شود.

برای آگاهی از اینکه چگونه توانمندی ابزارهای فعلی آرشیوسازی وب در حد کارکرد آرشیو کردن وب اجتماعی نیست معماری ساده شده یک خزشگر وب سنتی (مثل هریتریکس(2) [23]) را که در تصویر 1 نشان داده شده است، در نظر بگیرید یک خزشگر وب (که آن را عنکبوت یا رویات نیز می نامند) نوعی برنامه رایانه است که وب را به گونه ای روشمند وارسی میکند و مدارک مورد نظر را می یابد. خزشگرهای سنتی از حیث مفهومی به شیوه ای بسیار ساده در وب خزش میکنند آنها کار خود را از یک فهرست یوآرال هسته(3) آغاز میکند که در صفی(4) نگه داری میشوند (یو آرال ممکن است صفحه نخست یک وبگاه باشد) ،سپس صفحه های وب یکی پس از دیگری از صف انتخاب و واکشی(5) میشوند پیوندها یا همان یوآرالهای موجود در محتوای این صفحه ها واکشی شده استخراج و برداشت میشوند. اگر این پیوندها در دامنهٔ کار آرشیو قرار بگیرند یو آرالهای تازه برداشت شده به صف اضافه میشوند. این فرآیند از زمان مشخصی یا هنگامی که دیگر یوآرال مناسب دیگری در صف نباشد متوقف می شود.

تصویر -1 زنجیره فرایند سنتی خزش در یک خزشگر وب

در رویکرد بالا- با چالشهای خزش برنامه های کاربردی وب مواجه نخواهیم بود. ماهیت برنامه کاربردی وب که مورد خزش واقع شده است یا محتوای مورد نظر در تصمیم گیریهای راهبردی

1- Web forums

2- Heritrix

3- Seed URLS (Uniform Resource Locator)

4- Queue

5- Fetch

خزش، مـدنظر گرفته نمی شود. برنـامه های کاربردی وب با محتوای پویا(1) (مثـل تالارهای گفت وگوی وب، وب نوشت ها و ماننـد آن) ممکن است به گونه ای ناکارآمـد خزش شونـد؛ اگر محتوا به گونه ای باشـد که فقـط با عملیات پیچیـده خزش (مثل پرس و جوی AJAX ارسال فرم) قابل دسترس باشد، بعضی از بخشهای آن ممکن است از دست برود.

به عنوان نمونه تالارهای گفت و گوی وب دارای ویژگی پویایی هسـتند به این معنا که برای برداشت محتوای معنایی یا بهبود عملکرد خزش در مورد آنها ماهیت آنها بایـد شـناخته شود. محتوای تالارهای گفت و گوی وب اغلب در یک پایگاه داده نگهداری می.شود. هنگامی که یک کاربر، پرس و جویی ارسال می کند صفحهٔ پاسخ به صورت خودکار و با استفاده از یک قالب(2) از پیش تعریف شده تولید می شود. وقتی دو پرس و جو بخش واحـدی از محتوای چنین صـفحه ای را تقاضا میکننـد سـرور دو صـفحهٔ پویا با محتوای یکسان یا مشابه، ولی با یوآرال مختلف تحویل کاربر می دهد. ولی این صفحه های پویا موجب پدیده افزونگی(3) میشوند که ممکن است زیان آور باشد؛ از این جهت که موجب میشونـد منابع بیشتری برای خزش وجود داشـته باشـند و در نتیجه آرشـیو نهایی از کیفیت خوبی برخوردار نخواهـد بود. خـدمات وب نوشت نیز اطلاعات مکرر در خود دارنـد به عنوان مثال به صورت ماهانه و سالانه آرشـیو می شونـد و دربردارنده محتواهایی هسـتند که دارای تغییرات جزئی است و مکرر محسوب میشود در صورت خزش تالارهـای گفت و گو و وب نوشت هـا بـا رویکرد سـنتی ،خزش بـا موارد بسـیار زیـادی از منابع مکرر مواجه میشویم در نهایت خزشـگر در تله خزش(4) گیر میافتـد چرا که بایـد بینهایت پیوند را خزش کنـد. همچنین پیوندهای دارای اختلال(5) نظیر صـفحه های مناسب ،چاپ یا ،تبلیغات و ماننـد آنها وجود دارنـد که بهتر است از گردآوری و ذخیره آنهـا هنگـام ایجـا آرشـیو خودداری کرد. با رویکردهای سـنتی خزش، از برنامه های کاربردی وب نیز که در حـد بالایی پردازه نویسی(6) شده اند یا از وب عمیق(7) (داده های قابل دسترس از ورای فرمها) داده نمیتوان برداشت کرد.

دست آخر اینکه خزشـگرهای آرشـیو وب که در رویکرد سـنتی طراحی شده اند، برای برداشت داده به طرق گوناگون تلاشی نمیکنند؛ حال آنکه آرشـیو داران و کاربران آرشـیوهای وب دوست دارند به ابعاد معنایی بیشتری دربارهٔ محتوای آرشیو نظیر برچسبهای زمان(8) پیامهای وب نوشـتها، و توصـیفگرهای نویسـندگان دست یابند حتی اگر این کار با اسـتفاده از ناوبری های(9) پیچیـده در صـفحه ها انجام شود که اطلاعـات مربوط از صـفحه ها مختلف را سازمان می دهنـد. به عنوان مثال، یک برنامه کاربردی وب که محتوای خود را به گونه ای سامان میدهد که برای برداشت اطلاعات مورد نظر مرور تقویم ضروری باشد،

1- Dynamic

2- Template

3- Redundancy

4- Spider trap

5- Noisy

6- Scripting

7- Deep Web

8- Timestamps

9- Navigation

یـا در مورد یک تالار گفت و گوی وب جایی که یک روند(1) میتواند ورودیهای مختلفی متشـکل از صفحه ها مختلف داشته باشد و برای برداشت این اطلاعات ناوبری مؤثر صـفحه ضروری باشد رویکرد سنتی خزش برای انجام کارآمد این فرآیندها با محدودیتهایی مثل صرف زمان بیشتر در خزش برای تعداد اندکی صفحهٔ مناسب بدون در برداشتن اطلاعات معنایی در محتوا، مواجه خواهد بود.

در ادامه به طور خلاصه وضـعیت فعلی فناوری خزش در وب را بررسی میکنیم سپس در بخش3 رویکرد پیشنهادی ارائه میشود که عبارت است از وارد کردن یـک راهنمای آگاه از برنامه کاربردی(2) در فرآیند خزش که به خزشـگر در تمام فرآیند خزش کمک و خزش مؤثر داده ها را تضـمین میکند. در بخش 4 با توصیف الگوهای اکتشاف(3) برنامه کاربردی وب پیشـنهادی مان که ساختاری برای پایگاه دانش(4) یک برنامه کاربردی وب است میکوشـیم تا روش شناسی خود را به تفصیل شرح دهیم. همچنین توضیح خواهیم داد که چه نوع از عملکرد خزش مورد نظر ماست. در بخش ه نتایج اولیه را برخواهیم شمرد و مقاله را با بحثی دربارهٔ پژوهشهای آینده به پایان میبریم.

2.وضعیت فعلی فناوری خزش

خزش در وب مسـئله ای است که مطالعات خوبی درباره آن صورت گرفته، ولی همچنان چالش برانگیز است. ژولین ماسانه(5) در [18] مروری بر حوزهٔ آرشیو وب و خزش برای ایجاد آرشیو وب انجام داده است. او به ویژه دربارهٔ خزش وب عمیق بحث کرده و بر نیاز به آرشیو داده از سطح و عمق وب برای حفاظت وب تأکید میکند.

یک خزشگر کانونی(6) بر اساس مجموعه ای از موضوعات از پیش تعیین شده خزش می کند [7]. در این روش رفتار خزشگر نه بر مبنای ساختار برنامه کاربردی- وب که هـدف ماست - بلکه بر مبنای محتوای صـفحه های وب تنظیم میشود در رویکرد ما هـدف خزش کانونی نیست، بلکه جـای آن را رویکرد بهـتری برای برنـامه هـای کـاربردی وب میگیرد. هر دو راهبرد، روشـهای تکمیـل کننـده ای برای ارتقـای عملکرد خزشگر سنتی به حساب می آیند.

در یک برنامه کاربردی وب یا یک سامانه مـدیریت محتوا محتوا با توجه به یک قالب(7)(اجزای قالب به عنوان مثال شامل نوار سمت چپ یا راست صفحهٔ ،وب نوار ناوبری ،صفحه سرصفحه و پاصفحه منوی اصلی و... است).

از میان آثار متعدد دربارهٔ برداشت قالب(8)، گیبسون پونرا و تامکینز [10] زمینهٔ محتوای مبتنی بر قالب را در وب مورد بررسی قرار داده اند آنها دریافته اند که 40 تا 50 درصد محتوای وب در (2005) مبتنی

1- Thread

2- Application-aware helper

3- Detection patterns

4- Knowledge base

5- Julien Masanes

6- Focused or goal-directed crawler

7- Template

بر قالب است یعنی بخشی از یک برنامه کاربردی وب است. یافته های آنها همچنین نشان میدهد که صفحه های وب با نرخی معادل 6 تا 8 درصد در سال در حال رشد هستند. این پژوهش با قوت به منافع خزشگری که بتواند به شیوه ای خاص برنامههای کاربردی وب را خزش کند اشاره دارد.

گرچه جایی که ما میدانیم هنوز خزش آگاه از برنامه های کاربردی به صورت عام مورد توجه قرار نگرفته است، تلاشهایی برای برداشت محتوا از تالارهای گفت وگوی وب صورت پذیرفته است [6، 11]. اولین مورد که خزش تالار گفت و گو(1) نامیده میشود ،[11] ساختار سازمان یافته تالارهای گفت و گوی وب را هدف خزش قرار میدهد و رفتار کاربر را در فرآیند برداشت اطلاعات شبیه سازی میکند. خزش تالار گفت و گو با مسئله به طور بسیار مؤثری سروکار پیدا میکند ولی هنوز دچار محدودیتهایی است که ناشی از داشتن قواعدی ساده است و تنها میتواند در مورد تالارهایی با ساختار ساماندهی مشخص به کار رود رویکرد دوم [6] وابسته به ساختار تالار گفت و گوی وب نیست در این روش سامانه iRobot فرآیند برداشت را با فراهم کردن نقشه سایت(2) برنامه کاربردی وب که باید خزش شود، همراهی می کند. نقشه سایت از طریق خزش تصادفی چند صفحه از برنامه کاربردی ساخته میشود. این فرآیند به شناسایی مناطقی که به شدت تکرار پذیرند کمک و بعد آنها را بر مبنای آرایششان خوشه بندی می کند [25]. بعد از تولید نقشه سایت iRobot ساختار تالار گفت و گوی وب را در شکل یک گراف جهت دار(3) متشکل از رئوس(4) (صفحه ها وب) و یالهای(5) جهت دار (پیوندهای میان صفحه ها مختلف وب) می سازد. علاوه بر این مسیر نیز به منظور فراهم کردن بهترین مسیر عبور که فرآیند برداشت را هدایت کند، و از واکشی صفحه های مکرر و بی ارزش نیز جلوگیری نماید مورد تجزیه و تحلیل قرار می گیرد. هدف ما توسعه فناوریهای مشابه برای برنامههای کاربردی اختیاری(6) بوده است و نه فقط برای تالارهای گفت و گوی وب در اینجا توضیح این نکته لازم است که رویکرد ما برای کشف نوع برنامه کاربردی وب که خزش شده است، بر مکانیسمی عام(7) استوار است آثاری دربارۀ وب نوشت ها و تالارهای گفت و گوی خاص نگارش یافته اند. در [14] به ویژه از مدل SVM(8) برای کشف صفحه ای که متعلق به یک وب نوشت است استفاده شده .است 5 19] SVM] برای دسته بندی متون بسیار مورد استفاده قرار می گیرد. در [14] SVMها با استفاده از بردارهای مختلف سنتی که از دسته های(9) واژگان یا n-grams موجود در در محتوا تشکیل شده اند آموزش داده میشوند. پاره ای از ویژگیهای(10) جدید نیز برای کشف وب نوشت، نظیر دسته های یو آرال پیوند داده شده و دسته های گرانواژه(11) معرفی می شوند. برای انتخاب ویژگیها

1- Board Forum Crawling

2- Sitemap

3- Directed graph

4- Vertices

5- Arcs

6- Arbitrary

7- Generic mechanism

8- Support Vector Machine

9- Bags

10- Features

از آنتروپی(1) نسبی استفاده می شود. براساس نتایجی که در این نوشته آمده است، شناسایی وب نوشت با استفاده از ویژگیهایی مرکب از یوآر الها گرانواژه ها و ابر برچسبها بهتر انجام شده است. با این حال در موضوع کشف برنامه های کاربردی وب روشهایی برای کشف منابع وب پنهان وجود دارد [3 و 4]از جمله تجربیات در این زمینه خزشگر فرم کانون(2)- [3] برای کشف پایگاههای اطلاعاتی برخط بوده است. در این ،روش از یک خزشگر کانونی برای بارگذاری منابع در موضوعات مورد نظر به کمک یک برنامه دسته بندی پیوند(3)، استفاده میشود برنامه دسته بندی ،پیوند فرمهای قابل جست وجو را کشف و پیوندهای موجود در آنها را الویت بندی میکند در این روش محدودیتهایی نیز وجود داردکه از آن جمله اند: تنظیم دستی وابستگی به برنامه دسته بندی پیوندها و نتایج ناهمگن. در روشی که اخیراً به کار رفته [14] و سازگاریافته تر است خزشگر سازگار یافته برای منابع وب پنهان در مورد این محدودیتها ملاحظاتی صورت گرفته است در این روش نقاط ورود به وب پنهان به گونه ای کارآمد و با الگویی ناشناخته اکتشاف میشود و این تجربه به طور خودکار به فرآیند یادگیری افزوده می شود.

آثاری [1، 15، 16] نیز به طور عام معطوف به شناسایی دسته بندیهای عام (مثل وب نوشت، وبگاههای دانشگاهی شخصی و مانند آن) برای یک وبگاه با استفاده از برنامه دسته بندی مبتنی بر ویژگیهای ساختاری صفحه وب است که میکوشد تا کارکرد پذیریهای(4) این صفحه های وب را کشف کند این امر مستقیماً در تنظیمات کار ما قابل به کارگیری نیست؛ نخست به این علت که این متون در سطح وبگاه قابل به کارگیری اند (مثلا) مبتنی بر صفحه نخست(اند و نه در سطح صفحه های منفرد کاملا روشن است که همهٔ این فنون شناسایی برنامه های کاربردی بر برنامههای دسته بندی آموزش دیده متکی هستند. این مسئله میتواند یکی از جهات ممکن برای مکانیسم کشف برنامهٔ کاربردی وب در این پژوهش باشد.

3.رویکرد پیشنهادی

مدعای اصلی ما این است که برای انواع برنامه های کاربردی وب فنون خزش متناسب و متفاوتی به کار برده شود؛ یعنی داشتن راهبردهای خزش متفاوت برای انواع وبگاههای اجتماعی (وبنوشتها، ویکیها، شبکه های اجتماعی نشانکهای اجتماعی(5)، ریز بلاگها(6) ، شبکه های موسیقی(7)، تالارهای گفت و گوی ،وب، شبکههای عکس(8)، شبکه های ویدئویی(9)، و مانند آن) برای سیستمهای ویژه مدیریت محتوا (مانند Wordpress, php BB) و برای سایتهای خاص مانند توئیتر و فیس بوک در رویکردی که ما در این پژوهش در پیش گرفته،ایم نوع برنامهٔ کاربردی وب (نوع به طور کلی سیستم مدیریت محتوا، یا

1- Relative entropy

2- Form-focused crawler

3- Link classifier

4- Functionalities

5- Social bookmarks

6- Microblogs

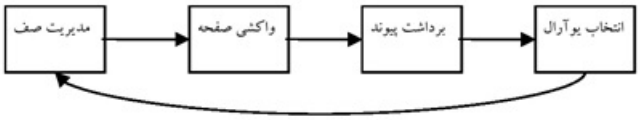
7- Music networks

8- Photo networks

عکس

سایت) حین عملکرد خزش ، و نوع صفحهٔ وب درون برنامهٔ کاربردی (مانند پروفایل کاربر در یک شبکهٔ اجتماعی) کشـــف و در مورد عملیات مورد نیاز برای خزش بعدی (تعقیب پیوند، استفاده از یک API، ارسال فرم، برداشت محتوای ساختارمند) تصمیم‌گیری می‌شود. برای سازگار کردن یک خزشگر سنتی با الزامات و نیازمندی‌های این پژوهش، ما معماری سنتی یک خزشگر وب را آن گونه که در شکل ۲ نشان داده شده اســـت بسط دادیم. ماژول واکشی[1] صفحه (نگاه کنید به شکل ۱) به ماژول واکشی دقیق‌تری بسط داده شده است که قادر است منابعی را بازیابی کند که با پرس‌وجوی HTTP GET قابل دسترسی نیســـتند (بلکه با یک رشته پرس و جو، یا با استفاده از یک API)، یا اشیای وبی منفردی در یک صفحهٔ وب هستند (مانند یک پست وب‌نوشت، یک دیدگاه[2]، یا نام یک پست گذار[3]). در واقع، به‌جای کارکرد معمول برداشت پیوند، ماژول راهنمای آگاه از برنامه کاربردی به‌کار برده می‌شود تا در اثنای خزش بتواند برنامهٔ کاربردی را شناسایی کند و ضمن تقسیم‌بندی عملکردهای خزشی، که می‌توان برای هر برنامه انجام داد، در این مورد تصمیم‌گیری کند.

این تغییرات در دو خزشـــگر اعمال می‌شوند: خزشگر تحت مالکیت بنیاد حافظهٔ اینترنت[4]، که ما با آنها همکاری نزدیکی داریم؛ و نسخهٔ سفارشی خزشگر هریتریکس[5] [۲۳]، که توسط آزمایشگاه پژوهشی ATHENA در چارچوب طرح ARCOMEN [۲] تهیه شـــده است. در بخش بعدی این نوشتار ماژول راهنمای آگاه از برنامه‌های کاربردی با جزئیات بیشتری تشریح می‌شود.



شکل ۲- معماری بسط داده شدهٔ خزشگر

٤. روش‌شناسی

در این بخش، ماژول راهنمای آگاه از برنامه‌های کاربردی معرفی می‌شود. این ماژول خزشگر آرشیو وب را در فراهم‌آوری محتوا از وب اجتماعی به شیوه‌ای هوشمند و سازگاریافته یاری می‌رساند. این ماژول قابلیت‌های کارکردی خزشگر را افزایش می‌دهد و فرآیند خزش را کارآمدتر می‌کند.

پایـــگاه دانش در برنامه کاربردی وب. خزشـــگر با پایگاهی از دانـــش در مورد برنامه‌های کاربردی وب یاری می‌شودکه در آن نحوهٔ خزش هوشمند وبگاه‌ها توصیف می‌شود. این پایگاه دانش چگونگی اکتشاف برنامه‌های ویژهٔ کاربردی وب و عملکرد خزشی را که برای آن لازم است به اجرا گذاشته شود،

1. Fetching
2. Comment
3. Poster
4. Internet Memory Foundation
5. Heritrix

سـایت) حین عملکرد خزش و نوع صـفحه وب درون برنـامه کـاربردی (ماننـد پروفایـل کـاربر در یک شـبکه اجتماعی) کشف و در مورد عملیات مورد نیاز برای خزش بعدی تعقیب ،پیوند استفاده از یک API ارسال فرم برداشت محتوای ساختارمند تصمیم گیری میشود برای سازگار کردن یک خزشـگر سـنتی با الزامات و نیازمندیهای این پژوهش ما معماری سنتی یک خزشگر وب را آن گونه که در شکل 2 نشان داده شده است بسط دادیم. ماژول واکشی(1) صفحه (نگاه کنید به شکل 1)به ماژول واکشـی دقیقتری بسط داده شده است که قادر است منابعی را بازیابی کنـد که با پرس وجوی HTTP GET قابل دسترسـی نیسـتند بلکه با یک رشـته پرس و جو یا با استفاده از یک (API) یا

اشـیای و بی منفردی در یک صفحهٔ وب هستند (مانند یک پست وب نوشت، یک دیدگاه(2)، یا نام یک پست گذار(3)). در واقع، به جای کارکرد معمول برداشت پیوند ماژول راهنمای آگاه از برنامه کاربردی به کار برده میشود تا در اثنای خزش بتواند برنامه کاربردی را شـناسایی کند و ضمن تقسیم بندی عملکردهای خزشی که میتوان برای هر برنامه انجام داد، در این مورد تصمیم گیری کند.

این تغییرات در دو خزشـگر اعمال میشوند خزشگر تحت مالکیت بنیاد حافظه اینترنت(4)، که ما با آنها همکاری نزدیکی داریم؛ و نسـخهٔ سفارشی خزشگر هریتریکس(5) [23] که توسط آزمایشـگاه پژوهشی ATHENA در چـارچوب طرح ARCOMEN [2] تهیه شده است. در بخش بعدی این نوشتار ماژول راهنمای آگاه از برنامههای کاربردی با جزئیات بیشتری تشریح میشود.

شکل -2 معماری بسط داده شده خزشگر

4.روش شناسی

در این بخش مـاژول راهنمـای آگـاه از برنـامه های کاربردی معرفی میشود این ماژول خزشـگر آرشـیو وب را در فراهم آوری محتوا از وب اجتماعی به شـیوهای هوشـمند و سازگاریافته یاری میرسانـد. این ماژول قابلیتهای کارکردی خزشـگر را افزایش میدهـد و فرآیند خزش را کارآمدتر میکند.

پایگـاه دانش در برنـامه کـاربردی وب خزشـگر بـا پایگـاهی از دانش در مورد برنـامه های کاربردی وب یاری میشود که در آن نحوهٔ خزش هوشـمند وبگاهها توصیف می.شود این پایگاه دانش چگونگی اکتشاف برنامه های ویژه کاربردی وب و عملکرد خزشی را که برای آن لازم است به اجرا گذاشته شود،

1- Fetching

2- Comment

3- Poster

4- Internet Memory Foundation

5- Heritrix

مشـخص می.کند پایگاه دانش از تقسـیم بندیهای عمومی گرفته تا موارد خاص (وبگاهها) یک برنامه کاربردی به شیوه سلسله مراتبی تنظیم میشود برای مثال وبگاههای شـبکه های اجتماعی را میتوان به وب نوشـتها تالارهای گفت و گوی ،وب.ریز بلاگها شـبکه های ،ویدئویی و ماننـد آن تقسـیم بندی کرد غیر از این میتوان برنامه های کاربردی را براساس سامانه های مدیریت محتوا به گونه دیگری تقسـیم بندی .کرد. مثلاً Wordpress و Movable Type نمونه هایی از سامانه های مدیریت وب نوشت هسـتند. phpBB و vBulletin نیز نمونه هایی از سامانه های مدیریت محتوای تالارهای گفت وگوی وب.

علاوه بر این هر برنامه کاربردی وب معمولاً از انواع مختلف صـفحه وب تشکیل می شود. در یک تالار گفت و گوی ،وب صفحه ها وجود دارنـد که فهرستی از تالارها را ارائه میکننـد صـفحه هایی که فهرستی از پسـتهای مربوط به تالارهای گفت و گوی خاص را نشان میدهنـد، صفحه هایی که به پسـتهای منفرد، همراه با دیـدگاههای آنها ارجاع می دهنـد. بنابراین، پایگاه دانش، انواع مختلف صـفحه های وب تحت یک برنامهٔ کاربردی را توصیف میکند و بر این مبنا میتوانیم عملکردهای مختلف خزشی را که باید در قبال هر نوع از صفحه به اجرا بگذاریم – تعیین کنیم.

پایگاه دانش باید در زبانی روان و گزاره ای بیان شود تا به آسانی قابل روز آمدسازی و اشتراک باشد ناآشـنایان به برنامه نویسـی هم بتوانند با آن کـار کننـد و حتی اگر بشود از مثالهـا به طور خودکار یادبگیرنـد کنسرسـیوم وب جهانی(1)، نوعی زبان توصـیف برنامـهٔ کاربردی به نام WADL(2) [12] را تهیه و برای بهره برداری ارائه کرده است که توصـیف منابع سازگار با انتقال در پروتکل HTTP را در قالبی با قابلیت پردازش ،ماشینی ممکن می سازد WADL برای توصیف مجموعه های منابع روابطشان با یکدیگر روشی که برای هر منبع باید به کار رود و قالبهـای نمایش منبع به کار میرود WADL را می توان به عنوان جزء قابـل کـاربرد و قـالب ارسال برای پایگاه دانش به کار برد، ولی همه نیازهای ما را مرتفع نمی سازد؛ از جمله توصیف الگوهای بازشناسی برنامه کاربردی وب و تعاملات برنامه کاربردی که ورای یک تقاضای GET و POST ساده جریان دارند در نتیجه پایگاه دانش ما باید در قالب XML توصیف شود و به خوبی با ساختار درختی سلسله مراتب برنامه کاربردی وب و سطوح مختلف صفحه ها انطباق یابد.

مـاژول کشف برنـامه کـاربردی وب تشـخیص برنامه کاربردی وب و پس از آن در پیش گرفتن بهترین راهبرد خزش متناسب با آن مهمترین چالش در خزش و برداشت محتواست. در مورد شـناسایی برنامه های کاربردی وب کار زیادی انجام نشده ولی تلاشهایی برای دسته بندی صفحه هـا وب تحت برنـامه های کاربردی وب مختلف صورت گرفته است ،[11 ، 15 ، 16] برای کشف یک برنامه کاربردی وب خاص پایگـاه دانش بـا توصـیف قواعـد مختلف برمبنای الگوهای یوآرال فراداده های HTTP، محتوای متنی، الگوهـای xPath(3) ارجاعات به سامانه دسته بندی و در صورت امکان ویژگیهای مبتنی بر گراف

1- World Wide Web Consortium

2- Web Application Description Language

3- زبان مسیر ،xml زبان پرسشی برای انتخاب گره ها از سند xml است که توسط WWC تعریف شده است.

وب(1) را میسر می سازد. شناسایی سطح صفحه درون هر برنامه کاربردی وب ممکن است با طبقه بندی صفحه متناسب با ویژگیهای ساختاری انجام شود.

بد نیست در اینجا سامانه مدیریت محتوای تالار گفت وگوی وب موسوم به vBulletin را ذکر کنیم که به عنوان نمونه از طریق جست و جوی یک ارجاع به پردازه نوشته(2) جاوای vbulletin_global.js با استفاده از عبارت xPath ساده //script/@src قابل شناسایی است.

صفحه های سطح «فهرست تالار گفت و گو هنگامی که با عبارت xPath //a[@class"forum"]@href جور در بیایند قابل شناسایی خواهند بود.(3)

خزش و برداشت. بعد از کشف برنامه کاربردی که صفحه وب متعلق به آن است، مرحله بعدی این است که عملکردهای مناسب خزش را تعیین کنیم دامنهٔ عمل خزش فراتر از افزودن فهرست یوآرال به صف یوآرالهای در دست خزش است. این امر شامل هر عملکردی است که در فرآیند خزش به نحوی دخیل است استفاده از API برای برداشت دادههای مرتبط از سایتهای شبکه های اجتماعی نظیر توئیتر یا انجام تعاملات پیچیده با برنامه های کاربردی برمبنای AJAX یا شناسایی موجودیتهای وبی، به ویژه برنامه های کاربردی وب عملکردهای خزش به معنای خاص آن به دو نوع اند:

عملکردهای ناوبری ناوبری به یک صفحه وب یا منابع وبی.

عملکردهای برداشت برداشت موجودیتهای معنایی منفرد از صفحه های وب (مانند برچسبهای

زمان(4)، پستهای وب نوشت، دیدگاهها).

مشابه همین ما یک زبان اعلانی(5) برای توصیف همهٔ عملکردهای خزش (مطلوب است یک پایگاه دانش که به سادگی قابل نگهداری ،باشد از جمله نگهداری ماشینی داشته باشیم.) میخواهیم بنابراین به یک زبان برنامه نویسی برای ناوبری و برداشت اطلاعات نیاز داریم که قادر باشد به داده های وب عمیق نیز دست یابد همان گونه که به یوآرالهای معمولی دسترسی پیدا میکند.

ما از oxPath[9] استفاده میکنیم که بسط یافته xPath است و دارای امکاناتی برای تعامل با برنامه های کاربردی وب و برداشت دادههای مرتبط است. این ،زبان شبیه سازی عملکردهای کاربر را در تعامل با رابطهای کاربری چند صفحه ای پردازه نویسی شده(6) برنامه های کاربردی وب (ارزیاب(7) یا با مرورگر موزیلا- و یا با مرورگر Webkit- کار میکند) را ممکن می سازد ،oxPath خصوصیاتی مشابه xPath دارد و استفاده از گزینشگرهای مبتنی بر الگوهای آبشاری(8) با آن میسر است و میتوان با آن در چندین صفحهٔ مختلف با کلیک ناوبری کرد و حتی از صفحه ها قبلی اطلاعات برداشت پیاده سازی شده

1- Web-graph-based-features

2- Script

3- مثال برای ارائه ساده شده است. ولی در واقع ما با چندین چینش (layout) سروکار داریم که vBulletin می تواند تعریف کند.

این برنـامه بـا کـد منبع باز در دسترس است که در سامانه ما نیز مورد اسـتفاده قرار گرفته است. نمونه ای از یک عملکرد ساده oxPath که میتوان آن را بر vBulletin اجرا کرد عبـارت است از: /click //@href/forum.a/ که به هر پیونـد به تالاـر گفت وگوی وب کلیک میکند برای نمونههای بهتر از متغیرهای oxPath به [9] نگاه کنید.

همچنین چندین جایگزین برای oxPath را بررسـی خواهیم کرد؛ به ویژه مواردی که در شماره های 17، 20، 21 22 23 منابع این مقاله ذکر شده.اند این روشها به جز مورد مطرح در شماره [24] با تعامل با وب ،پنهان نظیر ارسال فرم و ناوبری در صـفحه ها کاری ندارند. در [22] از Datalog به عنوان زبان برنامه نویسـی برای برداشت داده از صـفحه ها وب استفاده شده است. در این روش زبان Xlog به عنوان برنامه کاربردی Datalog که شامـل محمولات از پیش تعریف شـده ای در برداشت داده است معرفی می شود. این ،روش پنجره ای برای محققان در استفاده از Datalog به عنوان پایه ای برای فرآیند برداشت می،گشاید ولی هنوز تلاشـهای زیادی تا وب پنهان در آن لازم است [21] نیز با همین چالش روبه روست در این ،روش محتوا از یک صـفحهٔ ساده یا از صـفحه های کتابشناختی برداشت می.شود نیز در این روش هیچ عملی برای پر کردن فرم یا ناوبری در صفحه ها شبیه سازی نمی شود و در آن از یک زبان با نام Wraplet استفاده شده است که دادههای ساختارمنـد را از صـفحه ها وب در قالب HTML برداشت می.کنـد این ،زبان با عبارتهای پردازه نویسـی Wraplet عبارتهای برداشت (داده نوشـته شده است و یک سند HTML را به عنوان ورودی گرفته خروجی را در XML میسازد. متأسفانه این روش فقط برای صفحه ها تک قابل استفاده است که دارای ویژگی پویایی وب نیستند روشی که از حیث امکانات و کارکرد، نظیر ارسال فرم و ناوبری بسیار به oxPath شبیه است در [24] معرفی شده است. در نوشـته اخیر نمونه های مختلفی برای روشن شدن مفاهیم ارائه شده اند. نویسندگان در آن سامـانه ای را معرفی کرده انـد و آن را سامانه برداشت داده مرور–گرا(1) نامیده اند که اطلاعات را از صـفحه های وب برداشت و از پیوندها برای ناوبری به صـفحهٔ بعدی برای برداشت اطلاعات استفاده می.کند سامانه مذکور برنامه های کاربرد را حتی هنگامی که در حال انجام کارکردهای پردازه نویسـی ماننـد جاوا یا AJAX هسـتند گردآوری میکنـد تا به محتوای صـفحهٔ بعـدی دست پیـدا کند این سامانه همچنین عملکردهـای کـاربران را برای تعامل با وب پنهان شبیه سازی می.کنـد علیرغم همـهٔ این مزایا این سامانه محـدودیتهای ناشـی از مـدیریت حافظه را مـد نظر نمی گیرد و به همین علت برای سامانه مورد نیاز ما که لازمهاش خزش و آرشـیو کردن مداوم مقادیر عظیم داده است مناسب نیست در این سامانه برای ناوبری متعدد صـفحه ها کار مرورگر تکرار میشود تا عملکرد بهینه شود. در سامانه ای که ما تدارک دیـده ایم با برداشت داده در مقیاس وسـیع سـروکار داریم. همین علت به سامانه ای نیاز داریم که مراقب عملکرد و مدیریت حافظه باشـد xPath از حیث حافظه زمان به خوبی از عهده این حجم کار بر می آید.

1- BODE (Browser-Oriented Data Extraction System)

5.نتایج

از آنجا که این رساله دکتری تنها چند ماه پیش آغاز شده است، ما بر توسعه یک معماری متمرکز بودیم که روش شناسی آن در قسمتهای قبل ذکر شد. با این پیش فرض که این کار شدنی است و با بازشناسی آن در دو برنامه کاربردی وب، یک پیش نمونه اولیه از راهنمای آگاه از برنامه کاربردی وب پیاده سازی شد. بدین ترتیب که مثلا پیش نمونه در مواجهه با برنامه کاربردی vBulletin و وبگاههای متعددی که از این سامانه مدیریت محتوا استفاده میکردند مورد ارزیابی قرار گرفت. در حال حاضر سامانه قادر است نوع برنامه کاربردی وب و سطح [صفحه] در برنامه کاربردی را تشخیص دهد و عملکردهای مناسب خزش را به اجرا بگذارد. به تازگی سامانه 8] Yfilter] (سامانهٔ فیلترینگ مبنی بر NFA(1)) را برای نمایه سازی کارآمد الگوهای اکتشاف و به منظور یافتن برنامه های کاربردی وب ،مرتبط به سامانه ملحق کرده ایم هم اکنون، سامانه قادر است صفحه ها مرتبط بیشتری را با صرف حداقل ،توان در مقایسه با روشهای قبلی خزش، برداشت .کند هنوز لازم است بتوانیم عملکردهای خزش را با استفاده از یک ارزیاب oxPath به منظور تبدیل آنها یوآرال یا عبارتهای xPath ، در صورت امکان به اجرا درآوریم و راهنمای سامانه را رابط خزشگر قرار دهیم ولی نتایج اولیه برای پژوهشهای آینده راضی کننده و امیدبخش بوده اند.

5.پژوهشهای آینده

چالشهای جالبی در این حوزه وجود دارد که پژوهشهای بیشتری را طلب میکند که در ادامه این رسالهٔ دکتری در دسور کار ما خواهد بود:

1. استفاده از عبارتهای xPath 100 برای کشف الگوها با برخی محدودیتها در تبیین(2) مواجه است:

مثلاً در بعضی موارد ممکن است عبارتهای معمول برای شناسایی یک برنامه کاربردی وب مورد نیاز باشد میتوانیم به عبارت های 20 xPath برگردانیم و از آنها استفاده کنیم یا کارکردهای بسط به آنها برای این منظور بیفزاییم ولی باید با اهداف بهینه سازی باید بکوشیم زبان برنامه نویسی را حتی الامکان اعلانی نگه داریم.

2.از چالشهای مهم تحقیق در امکان خودکار کردن بدون نظارت آموزش برنامه های کاربردی وب جدید ب) واسط الگوهای همگانی و انطباق با تغییرات اندک در قالبهایی است که پوششها را غیر قابل استفاده میکند.

3.همچنین به طور قطع باید در سراسر این کار تلفیق دقیق خزشگرها را در سامانه با تهیه و توسعهٔ مکانیزمی برای تعامل با سایر اجزا صورت گیرد از جمله چالشهای این حوزه اینکه که به علت لزوم

1- Nondeterministic Finite در تئوری محاسبات اتوماتون تعین ناپذیر متناهی یا اتوماتون غیر قابل تعین متناهی یا اتوماتون پشته ای یا آن اف ای به او توماتونهایی گفته میشود که در مورد آنها برای برخی از دوتاییهای حالت و سمبل ورودی امکان عبور به بیشتر از یک حالت جدید اجازه داده شده باشد.NFA (Automaton

2- Expressiveness

پایبندی به مسئلهٔ «اخلاقیات»(1) هنوز خزشگر همهٔ تعاملات وبی را برعهده دارد، آن هم در جایی بعضی عملکردهای خزش ممکن است مستلزم ورود به یک برنامهٔ خارجی(2)(یک خزشگر که مثلاً API یا یک ارزیاب oxPath) باشد.

بدیهی است که چالش دیگر ذکر مقیاسهایی است که با آنها عملکرده سامانه ما، هم از حیث کارآمدی و هم از حیث اثربخشی، با توجه به رویکردهای کلاسیک خزش مورد ارزیابی قاعده مند قرار می گیرند.

قدردانی

این پژوهش با پشتیبانی مالی هفتمین برنامه اتحادیه اروپایی(3) تحت تفاهم نامه مالی شماره 270239 (ARCOMEM) انجام شده است. نگارندگان همچنین مراتب قدردانی خود را به ژولین ماسانه (از مؤسسة حافظه اینترنت) برای ارائهٔ مباحثی در موضوع این پایان نامه ابلاغ میکنند.

منابع

1. E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer. The connectivity sonar:
Detecting site functionality by structural patterns. In HT, 2003.

2. ARCOMEM Project. http://www.arcomem.eu/, 2011-2014.

3. L. Barbosa and J. Freire. Searching for hidden-Web databases. In WebDB, 2005.

4. L. Barbosa and J. Freire. An adaptive crawler forlocating hidden-web entry points. In
WWW, 2007.

5. B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers.
In COLT, 1992.

6. R. Cai, J.-M. Yang, W. Lai, Y. Wang, and L. Zhang. iRobot: An intelligent crawler for
Web forums. In WWW, 2008.

7. S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: A new approach to
topic-specific Web resource discovery. Computer Networks, 31(11-16), 1999.

8. Y. Diao, M. ALTINEL, M. J. Franklin, H. Zhang, and P. Fischer. Path sharing and

.predicate evaluation for high-performance XML filtering. ACM TODS, 2003

T. Furche, G. Gottlob, G. Grasso, C. Schallhart, and A. J. Sellers. OXPath: A language for .9

.scalable, memory-efficient data extraction from web applications.PVLDB, 4(11), 2011

.D. Gibson, K. Punera, and A. Tomkins. The volume and evolution of web page templates .10

Politeness -1

External programm -2

.(European Union Seventh Framework Programm (FP7/2007-2013 -3

In WWW, 2005. [11] Y. Guo, K. Li, kai Zhang, and G. Zhang. Board forum crawling: A

.Web crawling method for Web forums. In WIC, 2006

./M. Hadley. Web application description language. http://www.w3.org/Submission/wadl .11

/International Business Times. http://www.ibtimes.com/articles/175488 .12

.obama-twitter-townhall.htm, 2011/20110706 .13

P. Kolari, T. Finin, and A. Joshi. Svms for the Blogosphere: Blog Identification and Splog .14

.Detection. In AAAI, 2006

C. Lindemann and L. Littig. Coarse-grained classification of Web sites by their structural .15

.properties. In CIKM, 2006

.C. Lindemann and L. Littig. Classifying Web sites. In WWW, 2007 .16

.M. Liu and T. W. Ling. A rule-based query language for HTML. In DASFAA, 2001 .17

.J. Masanè s. Web archiving. Springer, 2006 .18

E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector .19

.machines. In Workshop on Neural Networks for Signal Processing, 1997

A. Sahuguet and F. Azavant. Building light-weight wrappers for legacy Web data-sources .20

.using W4F. In VLDB, 1999

N. Sawa, A. Morishima, S. Sugimoto, and H. Kitagawa. Wraplet: Wrapping your Web .21

.contents with a lightweight language. In SITIS, 2007

W. Shen, A. Doan, J. F. Naughton, and R. Ramakrishnan. Declarative information .22

.extraction using Datalog with embedded extraction predicates. In VLDB, 2007

.K. Sigurðsson. Incremental crawling with Heritrix. In IWAW, 2005 .23

J.-Y. Su, D.-J. Sun, I.-C. Wu, and L.-P. Chen. On design of browser-oriented data .24

.extraction system and plug-ins. In JMST, 2010

S. Zheng, R. Song, J.-R. Wen, and D. Wu. Joint optimization of wrapper generation and .25

.template detection. In SIGKDD, 2007

چکیده

امروزه دسته بندی صفحات وب برای بازیابی و مدیریت اطلاعات وب ،ساختار،حفاظت یا گسترش فایلهای وب (توالی وب) ، بهبود کیفیت نتایج تحقیق، و کاهش زمان جست و جو استفاده در موتورهای جست وجوی عمودی و تمرکز خاص ناحیه ای در این موتورها و فیلتر کردن محتوای وب با استفاده از مرور کردن وب کاربرد و ضرورت فراوانی دارد ما برای بهبود دسته بندی مفهومی صفحات وب از رویکرد ارزش واژه در روشی مبتنی بر اتوماتای یادگیر توزیع شده استفاده می.کنیم در روش پیشنهادی به هر صفحه یک اتوماتای یادگیر تخصیص داده می شود که وظیفه آن استفاده از وزن کلمات کلیدی صفحات به منظور یادگیری میزان ارتباط آن صفحه با سایر صفحات وب دیگر است که این امر موجب دسته بندی سلسله مراتبی صفحات میشود برای ارزیابی روش پیشنهادی آن را بر روی داده های مختلفی آزمایش کردیم که نتایج خوبی حاصل شد و همچنین الگوریتم پیشنهادی از سرعت خوبی برخوردار بود.

کلیدواژه: اتوماتای یادگیر اتوماتای یادگیر توزیع شده دسته بندی صفحات وب وزن واژه.

دسته بندی مفهومی صفحات وب با استفاده از اتوماتای یادگیر توزیع شده

روح اله گودرزی (1)، مصطفی پیرهادی(2)

1 - مقدمه

با افزایش روزافزون تعداد صفحات وب ، دستیابی به صفحههای مورد نیاز و همچنین تفسیر آنها به عنوان یک چالش فراروی بازیابی اطلاعات و داده کاوی مورد توجه قرار گرفته است. بنابراین دسته بندی کردن صفحات وب میتواند نقش مهمی در افزایش جست وجو تفکیک خلاصه سازی و تفسیر وب داشته باشد. دسته بندی صفحات وب نوع نظارت شده ای از مسئله آموزشی است که به منظور دسته بندی این صفحات به مجموعه ای از دستههای از پیش تعریف شده به کار میرود که بر اساس داده های آموزشی برچسب دار می باشند . دسته بندی وظایف شامل اختصاص یافتن اسناد بر اصول موضوع، عملکرد ، نظر ، نوع و غیره میباشد بر خلاف بسیاری از دسته بندیهای متنی کلی، روش های دسته بندی صفحات وب دارای مزیت محتوایی یکسان ساختاری و ارتباطی با دیگر صفحات در وب است . در سالهای اخیر کارهایی در زمینه دسته بندی صفحات وب گزارش شده است که برخی از آنها عبارتند از: استفاده از نزدیکترین درخت مجاورتK-NN(2003.Kwon et al) استفاده از شبکه های عصبی

1- کارشناسی ارشد مهندسی کامپیوتر مدرس گروه کامپیوتر آموزشکده سما واحد بروجردruhollahgudarzi@yahoo.com

2- کارشناسی ارشد مهندسی کامپیوتر ، کارشناس سازمان فناوری اطلاعات شهرداری بروجرد mpirhadi89@yahoo.com

عکس

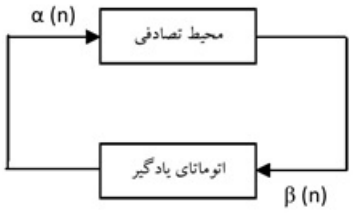
(Selamat et al.۲۰۰٤) (Anagnos et al.۲۰۰٤) ، استفاده از تئوری‌های ناهنجار و نامعلوم و سیستم های فازی برای کاهش داده های زائد (Jensen et al.۲۰۰٤) ، استفاده از مدل SVM دستگاه حفاظتی بردار (Chen et al.۲۰۰٦) (Wakaki et al.۲۰۰٦)، استفاده از وزن کلمات با الگوریتم TFIDF در صفحات وابسته به یک دیکشنری (Liang et al.۲۰۰٦) (Ulmer et al.۲۰۱۰) ، استفاده از درخت های تعیین کننده مبتنی بر فاصله (Estruh et al.۲۰۰٦) و استفاده از الگوریتم ژنتیک (Ozel et al.۲۰۱۰).

در این مقاله روشی مبتنی بر اتوماتای یادگیر توزیع شده که برای دسته بندی صفحات وب از وزن کلمات کلیدی که برمبنای الگوریتم TFIDF محاسبه می شوند، پیشنهاد می گردد . در این روش به هر صفحه وب یک اتوماتای یادگیر اختصاص داده می شود که وظیفه اش یادگیری ارتباط آن صفحه با دیگر صفحات و در نتیجه تعیین دسته مطلوب آن صفحه می باشد که اتوماتا برای این کار باید از وزن کلمات کلیدی صفحه که با استفاده از الگوریتم TFIDF به دست می آید استفاده نماید.

روش پیشنهادی ضمن داشتن کارایی مناسب و صحت مدل ، پارامترهای یادگیری در اتوماتای یادگیر توزیع شده را با توجه به تعداد اسناد وب و با توجه به تعداد کلمات کلیدی به صورت پویا تنظیم می‌کند و اجرای آن بر روی صفحات وب آزمایشی نتایج خوبی را نشان می دهد. ادامه مقاله به دین صورت سازماندهی شده است : در بخش ۲ اتوماتای یادگیر و اتوماتای یادگیر توزیع شده به اختصار بیان می‌شود. در بخش ۳ روش TFIDF برای محاسبه شباهت سند و وزن یابی واژه مورد بحث قرار می گیرد. در بخش ٤ روش پیشنهادی ارائه خواهد شد و در بخش ٥ نتایج حاصل از ارزیابی روش پیشنهادی بیان می‌شود و بخش نهایی هم نتیجه گیری است.

۲- اتوماتای یادگیر(Narendra et al.۱۹۸۹ , Lakshmivarahan et al.۱۹۸۱)

اتوماتای یادگیر یک مدل انتزاعی است که تعداد محدودی عمل را می تواند انجام دهد . هر عمل انتخاب شده توسط محیطی احتمالی ارزیابی شده و پاسخی به اتوماتای یادگیر داده می شود. اتوماتای یادگیر از این پاسخ استفاده نموده و عمل خود را برای مرحله بعد انتخاب می کند . شکل ۱ ارتباط بین اتوماتای یادگیر و محیط را نشان می دهد.



شکل ۱: ارتباط بین اتوماتای یادگیر و محیط

(Selamat et al.200) (Anagnos et al. 2004) ، استفاده از تئوریهای ناهنجار و نامعلوم و سیستم های فازی برای کاهش دادههای زائد) .Jensen et al) ، استفاده از مدل SVM دستگاه حفاظتی بردار (Wakaki et al.,2006) (Chen et al.2006) استفاده از وزن کلمات با الگوریتم TFIDF در صفحات وابسته به یک دیکشنری .Liang et al) (Ulmer et al.2010) ، استفاده از درختهای تعیین کننده مبتنی بر فاصله (Estruh et al.2006) و استفاده از الگوریتم ژنتیک (Ozel et al.2010).

در این مقاله روشی مبتنی بر اتوماتای یادگیر توزیع شده که برای دسته بندی صفحات وب از وزن کلمات کلیدی که برمبنای الگوریتم TFIDF محاسبه میشوند پیشنهاد میگردد . در این روش به هر صفحه وب یک اتوماتای یادگیر اختصاص داده میشود که وظیفه اش یادگیری ارتباط آن صفحه با دیگر صفحات و در نتیجه تعیین دسته مطلوب آن صفحه میباشد که اتوماتا برای این کار باید از وزن کلمات کلیدی صفحه که با استفاده از الگوریتم TFIDF به دست می آید استفاده نماید.

روش پیشنهادی ضمن داشتن کارایی مناسب و صحت مدل پارامترهای یادگیری در اتوماتای یادگیر توزیع شده را با توجه به تعداد اسناد وب و با توجه به تعداد کلمات کلیدی به صورت پویا تنظیم میکند و اجرای آن بر روی صفحات وب آزمایشی نتایج خوبی را نشان میدهد. ادامه مقاله به دین صورت سازماندهی شده است: در بخش 2 اتوماتای یادگیر و اتوماتای یادگیر توزیع شده به اختصار بیان میشود. در بخش 3 روش TFIDF برای محاسبه شباهت سند و وزن یابی واژه مورد بحث قرار میگیرد در بخش 4 روش پیشنهادی ارائه خواهد شد و در بخش 5 نتایج حاصل از ارزیابی روش پیشنهادی بیان میشود و بخش نهایی هم نتیجه گیری است.

2-اتوماتای یادگیر (Narendra et al.1989 ,Lakshmivarahan et al.1981)

اتوماتای یادگیر یک مدل انتزاعی است که تعداد محدودی عمل را میتواند انجام دهد . هر عمل انتخاب شده توسط محیطی احتمالی ارزیابی شده و پاسخی به اتوماتای یادگیر داده میشود. اتوماتای یادگیر از این پاسخ استفاده نموده و عمل خود را برای مرحله بعد انتخاب میکند. شکل 1 ارتباط بین اتوماتای یادگیر و محیط را نشان می دهد.

شکل :1 ارتباط بین اتوماتای یادگیر و محیط

-2-1- اتوماتای یادگیر با ساختار متغیر

اتوماتای یادگیر با ساختار متغیر توسط 4 تایی .app نشان داده میشود که در آن ...,,a} a, a { } - a, مجموعه عملهای اتوماتا. { . ...,,,, } = B مجموعه ورودیهای اتوماتا {p, ...,,,,}=p بردار احتمال انتخاب هر یک از عملها و (1) n (Tan) الگوریتم یادگیری میباشد . در این نوع از اتوماتاها اگر عمل در مرحله nام انتخاب شود و پاسخ مطلوب از محیط دریافت نماید ، احتمال (n) افزایش یافته و سایر احتمالها کاهش می یابند و برای پاسخ نا مطلوب احتمال (n) کاهش یافته و سایر الگوریتم زیر یک نمونه از الگوریتمهای یادگیری خطی در اتوماتای با ساختار متغییر است .

الف - پاسخ مطلوب

ب - پاسخ نامطلوب

در روابط فوق ، a پارامتر پاداش و پارامتر جریمه می باشد.

2-2- اتوماتای یادگیر توزیع شده

یک اتوماتای یادگیر توزیع شده شبکه ای از اتوماتاهای یادگیر است که برای حل یک مسأله خاص با یکدیگر همکاری دارند در این شبکه اتوماتاهای یادگیر همکار در هر زمان تنها یک اتوماتا فعال است تعداد اعمال قابل انجام توسط یک اتوماتا در DIA برابر با تعداد اتوماتاهایی است که به این اتوماتا متصل شدهاند. انتخاب یک عمل توسط اتوماتای یادگیر در این شبکه باعث فعال شدن اتوماتای یادگیر متصل شده به این اتوماتای یادگیر متناظر با این عمل گردد . به عبارت دیگر انتخاب یک عمل توسط یک اتوماتای یادگیر در این شبکه متناظر با فعال شدن یک اتوماتای یادگیر دیگر در این شبکه است. یک DIA توسط یک گراف که هر یک از رئوس آن یک اتوماتای یادگیر است، نشان داده میشود وجود یال (LALA در این گراف بدین معناست که انتخاب عمل توسط LA باعث فعال شدن LA می گردد . تعداد اعمال قابل انتخاب توسط LA بصورت pf ... pt pt { = palpk .... p, p. = {pk نمایش داده شود. در این مجموعه عدد نشان دهنده احتمال مربوط به عمل aman است انتخاب عمل توسط LA باعث فعال شدن LA میشود . r تعداد اعمال قابل انجام توسط اتوماتای LA را نشان میدهد . برای کسب اطلاعات بیشتر راجع به اتوماتاهای یادگیر توزیع شده و کاربردهای آن میتوان به مراجعه(Narendra et al.1989)،(Lakshmivarahan et al.1981) نمود.

3- شباهت سند در روش TFIDF (Ulmer et al.2010)

شباهت سند با استفاده از اهمیت اصطلاح ، یک روش بازیابی اطلاعات است که به صورت زیر بیان

میشود:

عکس

می‌شود:

۱.اصطلاحات مهم و معتبر در هر سند مانند گویا ترین اصطلاحات ،بیشترین اهمیت و اعتبار را دریافت کنند.

۲. هر سند توسط یک بردار اهمیت اصطلاح نشان داده شود.

۳. مقایسه ی اسناد با همدیگر با استفاده از یک مقیاس شباهت در فضای برداری اصطلاح .

یک طرح موثر و شناخته برای هدف، tfidf است.اعتبار tfidf شامل دو قسمت است:tf،idf. تکرار tf ارزیابی می کند که چند وقت یکبار اصطلاح خاص‌tنسبت به همه ی اصطلاحات Vدرصفحه d رخ می دهد:

$$Tf(T_{\cdot\cdot} d)= \frac{count(t_{\cdot} d)}{\sum_{v\epsilon d} count(v_{\cdot} d)} \frac{count(t_{\cdot} d)}{\sum_{v\epsilon d} count(v_{\cdot} d)} \quad (۳)$$

تکرار سند معکوس (idf)،نسبت اسناد d راکه tدر آنها وجود دارد درمجموعه D ارزیابی می کند.

$$idf(t) = \frac{log|D|}{|\{d \epsilon D : t \epsilon d\}|} idf(t) = \frac{log|D|}{|\{d \epsilon D : t \epsilon d\}|} \quad (٤)$$

Tfidf بیشترین اهمیت (وزن) اصطلاح t را که غالباً در سند d وجود دارد تعیین می کند این اعتبار در اسناد دیگر،کمتر وجود دارد.

$$Tfidf(t_{\cdot\cdot} d)=tf(t_{\cdot\cdot} d).idf(t) \quad (٥)$$

در مدل فضای بردار ،سند d با بردار vd متعلق به tfidf مشخص می شود .هر قسمت i از vd، امتیاز iام اصطلاح در مجموعه سند tfidf را می گیردشباهت بین اسناد d و a با استفاده از کسینوس زاویه θ بین بردارهای vbوva در tfidf محاسبه می شود.

$$sim(d,a)= \cos(\theta_{v_d}, v_a (\theta_{v_d}, v_a) \quad (٦)$$

یعنی

$$sim(d_{\cdot\cdot} a)= \frac{va.va}{||vd||\ ||va||} \frac{va.va}{||vd||\ ||va||} \quad (٧)$$

یا به طور هم ارز

$$sim(d,a) \frac{\sum_{t\epsilon d\cap a} tfidf(t,d).tfidf(t,a)}{\sqrt{\sum_{t\epsilon a} tfidf(t,d)^2}\ \sqrt{\sum_{t\epsilon a} tfidf(t,a)^2}} \quad (٨)$$

٤- الگوریتم پیشنهادی

الگوریتم پیشنهادی برای دسته بندی صفحات وب از روشی مبتنی بر اتوماتای یادگیر توزیع شده و از وزن کلمات کلیدی استفاده می کند. اختلاف این الگوریتم با الگوریتم‌های پیشین مبتنی بر اتوماتای یادگیر توزیع شده (Baradaran et al،۲۰۰۷) ، (Anari et al،۲۰۰۷) در این است که در الگوریتم‌های قبلی در حقیقت خوشه بندی صفحات انجام می گیرد و اتوماتا یاد می‌گیرد در بین صفحات وبی که در حال

1. اصطلاحات مهم و معتبر در هر سند مانند گویاترین اصطلاحات بیشترین اهمیت و اعتبار را دریافت کنند.

2.هر سند توسط یک بردار اهمیت اصطلاح نشان داده شود.

3.مقایسه ی اسناد با همدیگر با استفاده از یک مقیاس شباهت در فضای برداری اصطلاح .

یک طرح موثر و شناخته برای ،هدف tfidf .است اعتبار tfidf شامل دو قسمت است tf و idf. تکرار

f ارزیابی میکند که چند وقت یکبار اصطلاح خاص نسبت به همه ی اصطلاحات 7 درصفحه d رخ می دهد:

تکرار سند معکوس (idf) نسبت اسناد d را که در آنها وجود دارد در مجموعه D ارزیابی می

کند.

Tidf بیشترین اهمیت (وزن) اصطلاح t را که غالباً در سند d وجود دارد تعیین می کند این اعتبار در اسناد دیگر کمتر وجود دارد. در مدل فضای بردار سند d با بردار vd متعلق به tfidf مشخص میشود هر قسمت از vd امتیاز نام اصطلاح در مجموعه سند tfidf را می گیرد شباهت بین اسناد d و a با استفاده از کسینوس زاویه 0 بین بردارهای va gvb در tfidf محاسبه می شود.

یعنی

یا به طور هم ارز

4- الگوریتم پیشنهادی

الگوریتم پیشنهادی برای دسته بندی صفحات وب از روشی مبتنی بر اتوماتای یادگیر توزیع شده و از وزن کلمات کلیدی استفاده میکند اختلاف این الگوریتم با الگوریتمهای پیشین مبتنی بر اتوماتای یادگیر توزیع شده ,.Anari et al) ,(Baradaran et al.,2007) در این است که در الگوریتمهای قبلی در حقیقت خوشه بندی صفحات انجام میگیرد و اتوماتا یاد میگیرد در بین صفحات وبی که در حال

حاضـر موجود است کدام ورودی ها را با هم در یک دسـته قرار دهد اما در این الگوریتم عمل دسته بندی بدین معنا انجام میگیرد که برای هر دسـته از روی تشابه مفهومی صفحات آن دسته و با توجه به وزن کلمات کلیدی آن صفحات با استفاده از الگوریتم TFIDF ، یک تابع عضویت دسـته تعریف می شود و هر صفحه جدیدی که وارد شود در این تابع تست می گردد تا عضویت آن در دسته بررسی شود که البته با توجه به تعـداد کلمات انتخاب شـده، میزان شـباهت صـفحات به یکـدیگر نیز به صورت سلسـله مراتبی قابل تعیین است. رونـد اجرای الگوریتم پیشنهادی بصورت شکل 2 می باشد.

1 اتوماتای یادگیر توزیع شده متناظر با ساختار اسناد وب ایجاد کن

2.بردار احتمالاـت اتوماتـای یـادگیر موجود در اتوماتای یادگیر توزیع شـده را مقـدار دهی اولیه کن 3. تکرار ft را برای همه اسـناد مطابق فرمول 3 محاسبه کن

4.به ازای هر کاربر موجود در لاگ فایل انجام بده

4 - 1- به ازای هر حرکت به صورت nm در طول مسیر انجام بده

1-1-4- بردار احتمال اتوماتای متناظر با سند m را مطابق روابط زیر بروز کن

2-1-4 - تکرار سند معکوس) (fdi را طبق فرمول (4) برای مجموعه اسناد محاسبه کند.

ه به ازای صفحات فاقد کاربر در لاگ فایل مطابق فرمول (8) عمل کن وسپس برو به 4-1-1

شکل 2 - الگوریتم پیشنهادی

5-شبیه سازی و ارزیابی الگوریتم پیشنهادی

برای شبیه سازی الگوریتم پیشنهادی از مدل معرفی شده در(Liu et al.2004) استفاده می شود که در آن Liu و همکارانش نظم موجود در رفتارهای کاربران در محیط وب را با استفاده از یک مدل مبتنی بر عامل مشخص و اعتبار مدل خود را با استفاده از اطلاعات استفاده از وب چندین سایت وب بزرگ مانند مایکروسافت تایید کرده اند. آنها به جای استفاده از صفحات وب واقعی و داده های واقعی کاربران وب از این مدل اسـتفاده کردهاند این مدل محیطی شامل صـفحات وب و کاربران آن را فراهم می کند. مزیت استفاده از این مدل آن است که تشـخیص کاربران و بازدیـدهای انجام شده از صـفحات وب با اسـتفاده از این مدل بسـیار دقیق تر میباشد و به عملیات پالایش دادهها نیز احتیاجی نخواهد بود

البته پارامترهای معرفی شـده در این مـدل بایستی بدقت تنظیم گردند تا نتیجه حاصل از آن مشابه با محیط واقعی .گردد هر صـفحه وب در این ،مـدل دارای برداری است که هر مولفه آن میزان ارتباط این صـفحه با موضوع متناظر با این مولفه را نشان می دهـد (تعـداد موضوعات ثابت و قابل تعریف است). میزان ارتباط هر صفحه با یک موضوع به صورت عددی بین صفر و یک بیان می شود به طوری که مجموع آنها برای همه موضوعات برابر با یک است. همچنین هر صفحه دارای پیوندهایی با صفحات دیگر است برای آزمایشها پروفایل علاقه کاربران به صورت توزیع توانی و توزیع محتوای اسـناد به صورت توزیع نرمال در نظر گرفته شده است سایر پارامترهای استفاده شده در این مدل برای شبیه سازی انجام شده در این قسمت در جدول (1) نشان داده شده است.

جدول شماره 1

برای ارزیابی الگوریتمهای پیشنهادی تعیین ساختار اطلاعاتی اسناد وب از معیار کورولیشن(1) استفاده می شود. کورولیشن معیاری برای بدست آوردن وابستگی خطی بین دو بردار است و به صورت زیر تعریف میشود:

عکس

. البته پارامترهای معرفی شده در این مدل بایستی بدقت تنظیم گردند تا نتیجه حاصل از آن مشابه با محیط واقعی گردد. هر صفحه وب در این مدل، دارای برداری است که هر مولفه آن میزان ارتباط این صفحه با موضوع متناظر با این مولفه را نشان می دهد (تعداد موضوعات ثابت و قابل تعریف است). میزان ارتباط هر صفحه با یک موضوع به صورت عددی بین صفر و یک بیان می شود به طوری که مجموع آنها برای همه موضوعات برابر با یک است. همچنین هر صفحه دارای پیوندهایی با صفحات دیگر است. برای آزمایش‌ها پروفایل علاقه کاربران به صورت توزیع- توانی و توزیع محتوای اسناد به‌صورت توزیع نرمال در نظر گرفته شده است. سایر پارامترهای استفاده شده در این مدل برای شبیه سازی انجام شده در این قسمت درجدول(۱) نشان داده شده است.

جدول شماره ۱

۰/۷	حد آستانه ایجاد اتصال
۲۰۰	تعداد کاربران
۵۰۰	تعداد اسناد
۲۵	تعداد موضوع ها
۰/۲	T_c مقدار ثابت سند اولیه (صفحه اولیه سایت) در موضوعات مختلف
۱	a_u پارامتر توزیع قانون - توانی توزیع احتمال علایق کاربران
۰/۹	a ضریب پاداش دریافتی از مشاهده یک سند
۰/۱	b ضریب جریمه دریافتی از پیمایش یک دور
۰/۵	λ ضریب جذب اطلاعات از یک سند توسط یک کاربر
۵/۹۷	μ_m میانگین توزیع نرمال $M\Delta^{v}_{t}$
۰/۲۵	σ_m واریانس توزیع نرمال $M\Delta^{v}_{t}$
۳	a_p پارامتر توزیع قانون - توانی توزیع احتمال وزنهای مطالب برای هر سند
۰/۲۵	σ_t واریانس توزیع نرمال برای مقدار افزایش یک گره برای یک موضوع خاص
۱	θ ضریب کاهش علاقه کاربر
۰/۲	حداقل اشتیاق کاربر برای ادامه جستجو

برای ارزیابی الگوریتم‌های پیشنهادی تعیین ساختار اطلاعاتی اسناد وب از معیار کورولیشن[1] استفاده می‌شود.کورولیشن معیاری برای بدست آوردن وابستگی خطی بین دو بردار است و به صورت زیر تعریف می‌شود:

1. Correlation

Correlation -1

شکل (3) کارایی الگوریتم پیشنهادی را با معیار کورولیشن در مقایسه با الگوریتمهای ارائه شده (Anari et al.2007 نشان میدهد محور افقی تعداد کاربران و محور در(Baradaran_et al.2007) عمودی میزان کورولیشن را نشان میدهد. در ارزیابی فوق تعداد صفحات 20, تعداد موضوعات 5 و تعداد کاربران 20000 در نظر گرفته شده است.

شکل (3) – مقایسه روش پیشنهادی

عکس

$$Corr(p,p') = \frac{\sum pp' - (\sum p \sum p')/n}{\sqrt{(\sum p^2 - (\sum p)^2/n)(\sum p'^2 - (\sum p')^2/n)}}$$

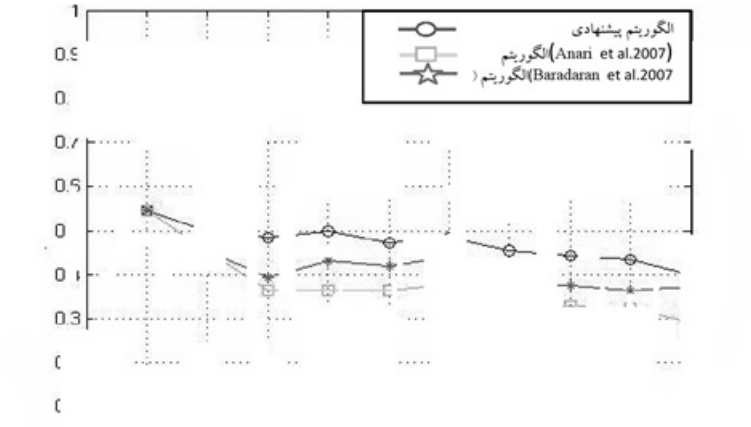
$p=\{p_{ij} \mid i,j=1,2,3,\ldots,n,\ i\neq j\}$

$$p_{ij}=\frac{d_{ij}^{-2}}{\sum_{k=1}^{n} d_{ik}^{-2}} \quad (9)$$

$$d_{ij} = \sqrt{\sum_{k=1}^{m}\left(cw_i^k - cw_j^k\right)}\sqrt{\sum_{k=1}^{m}\left(cw_i^k - cw_j^k\right)}$$

$p' = \{p'_{ij} \mid i,j=1,2,3,\ldots,n,\ i\neq j\}$

شکل(۳) کارایی الگوریتم پیشنهادی را با معیار کورولیشن در مقایسه با الگوریتم‌های ارائه شده در(Baradaran et al.۲۰۰۷) و (Anari et al.۲۰۰۷) نشان می دهد. محور افقی تعداد کاربران و محور عمودی میزان کورولیشن را نشان می‌دهد.در ارزیابی فوق تعداد صفحات ۲۰ , تعداد موضوعات ۵ و تعداد کاربران ۲۰۰۰۰ در نظر گرفته شده است.



شکل (۳)- مقایسه روش پیشنهادی

6- نتیجه گیری

با توجه به اهمیت دسته بندی صفحات وب در این مقاله روشی مبتنی بر اتوماتای یادگیر توزیع شده و ارزش کلمات پیشنهاد گردید و کارایی آن با توجه به معیار کورولیشن و در نتیجه وابستگی خطی در مقایسه با الگوریتمهای ارائه شده در(Baradaran et al.2007) و (Anari et al.2007) مورد ارزیابی قرار گرفت که الگوریتم پیشنهادی از کورولیشن بالاتری نسبت به این دو روش برخوردار است. ایده اصلی این الگوریتم این بود که صفحاتی که با همدیگر ارتباط مفهومی بیشتری دارند ، پاداش بیشتری را دریافت مینمایند . از نتایج بدست آمده از این الگوریتم میتوان به عنوان ابزاری برای یافتن صفحات مشابه و مرتبط با یک موضوع خاص استفاده نمود.

منابع

Kwon,O.w., J.H. Lee. (2003). Text categorization based on k-nearest neighbor approach for .Web site classification. Information Processing and Management, 39:25-44

Anagnostopoulos, I., C. Anagnostopoulos, V. Loumos, and E. Kayafas. (2004). Classifying .(Web pages employing a probabilistic neural network. IEEE Proc.Softw, 151(3

Selamat, A., S. Omatu.(2004).Web page feature selection and classification using neural .networks. Information Sciences, 158:69-88

Jensen, R., Q. Shen.(2004). Fuzzy-rough attribute reduction with application to web .categorization. Fuzzy Sets and System, 141:469-485

Chen, R.C., C.H. Hsieh. (2006). Web page classification based on a support vector machine .using a weighted vote schema.Expert System with Applications, 31:427-435

-Wakaki, T., H. Itkura, M. Tamura, H. Motoda, and T.Washio. (2006).A study on rough set aided feature selection for automatic web-page classification. Web Intelligence and Agent .System, 4: 431-441

Liang, C. Y., L. Guo, Z.J. Xia, F.G. Nie, X.X. Li, L. Su, and Z. Y. Yang.(2006). Dictionary-based ,text categorization of chemical web pages. Information Processing and Management

.42:1017-1029

Ulmer, C., M. Gokhale, B.Gallagher, Ph. Top, and T. Eliassi-rad. (2010). Massively parallel .acceleration of a document -similsrity classifier to detect web attacks. J.Parallel Distrib Comput

Estruh, V., C. Ferri, J. Hernanndes-Orallo, M,j. Ramirez-Quintana .(2006).Web categorization ,using distance-based decision trees. Electronic Notes in Theoretical Computer Science .157:35-40

Ozel,S.A. (2010). A Web page classification system based on a genetic algorithm using .tagged–terms as features. Expert Systems With Applications

:Lakshmivarahan,S.(1981). Learning algorithms: Theory and applications. New York .Springer–Verlag

Narendra,k.S., K.S. Thathachar (1989). Learning automata: An inter oduction. New York .Prentice Hall :

Anari, B., M.R. Meybodi. (2007). A new method based on distributed learning automata for determining web documents structure.Proceedings of the 12 Annual international CSI Computer Conference, CSICC2007, Tehran, Iran, pp.2281

Baradaran Hashemi, A., M.R. Meybodi. (2007).Web usage mining using distributed learning autoumata. Preceedings of the 12 Annual International CSI Computer .Conferenc, CSICC2007, Tehran, Iran, pp. 553–560

Liu, J., S. Zhang, J. Yang. (2004), Characterizing web usage regulaities with information .(foraging agenst. IEEE Transaction in Kniwlodge and Data Engineering, 16(5

Chao, M., H. Chen (2008). A machine learning approach to web page filtering using content .and structure analysis.Decision Support System, 44:482–494

چکیده

ایجاد ابزارهای خودکار در سازمانهایی که از فرآینـدهایی برخوردارنـد که به خوبی تعریف شـده اند، بهبود فعالیتها افزایش کارآیی، کاهش خطاها، و بهبود سطح خدمات آن سازمان خواهد شد.

در سال 2010 تیم آرشیو وب(1) کتابخانه کنگره با توسعه و ایجاد ابزاری با نام DigiBoard به خود کارسازی جریانهای کاری در آرشیو وب پرداخت. .DigiBoard نوعی برنامه کاربردی تحت وب است که به منظور پشتیبانی از جریان کاری آرشیو وب (نامزدی(2)، پردازش مجوزها(3)، و داوری(4)) در کتابخـانه کنگره به کـار میرود و به افزایش بهره وری تیمها و کاربران کتابخانه کنگره کمک می کنـد و کاهش میزان خطاهای معمول در روشهای دستی آرشیو وب را میسر می.سازد مقاله حاضر برخی کارکردهای ابزار جدید و سفارشی شده(5) تیم آرشیو وب کتابخانه کنگره را توصیف خواهد کرد.

1- Web Archiving Team

2- Nomination

3- Permission Processing

4- Reviewing

5- Custom-built

DigiBoard: ابزار افزایش کارآیی فعالیتهای پیچیده آرشیو وب درکتابخانه کنگره

*DigiBoard: ابزار افزایش کارآیی فعالیتهای پیچیده آرشیو وب درکتابخانه کنگره(1)

آبه گروتک(2) | جینا جونز (3)| ترجمه: سعیده اسلامی(4)

1 - مقدمه

کتابخانه کنگره از 10 سال پیش تا امروز به بایگانی کردن محتوای وب پرداخته است و در این سالها بیشتر از 14/000 وبگاه را در 40 مجموعه آرشیو وب موضوعی(5) و گزینشی(6) گنجانده و روالهای کاری ،نامزدی پردازش مجوز تعیین میدان خزش داوری ،کیفیت فهرستنویسی و دسترسی را یکپارچه کرده .است. البته کتابخانه کنگره به منظور خزش وب حکم و قانون و اسپاری معینی ندارد؛ بنابراین برای بایگانی برخی وبگاهها مانند وب نوشتها و سایتهای سازمانهای ،خبری ملزم به دریافت مجوز است، اما وبگاههای دیگر قابل آرشیو شدن هستند و فقط از طرف کتابخانه کنگره مطلع میشوند که محتوای آنها آرشیو خواهد شد. جهت دسترسی به آرشیو کتابخانه کنگره، کتابخانه ملزم به اخذ مجوز از تمامی وبگاههاست وبگاههای دولتی از این قاعده جدا هستند. اگر چه در سالهای اخیر کتابخانه قابلیت

1- DigiBoard: A Tool to Streamline Complex Web Archiving Activities at the Library of Congress

2- Abbie Grotke

3- Gina Jones

4- کارشناس ارشد مهندسی نرم افزار سازمان اسناد و کتابخانه ملی ایران(s-eslami@nlai.ir)

5- Thematic

6- Selective

خزشگری درجا(1) را پیاده سازی کرده است بیشتر محتوا از طریق یک عامل خزشگر غیر درجا(2) حاصل می شود.

2-ضرورت ایجاد ابزار جدید

در سال 2003 کتابخانه کنگره ابزاری ابتدایی را برای تأمین نیازهای مجوزدهی مستندات ساخت که فقط از جریانهای کاری نامزدی و پردازش مجوزها پشتیبانی میکرد تقاضای ساخت چنین ابزاری از طرف OGC(3) کتابخانه کنگره و اداره حق نشر ایالات متحده(4) ارائه شد این ابزار در بازه زمانی یک ماهه ساخته شد اما با افزایش چندین برابری تعداد کاربران به قدر کافی توسعه پذیر نبود، علاوه بر آن، به موجب تغییر حقوق قانونی اخذ مجوزها مرحله پردازش مجوزها بیش از حد تغییر می.کرد. در حال حاضر 5 کارمند تمام ،وقت 3 پیمانکار و بیش از 80 کتابدار در واشنگتن دی سی و نقاط مختلف جهان با هدف ساده سازی مراحل نامزدی و پردازش مجوزها از این ابزار استفاده می کنند.

کتابخانه کنگره لازم میداند از میزان مشارکت کانون وکلا در فعالیتهای آرشیو وب کتابخانه بکاهد؛ بدین ترتیب کاربران مختلف در خدمات کتابخانه(5)، کتابخانه قانون کتابخانه کنگره و خدمات تحقیقاتی کنگره ای(6) به حداقل آموزشها جهت انتخاب کردن وبگاهها نیاز خواهند داشت. علاوه براین تیم آرشیو وب برای مدیریت بهتر جریانهای کاری اش به ابزاری نیاز دارد جریانهای کاری آرشیو وب کتابخانه و مراحل آن به تدریج توسعه پیدا کرد و در مقایسه با سال 2003 که ابزار اولیه ساخته شد، رسمیت بیشتری پیدا کرده است. بنابراین کتابخانه کنگره نیازمند ابزاری با ویژگیهای زیر است:

1) داده های فعالیتهای مختلف را مدیریت کند نامزدی، مجوزها، خزش، بررسی کیفیت، گزارشگیری، و مانند آن؛ و

2) به منظور کاهش زمان پردازش URL برای نامزد کننده و بگاهها و تیم فعالیتهای دستی را خودکارسازی کند.

3- ایجاد و توسعه DigiBoard

DigiBoard ، با استفاده از زبان برنامه نویسی PHP و بانک اطلاعاتی MySQL در چند مرحله توسعه یافته .است نخستین ماژول آن برای نامزد کننده وبگاهها در سپتامبر 2009 راه اندازی شد یک پیمانکار مسئولیت توسعه ابزار اولیه را بر اساس نیازمندیهای مشخص شده توسط 5 عضو تیم آرشیو وب عهده دار شد. کتابداران نیازمندیهای دیگری را تدارک دیدند تا بهسازی مراحل کاری خاصشان را فراهم کند.

مدل

1- On-site

2- Off-site

3- Office of General Counsel

4- United state copyright office

5- Library services

6- Congressional Research Service

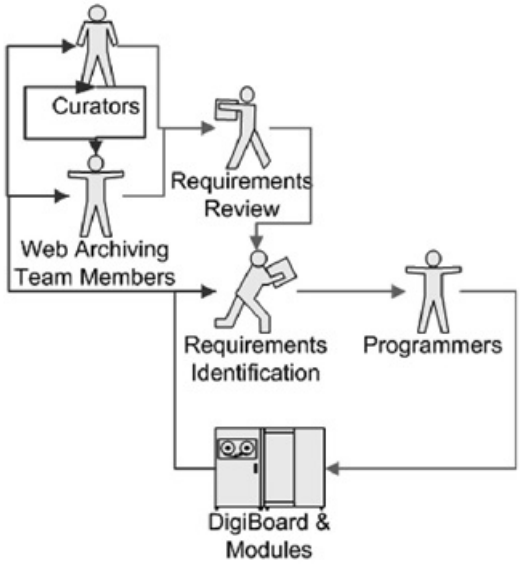
توسعه DigiBoard به صورت چابک(1) و فرآیند توسعه تکراری(2) (شکل 1) میباشد که نیازمندیها و مسائل شناسایی شده توسط کتابداران و اعضای تیم را پوشش میدهد نیازمندیهای هر مرحله از فرآیند تیم آرشیو وب به صورت ترتیبی و افزایشی به سیستم اضافه شده است؛ بنابراین کارآیی مورد اضافه شده توسعه سیستم را تحت تأثیر قرار داده است.

شکل 1 .مراحل توسعه دیجی بورد فرآیند چابک و تکراری توسعه

قابلیتهای ناشی از توسعه موفق آمیز DigiBoard در کتابخانه کنگره عبارت اند از لایق بودن تیم توسعه، دارا بودن دانش کافی درباره واسطهای کاربری اشراف بر تواناییهای پایگاه دادههای وب پویا، توانایی انتقال اطلاعات به توسعه دهندگان سیستم، تجربه بالای افراد تیم و علاقه مند به فعالیتهای خودکارسازی فرآیندها.

عکس

توسعه DigiBoard به‌صورت چابک[1] و فرآیند توسعه تکراری[2] (شکل ۱) می‌باشد که نیازمندی‌ها و مسائل شناسایی شده توسط کتابداران و اعضای تیم را پوشش می‌دهد. نیازمندی‌های هر مرحله از فرآیند تیم آرشیو وب به‌صورت ترتیبی و افزایشی به سیستم اضافه شده است؛ بنابراین، کارآیی مورد اضافه شده توسعه سیستم را تحت تأثیر قرار داده است.



شکل ۱. مراحل توسعه دیجی‌بورد: فرآیند چابک و تکراری توسعه

قابلیت‌های ناشی از توسعه موفق‌آمیز DigiBoard درکتابخانه کنگره عبارت‌اند: از لایق بودن تیم توسعه، دارا بودن دانش کافی درباره واسطهای کاربری، اشراف بر توانایی‌های پایگاه داده‌های وب پویا، توانایی انتقال اطلاعات به توسعه‌دهندگان سیستم، تجربه بالای افراد تیم و علاقه‌مند به فعالیت‌های خودکارسازی فرآیندها.

٤- بهبود فرآیند نامزدی و پردازش مجوزها

٤-١- فرآیند نامزدی

آرشیوهای وب کتابخانه کنگره بر مبنای مفاهیم مجموعه‌ها ساخته می‌شود. URL ها برای یک مجموعه و یا چندین مجموعه نامزد می شوند و حد کمینه فراداده برای نامزد کردن یک URL لازم است. قبل از

1. Agile
2. Iterative process

4- بهبود فرآیند نامزدی و پردازش مجوزها

4-1- فرآیند نامزدی

آرشیوهای وب کتابخانه کنگره بر مبنای مفاهیم مجموعه ها ساخته میشود URL ها برای یک مجموعه و یا چندین مجموعه نامزد میشوند و حد کمینه فراداده برای نامزد کردن یک URL لازم است قبل از

Agile -1

Iterative process -2

ارائه این ابزار لازم بود تا نامزد کننده ها دو فرم جداگانه خارج از ابزار را کامل کنند و سپس برای ادامه نامزدی وارد سیستم شوند اما اکنون کلیه مراحل یکپارچه و ساده تر شده است فرمی که توسط کتابداران کتابخانه کنگره استفاده میشود (به این فرم با نام نامزد کننده در این مقاله اشاره شده است) ساده تر بوده و اطلاعات بیشتری مانند موضوع زبان و دیگر اطلاعات درباره URLها را جمع آوری میکند. این فرم اجازه میدهد تا کارهای نیمه تمام کاربران ذخیره شده و در زمان دیگری ادامه داده شوند و امکان کپی رکوردها از یک مجموعه به مجموعه دیگر را فراهم می آورد کلیه دادهها درباره URL نامزد شده در بانک اطلاعاتی ذخیره شده و در هر زمان به راحتی قابل مشاهد و قابل جست وجوست. همچنین، نامزدکننده ها میتوانند پلاگینهایی را به منظور نامزد کردن وبگاهها برای شمول در یک مجموعه بارگذاری و استفاده کنند (شکل 2 را ببینید).

شکل .2 پلاگینهای موجود برای مرحله نامزدی در دیجی بورد

2-4- پردازش مجوزها

از آنجا که به دست آوردن مجوز از مالکان وبگاهها برای آرشیو در کتابخانه کنگره ضروری است فرآیند مجوز درون فرآیند نامزدی ساخته شده است از سال 2003، الزاماً باید تمام وبگاههای آمریکا به غیر از وبگاههای دولتی از قصد کتابخانه کنگره جهت آرشیو محتوای وبگاه شان مطلع شوند. علاوه براین، OGC کتابخانه کنگره صدور مجوز از طرف مالکان محتوا جهت گردآوری اطلاعات را برای برخی وبگاهها ضروری دانسته است ابزار قدیمی از طریق پست الکترونیکی مجوزها را بر اساس نوع مجوز مجوز گردآوری (مجوز نمایش محتوای غیر درجا) به مالکان محتوا ارسال می.کرد در سال 2006، OGC امکان مجوزهای روکشی(1) را فراهم آورد بدین معنی که اگر مالک محتوا مجوز گردآوری یا نمایش را اعطا کند، این مجوز برای مجموعه های آینده هم قابل استفاده مجدد خواهد بود در سالهای اخیر کتابخانه کنگره به ازای هر URL اضافه شده به مجموعه حتی اگر قبلاً مجوز کسب کرده باشد، اطلاعیه ای به مالکان محتوا ارسال میکند.

مدیریت مجوزهای روکشی در ابزار قدیمی به آسانی ممکن نبود و نامزد کننده ها ملزم بودند تا یک

1- Blanket permission

فایل خارجی حاوی فهرست وبگاههایی دارای مجوز کنکاش کننـد در برخی موارد، مجوزها مجـدداً درخواست میشدنـد در نتیجه حدود 15 تا 30 دقیقـه به ازای هر وبگـاه زمـان از دست می رفت. اگرچه مـاژول مجـوز DigiBoard امکـان ردیابی بهتر و قابلیت اسـتفاده از مجوزهای روکشـی را فراهم آورده است به دلیل پیچیدگی وب و نیاز به آموزشهای پیشرفته برای کاربران جهت اعمال مجوزهای روکشی به طور منـاسب به نامزدهای جدیـد این ماژول چالش برانگیزترین ماژول در مرحله پیاده سازی بود. به عنوان مثال URLهای مختلف از یک دامنه وبگـاه برای مجمـوعه هـای مختلفی نـامزد میشود و لاـزم است نـامزد کننـده هـا آموزش ببیننـد که گـاهی اوقـات زیر دامنه دامنه ها ،دایرکتوریهـا و فایلهـا به یک مالک محتوا تعلق دارد و گاهی بنا به ماهیتشان به همان مالک محتوا تعلق .نـدارد برای مثال، یک وب نوشـت درباره قانون با میزبانی دانشـگاه کلمبیا برای شـمول در آرشـیو وب دیوان عالی کشور(1) نـامزد شـده بود و مالک ،سایت مجوز آن را برای آرشیو به کتابخانه کنگره اعطا کرده بود وبگاه معماری دانشگاه کلمبیا برای یک آرشیو وب دیگر نامزد شده بود و اطلاعاتی در DigiBoard مبنی بر اینکه مالک وب نوشت مجوز اعطا کرده است به کتابـدار ارائه شـد و آنها مجوز روکشـی را به وبگاه کتابخانه اعمال کردند. به دلیل اینکه مالکان سایت به صورت دو موجودیت مختلف در نظر گرفته میشونـد آموزشـهایی جهت وضوح اینکه بخش های مختلف دانشـگاه کلمبیا ممکن است چندین مالک داشته باشند لازم است.

3-4- بهبود پاسخ مالک محتوا

به علت نحوه طراحی DigiBoard از زمان پیاده سازی آن افزایش قابل توجهی در میزان پاسـخهای مالکان محتوا به وجود آمده است. در حـالی که ابزار قبلی امکـان ارسال پست الکترونیکی و پاسـخ دادن از طریق یک فرم وب را برای مالکان محتوا فراهم آورده بود و گاه پسـت الکترونیکی به اشتباه به صورت اسپم(2) در نظر گرفته میشد علاوه بر این فرم و بی پاسخها با سایت آرشیو وب یکپارچه نبود. با ابزار جدید ایمیلهای تولید شده اجازه میدهند تا مالک محتوا نیز مسـتقیماً با کلیک بر یک پیوند ساده و یا با رجوع به صـفحه وبگاه و با وارد کردن یک شماره یکتا - که امکان پیگیری پاسخها را داشته باشد - پست الکترونیکی خود را ارسال کند (شکل3).

شکل 3 .بازیابی پست الکترونیکی مجوز

1- supreme court

2- spam

میزان پاسخ مجوزها با پیاده سازی دیجی بورد به میزان 50 درصد افزایش پیدا کرده است.

4-4- خصوصیات جدید

خصوصیات جدیدی به ماژولهای نامزدی و ماژول مجوز DigiBoard اضافه شده است که جریانهای کاری و پردازش URLها را بهبود میبخشد. در حال حاضر:

• دیجی بورد روشی برای کپی کردن یک URL در یک مجموعه جدید نگهداری تاریخچه تغییرات و ثبت عملیات انجام شده امکان ثبت تغییرات URL، ها امکان فهرستنویسی یکپارچه و مدیریت مجوزها را فراهم کرده است.

• DigiBoard مکانیزمی برای شناسایی و پیاده سازی آسان خصوصیات مشترک وبگاهها فراهم کرده است که پیچیدگی انتخاب مجوزها از دامنه زیر دامنه دامنه های سطح بالا- و میزبانهای وب را کاهش داده است. به عنوان مثال بیش از 250 وبگاه بلاگ اسپات در ابزار DigiBoard نامزد شده اند که هر کدام متعلق به یک مالک محتوا مختلف است. این ابزار امکاناتی فراهم آورده که به موجب آن، نیاز کمتری به تصمیم گیری نامزد کننده ها در مرتبط و نامرتبط بودن وبگاه وجود دارد.

• در DigiBoard نامزد کننده ها به رکوردهای خودشان شامل URLها در مجموعه های فعال، مجموعه های غیر فعال و نشانیهای وبگاههای آرشیو شده دسترسی دارند.

• DigiBoard امکان دسترسی تک - کلیکی را به منظور ارائه اطلاعات مدیریتی درباره مجموعه ها شامل مستندات برنامه ریزی شده مجموعه و اطلاعاتی درباره کنترل مجوزهای هر پروژه برای نامزد کننده ها فراهم می آورد.

• DigiBoard دارای واسط کاربری با قابلیت دسته بندی براساس مجموعه طبقه نامزد کننده بخش و مجوز میباشد.

• DigiBoard امکان ارسال چندین پست الکترونیکی مجوز به یک مالک محتوا را فراهم کرده است.

• DigiBoard امکان انقضای سالیانه و یا شش ماهه URL ها را فراهم میکند و طی آن لازم است نامزد کننده ها مجدداً URL ها را جهت گردآوری دوباره تأیید کنند.

• DigiBoard امکانی برای داوران فراهم میکند تا نامزدها را به منظور اعمال کار بیشتر و بررسی مجدد مجوزها به نامزد کننده ها برگشت دهند و وبگاههایی را که با راهنماهای انتخاب(1) مجموعه همخوانی ندارند رد کنند.

• DigiBoard امکان انتقال مالکیت یو آر ال از یک نامزد کننده به نامزده کننده دیگر را فراهم می کند که طی آن مسئولیت تصدیق URL و گردآوری نامزدی دوباره نسبت داده شده(2)، با نامزد کننده جدید خواهد بود.

1- Selection guidelines

2- Reassigned nomination

5.مدیریت هسته(1)

بررسی کیفیت پیش - خزش(2) فعالیتهای مدیریت ،هسته و توسعه دستور عملهای تعیین حوزه خزش برای خزشگرهای هر URL نامزد شده را در برمی گیرد. قبل از توسعه این ماژول در ابزار قبلی مرحله مدیریت هسته به صورت خارجی صورت میگرفت بدین صورت که به ازای هر مجموعه یک فایل متنی از هسته ها به همراه دستور عملهای تعیین حوزه خزش وجود داشت ،بنابراین ویرایش و به روز نگهداشتن فایل کار مشکلی بود و قالب بندی خاص URLها دستی انجام میگرفت که متمایل به خطا .بود با پیاده سازی ماژول مدیریت هسته DigiBoard چندین فعالیت خودکار شده را فراهم کرده است که به موجب آن به نسبت آن به نسبت 10:1 در مصرف زمان صرفه جویی میشود.

5-1 - وارسی خودکار در نقطه شروع هسته

بیشتر مواقع نامزد کننده ها URL وبگاهی را معرفی میکنند که شامل فایلی با نام index است (برای مثال http://www.loc.gov/index.html). در این حالت سیستم در مرحله مرور پیشخزش بررسی میکند که آیا تعداد بایتهای http://www.loc.gov و http://www.loc.gov/index.html برابر هستند یا نه اگر برابر باشند یا نه سیستم http://www.loc.gov را به عنوان نقطه شروع بهینه برای خزش معرفی میکند.

5-2- وارسی تک - کلیک برای پیوندهای سایت و نقشه سایتها

وارسی تک - کلیک برای پیوندها روش آسانی برای مرورگران تیم آرشیو وب فراهم میکند که به موجب آن میتوانند انتخاب و نمایش پیوندها برای تعیین میدانی برای شمول و همچنین برای انتخاب میدان مناسب برای محتوای سایت را به راحتی انجام دهند. برای مثال آرشیو انتخابات را در نظر .بگیرید وبگاههای فیس بوک توییتر ،فلیکر مای ،اسپیس و یوتیوب متعلق به کاندیدای انتخاب به عنوان وبگاه هایی در نظر گرفته میشوند که با دستور عملهای خزش مرتبط هستند. ماژول مدیریت هسته، شناسایی این URLها و تغییر فرمت مناسب برای خزشگر را آسان می کند.

5-3- قالب بندی فهرست هسته

قالب بندی URL ها در دامنه زیر دامنه مسیر و فایلهایی با فرمت SURT(3) که در حال حاضر در خزشگر هریتریکس(4) پشتیبانی میشود در ابزار DigiBoard به طور خودکار انجام می شود.

1- Seed management

2- Pre-crawl

3- Sort-friendly URI Reordering Transform

4- Heritrix

5-4- مدیریت تناوب(1) هسته

DigiBoard با پشتیبانی از امکان مـدیریت ،هفتگی ماهیانه سه ماهه شـش ماهه و سالیانه خزشـها، امکان ایجاد فهرست هسـته بر اساس تناوب و براساس مجموعه را فراهم میکند این ماژول مراحل جاری قرار داد بسـتن با عاملهای خزشـگر را آسان میکند فهرست هسته صادر شده و به عاملها تحویل داده می شود.

عکس

٥-٤- مدیریت تناوب[1] هسته

DigiBoard با پشتیبانی از امکان مدیریت هفتگی، ماهیانه، سه ماهه، شش‌ماهه، و سالیانه خزش‌ها، امکان ایجاد فهرسـت هسته براساس تناوب و براسـاس مجموعه را فراهم می‌کند. این ماژول، مراحل جاری قرارداد بسـتن با عامل‌های خزشگر را آسان می‌کند، فهرست هسته صادر شده و به عامل‌ها تحویل داده می‌شود.

٦- بررسی پس خزش[2]

DigiBoard از دو نوع بررسـی کیفیت پس خزش پشـتیبانی می‌کند: براساس نامزدکننده، براساس تیم آرشیو وب.

٦-١- بررسی کیفیت نامزدکننده

امکان انجام بررسی کیفیت برای نامزدکننده‌ها، باعث می‌شود تا تیم آرشیو وب از تعداد نامزدکننده‌هایی که محتوای گردآوری شده را بازرسی می‌کنند مطلع شوند. واسط کاربری DigiBoard پیوندی به آرشیو هر کدام از وبگاه‌های نامزد شده برای نامزدکننده‌ها فراهم می‌کند و بررسی نامزده‌کننده‌ها را با ثبت شناسه کاربری، و URL بررسی شده ضبط و ثبت می‌کنند. با ورود به آرشیو، پیوندی در بنر آرشیو وجود دارد که با کلیک روی آن می‌توانند با «ضبط موفقیت‌آمیز اسـت» و «موفقیت‌آمیز نیست» همراه با توضیحی پاسخ دهند (شکل ٤).

شکل ٤. فرم بررسی کیفیت نامزدکننده

1. Frequency
2. Post-Crawl

6- بررسی پس خزش(2)

DigiBoard از دو نوع بررسی کیفیت پس خزش پشتیبانی میکند بر اساس نامزدکننده بر اساس تیم آرشیو وب.

6-1 بررسی کیفیت نامزد کننده

امکان انجام بررسی کیفیت برای نامزد کننده ها باعث میشود تا تیم آرشیو وب از تعداد نامزدکنندههایی که محتوای گردآوری شده را بازرسی میکنند مطلع شوند واسط کاربری DigiBoard پیوندی به آرشیو هر کدام از وبگاههای نامزد شده برای نامزد کننده ها فراهم میکند و بررسی نامزده کننده ها را با ثبت شناسه کاربری و URL بررسی شده ضبط و ثبت میکنند با ورود به آرشیو پیوندی در بنر آرشیو وجود دارد که با کلیک روی آن میتوانند با ضبط موفقیت آمیز است و موفقیت آمیز نیست همراه با توضیحی پاسخ دهند (شکل 4).

شکل4.فرم بررسی کیفیت نامزد کننده

1- frequency

2- post-crawl

2-6.بررسی کیفیت پس خزش پس خزش تیم آرشیو وب

تأیید عمق و بگاهی که کتابخانه LOC ضبط میکند در طول مرحله بررسی کیفیت پس خزش رخ میدهد به دلیل مرحله پردازش مجوزها، خزشهای LC بـا دسـتور عملهـای صـریح به خزشـگر - با توجه به اینکه اجازه داره کجا در وب برود - انجام میشود. بررسـی کیفیت پس خزش به سؤالهای زیر پاسخ میدهد:

آیا قلمرو یا حوزه سایت به درستی مشخص شده است؟

آیا تمام وبگاههای مرتبط دیگر مشخص شده و حوزه آنها تعیین شده است؟

آیا ضبط مشکلی دارد؟

آیا این مشکلات شناخته شده و قابل شناسایی هستند یا به تحقیقات بعدی نیاز است؟

مرورگر حداقل یکبار در زمان گردآوری ،سایت به ازای هر URL انتخابی با استفاده از ماژول DigiQR(1) دیجی بورد و پلاگین QR فایر فاکس بررسـی کیفیت را انجام میدهـد. همان طور که در شـکل ه نشان داده شده است مرورگر میتواند مشـخص کند که سایت خوب درو (جمع آوری) شده است، یا به منظور لزوم بررسی بیشتر گزینه partial را انتخاب کرده و مسئله ای(2) را عنوان کند.

شکل5. بررسی کیفیت وبگاه

اگر مسـئله آرشیوی باشد مرورگر نوع مسئله و شدت آن را شناسایی می کند. مسئله ها به انواع خاصی از مسئله متعارف طی مرحله بررسی کیفیت طبقه بنـدی شـده .است پلاگین نوعی واسط کاربری را ایجاد کنـد که مسـئله را - در همان صـفحه ای که پیدا شده است - گزارش میدهد(شکل6).

1- Quality Reviwe

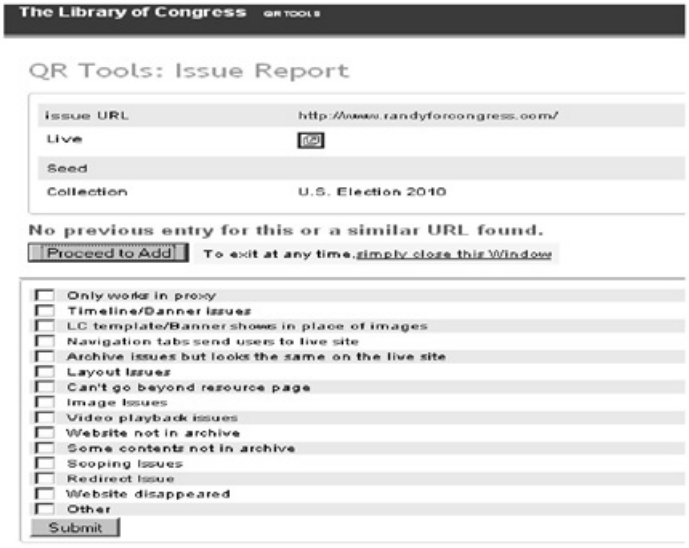
2- issue

شکل 6 . صفحه گزارش کردن مشکلهای در ماژول DigiQR

در مثال QR (شکل 7) وب سـرور مانع ضبط کردن وبگاه توسط خزشگر می شود. متخصص بررسی کیفیت، مسئله را شناسایی میکند و سطح شدت آن را تعیین می کند.

شکل .7 شناسایی مشکلات در ماژول DigiQR

عکس



شکل ٦. صفحه گزارش کردن مشکل‌های در ماژول **DigiQR**

در مثال QR (شکل ۷) وب سرور مانع ضبط‌کردن وبگاه توسط خزشگر می‌شود. متخصص بررسی کیفیت، مسئله را شناسایی می‌کند و سطح شدت آن را تعیین می‌کند.

شکل ۷. شناسایی مشکلات در ماژول **DigiQR**

این مسئله توسط یک عضو تیم آرشیو وب بیشتر بررسی میشود مراحل تحقیقات و نتایج آن ثبت می.گردد واسط جست و جوی DigiQR امکان جست وجو بر اساس ،مسئله براساس مجموعه، براساس URL و براساس خزش را فراهم می سازد.

پیش از توسعه ماژول DigiQR کلیه کارها به صورت مجزا صورت میگرفت و اعضای تیم آرشیو وب از صفحه گستردههای خودشان با اصطلاحات فنی متناقض جهت مستندسازی مسئله ها استفاده می کردند. اکنون DigiQR تیم را قادر میسازد تا مسئله های محتوای خزش شده را در یک چارچوب مدیریت دانش اشتراکی با کارآیی بیشتر ثبت و پیگیری نمایند. اعضای تیم یاد گرفته اند که مسئله های QR پیچیده تری را شناسایی کنند زیرا ابزار امکان دسترسی آسان به اطلاعات مسئله ها و نتایج تحقیقات را فراهم میآورد این ماژول زمان صرف شده در کپی کردن URL ها در صفحه گسترده ها را کاهش داده امکان کارآمدتری برای پیگیری مسئله فراهم میکنند این ماژول مکان رخ دادن مشکل را دقیقاً ثبت جست و جوی های بعدی را تسهیل می بخشد.

7-فعالیتهای اجرایی

ماژول های اجرایی از تمام فعالیتهای مدیریتی دیجی بورد پشتیبانی میکند، اما به طور قابل توجهی امکاناتی برای مدیریت نقشهای کاربران و مدیریت مجموعه های آرشیو وب فراهم شده است.

7-1- مدیریت کاربر

با توجه به تعداد تخمینی کاربران DigiBoard توانایی مدیریت فعالیتهای کاربران توسط تیم آرشیو وب ، امری حیاتی است به دلیل اینکه شمول یک URL در فهرست هسته به قیمت زمان کارمندان در مرحله QR، دسترسی و فهرستنویسی و صرف اعتبار مالی تمام میشود در سالهای اخیر یک مفهوم دو لایه از نامزد کننده ها و داوران به عنوان بخش مهمی از فرآیند جریان کاری به وجود آمد. دیجی بورد جهت تطبیق با این جریان کاری سه سطح کاربری دارد نامزد کننده داور و مدیر نامزد کننده می تواند هر منبعی از وب را برای مجموعه معرفی کند داور قادر است تصدیق کند آیا نامزدها با راهنمای گزینش منبع متناسب است یا نه. البته نامزد کننده مسئولیت فراداده ابتدایی مانند ،موضوع زبان وبگاه و فرآیند مجوزها را نیز دارد. مدیران نیز عضوهای تیم آرشیو وب هستند.

7-2- مدیریت مجموعه

در حال حاضر دیجی بورد دادههای 14,151 وبگاه را در 39 مجموعه آرشیو وب مدیریت میکند و گزارشهایی که از بخش مدیریت کتابخانه کنگره درباره پروژه آرشیو وب درخواست میشود از محتوای پایگاههای اطلاعاتی به راحتی قابل تولید است. دادههای قدیمی از ابزار قدیمی به DigiBoard انتقال یافته است، بنابراین، سوابق گذشته و دادههای جدید در یک سیستم یکتا قابل نگهداری است.

3-7- ارتباطات

زمانی که توسعه و ایجاد DigiBoard رشد پیدا کرد ضرورت ارتباط با کاربران به عنوان یک نیاز شناسایی شد. صفحه اصلی به منظور ایجاد فضایی برای اعلانها تغییر کرد و نوعی سیستم پیغام رسانی برای پیگیری نیز فراهم شد.

8- ماژولهای ویژه

ماژولهای مختلفی جهت پشتیبانی فعالیتهای فرعی در حال توسعه هستند این ماژولها با هدف مدیریت مجموعه ها و مدیریت محتوای آرشیو وب تدارک دیده شده اند.

8-1- ماژول نامزدهای انتخابات ملی آمریکا

ماژول نامزدها در اول جولای 2010 برای نخستین بار شروع به کار کرد این ماژول، آرشیو وب انتخابات را برای علم کتابداری (LS(1)) فراهم میکند و هر دو سال یکبار سایتهای جدید به آرشیو می پیوندند. در گذشته، داده های مبارزه های انتخاباتی فدرال(2) را به منظور بررسیهای بعدی در یک پایگاه داده اکسس وارد میکردند این پایگاه اطلاعاتی نوعی جریان کاری برای پردازش اطلاعات 2200 کاندیدا، پشتیبانی میکرد اما با ابزار تیم آرشیو وب یکپارچه نبود ،بنابراین گرفتن داده از یک پایگاه داده و وارد کردنش به پایگاه داده دیگر مستلزم پردازشهای دستی بود علاوه براین به دلیل اینکه نامزد کننده ها با پایگاه اطلاعاتی اکسس کار میکردند اطلاعات مجوز روکشی برایشان قابل دسترس نبود بنابراین دیگر به تعیین هویت اطلاعات تماس نامزدهایی که قبلاً پاسخ داده بودند نیاز نبود مقصود ماژول نامزدها توسعه ابزاری کارآمد و یکپارچه برای علم کتابداری و تیم آرشیو وب بود.

8-1-1- مجوزهای روکشی نامزدها

مجوزهای روکشی توسط نامزدها مدیریت میشوند؛ بدین ترتیب موجب سهولت مدیریت مجوزها برای نامزدهایی میشوند که در هر انتخابات URL متغییری دارند.

8-1-2- داده های گذشته

دادههای گذشته هر نامزد برای نامزد کنندههای ماژول در دسترس است تمامی داده های خزش شده داده های مبارزه های انتخاباتی فدرال و عملیات انجام شده حفظ خواهند شد و به راحتی در اختیار محققان آینده قرار خواهند گرفت.

1- Library Service

2- Federal Election campaign (FEC)

-8-1-3- فراداده استاندارد

این ماژول قابلیت استانداردسازی فراداده هر نامزد را فراهم می آورد داده مبارزه های انتخاباتی فدرال که به صورت دستی در پایگاه داده ها وارد شده است دارای خطاهایی در ورودی است که این ماژول با گذر ،زمان رکوردهای پایداری از آنها فراهم خواهد آورد.

-8-1-4- یکپارچگی DigiBoard

اگرچه ماژول نامزدها کاملاً با DigiBoard یکپارچه شده است از دیدگاه بصری متفاوت است و برای به انجام رسانیدن تکلیف ویژه ای طراحی شده است شناسایی نامزدها با وبگاههایی که کتابخانه کنگره تمایل دارد به آرشیو اضافه کند زمانی که کار در ماژول نامزدها تکمیل شد رکورد به عنوان رکوردی جدید به DigiBoard نیز انتقال می یابد تا مراحل آرشیو وب آن تکمیل می شود.

9- ماژولهای آینده

ماژولی با نام مدیریت محتوا در دست توسعه است که امکان ردیابی محتوا در زمان تولید آن توسط عامل خزشگر، انتقال به کتابخانه کنگره، کپی شدن در مخزن ذخیره سازی و دسترسی به سرورها را فراهم میکند این ماژول امکانی برای پاسخ دادن به پرسشهای مدیریتی درباره میزان محتوا آرشیو وب کتابخانه کنگره و مکان ذخیره سازی آن را فراهم می.کند این ابزار جهت تکمیل ابزارهای پیچیده ردگیری و ابزارهای سیاهه بندی و قفسه بندی – که توسط کتابخانه کنگره به منظور پشتیبانی از حفاظت رقومی پیاده سازی می شوند – به راحتی قابل اصلاح است.

فعالیتهای توسعه آینده شامل موارد زیر است:

یک ماژول فهرستنویسی که تولید خودکار رکوردهای متس را برای آرشیو وب پشتیبانی میکند؛ بنابراین با در نظر داشتن فراداده ای که توسط ابزار ضبط می،شود از مدت زمانی که توسط فهرستنویسان جهت پردازش این رکوردها مورد نیاز است کاسته میشود. ماژول دیگری نیز مد نظر است که از یک رویکرد انتخابی به منظور شناسایی منابع برای فعالیتهای حفاظت پشتیبانی کند.

10 انجمن آرشیو وب

10-1- ابزار کتابدار وب

ابزار متصدی وب (WCT)(1) برنامه مدیریت جریان کاری متن بازی برای آرشیو وب انتخابی است که در سال 2006 با همکاری مشترک کتابخانه ملی نیوزلند و کتابخانه بریتانیا توسعه یافت و نخستین ابزار کتابدار برای انجمن آرشیو وب شد تا زمانی که WCT برای پتانسیل کاربردیش ارزیابی می شد، کتابخانه کنگره با یک برنامه کاربری پست الکترونیکی برای اطلاع رسانی مالکان محتوا همکاری می کرد. سرانجام،

1- Web Curator Tool

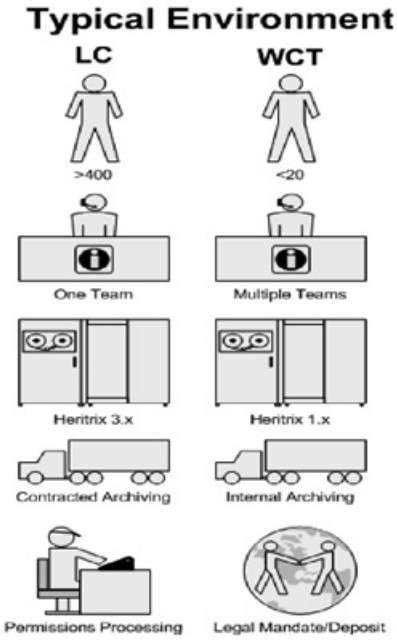
کتابخانه متقاعد شد که WCT به علت برخی تفاوتهای عملیاتی نیازهای آرشیو وب کتابخانه کنگره را مرتفع نخواهد کرد (شکل 8)

شکل .8 مقایسه محیط کاری LC و WCT

التزامهـای پردازش مجوزها در کتابخانه کنگره، تعـداد کارکنان درگیر در مرحله نامزدی و طبیعت اینکه چطور آرشـیو وب کتابخانه ساخته شده است ضرورت ایجاد DigiBoard و ماژولهای وابسته اش را تعیین کرد.

عکس

کتابخانه متقاعد شد که WCT به‌علت برخی تفاوت‌های عملیاتی، نیازهای آرشیو وب کتابخانه کنگره را مرتفع نخواهد کرد (شکل ۸).



شکل ۸. مقایسه محیط کاری LC و WCT

التزام‌های پردازش مجوزها در کتابخانه کنگره،تعداد کارکنان درگیر در مرحله نامزدی و طبیعت اینکه چطور آرشیو وب کتابخانه ساخته شده است، ضرورت ایجاد DigiBoard و ماژول‌های وابسته اش را تعیین کرد.

۱۰-۲- فعالیت‌های گردآوری درون سازمانی

در حال حاضر، برنامه‌ای جهت توسعه ماژولی به‌منظور مدیریت فعالیت‌های گردآوری درون سازمانی[1] با استفاده از خزشگر هریتریکس[2] کتابخانه کنگره وجود ندارد. WCT و نرم‌افزار سوئیت نت آرشیو که توسط کتابخانه ملی واسپاری و کتابخانه رویال دانمارک توسعه یافته است برای استفاده در مراحل مدیریتی خزش درون سازمانی، ارزیابی خواهد شد.

1. On-site
2. Heritrix Crawler

10-2- فعالیتهای گردآوری درون سازمانی

در حال حاضر برنامه ای جهت توسعه ماژولی به منظور مدیریت فعالیتهای گردآوری درون سازمانی(1)با استفاده از خزشگر هریتریکس(2) کتابخانه کنگره وجود ندارد WCT و نرم افزار سوئیت نت آرشیو که توسط کتابخانه ملی و اسپاری و کتابخانه رویال دانمارک توسعه یافته است برای استفاده در مراحل مدیریتی خزش درون سازمانی ارزیابی خواهد شد.

On-site -1

Heritrix Crawler -2

3-10. DigiBoard متن باز

اگر چه DigiBoard به منظور پشتیبانی از جریان کاری کتابخانه کنگره توسعه یافته است، غرض اولیه این ابزار توسعه یک واسط کاربری آسان برای تعداد عظیمی کتابدار است تا فعالیتهای گردآوری را انتخاب و مدیریت کنند نامزدی پردازش ،مجوزها مرور کیفیت و مدیریت خزش این جریانهای ،کاری با جریانهای ،کاری کشورهایی که قانون و اسپاری ندارند .همسوست کتابخانه کنگره تمایل دارد ماژولهای DigiBoard را برای نسخه متن باز آن استاندارد نماید.

11- نتیجه گیری

توسعه ابزار کتابدار که با جریان کاری منحصر به فرد ،کتابخانه نیازمندی مجوزها و محیط کاربران متناسب باشد گامی رو به جلو در آرشیو وب کتابخانه کنگره است خودکارسازی فعالیتهای دستی تیم آرشیو وب جهت افزایش قابلیت مقیاس پذیری فعالیتهای آرشیو وب کتابخانه کنگره بسیار حیاتی است. از طرف دیگر همزمان با بهبود DigiBoard و سهولت استفاده آن برای کاربران مختلف، می توان کارکنان کتابخانه کنگره را در پروژههای آرشیو وب بیشتری درگیر کرد.

12- منابع

[1] http://archive-access.sourceforge.net/projects/wayback/

[2] http://webcurator.sourceforge.net/

[3] http://crawler.archive.org/

[4] http://netarchive.dk/suite/

چکیده

حفاظت و آرشیو کردن وبگاهها چالشی عظیم از حیث فناوری است که باید بی وقفه دنبال شود تحول و تکامل در ،فناوری محتوا و رشد مستمر وب ما را به این معنا هدایت می کند که هیچ راه حل قطعی در آرشیوسازی وب نخواهیم یافت. رونوشت برداری از وبگاهها گزینشی در طراحی وب است. سه روش برای رونوشت برداری کامل از هر وبگاه وجود دارد اولین ،روش آرشیو از سرور، که از همه روشها دشوارتر است. در این روش باید با مدیران سایت تماس گرفت و آنها را متقاعد کرد که رونوشتی از فایلهای سامانه اطلاعاتی داخلی فرانماهای پایگاه داده و ویژگیهای سامانه را تهیه کنند روش دوم این است که این کار را در سرور انجام دهیم و تراکنشهای مربوط به عمل آرشیو کردن منابع وب را ضبط کنیم. آخرین روش، گردآوری خودکار اطلاعات به طور مستقیم از وبگاههاست، به همان ترتیبی که یک مرورگر معمولاً آن را انجام می دهد . در مقاله حاضر به بحث رونوشت برداری از وبگاهها از جنبههای مختلف نظیر روشهای رونوشت برداری ، تجزیه کننده هسته HTML، تجزیه کننده پردازه تجزیه کننده های گروه جاوا ، واکشی مدارک ، تایید وثوق ، اتصال پذیری و مدیریت روزآمد.

رونوشت برداری از وبگاه ها

*رونوشت برداری از وبگاه ها(1)

نوشته: خاویر روش(2) | ترجمه: فرزانه شادان پور(3)

1 – مقدمه فناوری رونوشت برداری از وبگاه ها

تفاوت اساسی میان رونوشت بردار(4) از داده ساختارها(5) یا سایتهای ftp و وبگاهها در طبیعت ژرف وب جهانی نهفته است. این نه به ویژگی «فهرست راهنما» بودن در پروتکل HTTP مربوط می شود، نه به ویژگی انتقال حجم انبوه و بگاهها؛ بلکه گزینشی در طراحی وب است که خود مجموعه ای است از منابع ناهمگن که لزوماً با یکدیگر در ارتباط نیستند. به عنوان مثال، مجموعه ای از صفحاتی که در محتوای یک پایگاه داده تولید میشوند قلمرو بی ثباتی از اطلاعات است. پایگاه داده که پاکسازی(6) شود، صفحات مذکور نیز ناپدید میگردند ،وب به طور کلی شکلی در حال تغییر است راهنماهای ftp تغییر می کنند ولی شما هرجا که هستید به آسانی میتوانید آنها را همزمان(7) کنید تا روز آمد باشند. این فقط مقداری دادهٔ ذخیره شده در یک انباره(8) فایل است. ولی یک صفحه وب بالقوه منحصر به فرد است ممکن است یک

1- Roche. Xavier (2006). "Copying websites", in Masanes, Julien (ed.), Web Archiving. Berlin Heidelberg .New York: Springer.pp.93-113

2- Roche. Xavier

3- مربی، عضو هیئت علمی سازمان سناد و کتابخانه ملی ایران

4- Coping

5- File structure

6- Flush

7- Synchronize

8- Repository

ساعت شمار(1)، اطلاعـات تحویلی بی درنگ بر حسب تقاضا یا نمایی(2) ویژه کاربر یا ویژه تراکنش(3) از یـک مجموعه کلیتر از داده باشـد میتوانـد هر چه شـما میخواهیـد باشد منطق درونی آن از ابزارهای کاربری و نیز کاربران این ابزارها پوشـیده است و ساده بگوییم یک سرور ftp مجموعه ای از فایلهاست که بیشتر به دیسک سختی که عمومی و قابل کنترل از راه دور است شباهت دارد.

یک سـرور وب را میتوان مجموعه ای از منابع منطقی(4) دانست که به مشتری ها محتوا تحویل می دهند. این منابع منطقی ممکن اسـت برنامههـایی باشـند که به پایگـاه داده و سـایر سامـانه هـا از طریق تراکنش بـا ترجیحات و نیازهای کاربر و محیط سـرور (پایگاه داده منابع ،بیرونی وضـعیت فعلی و غیره) متصل اند. مشتری راه دور هرگز این منطق را مشاهده نمی کند و فقط محتوای به دست آمده قابل دسترسی اوست.

بنـابراین سه روش برای رونوشت برداری کامـل از هر وبگاه وجود دارد نخستین آرشـیو از سـرور(5) از همه روشـها دشوارتر است در این روش بایـد با مدیران سایت تماس گرفت و آنها را متقاعد کرد که رونوشتی از فایلهای سامانه اطلاعاتی ،داخلی فرانماهای(6) پایگاه داده و ویژگیهای سامانه را تهیه و همان معماری را از حیث سـخت افزار نرم افزار و محیط رایانشـی (مانند منابع داده ای که میتوانند مورد اسـتفاده قرار گیرند) در آن ترتیب دهند.

این راه حل را که معمولا حتی برای خود مـدیران سایت نیز بسـیار دشوار است نمی توان به عنوان یک راه حل عمومی در نظر گرفت. گزینه دوم این است که این کار را در سرور انجام دهیم و همهٔ تراکنشهای(7) مربوط به عمل آرشیو کردن منابع وب را ضبط کنیم. آخرین روش، گردآوری خودکار اطلاعات به طور مسـتقیم از وبگاههاست به همان ترتیبی که یک مرورگر معمولاً آن را انجام میدهـد (آرشـیو از سـمت کاربر(8)). این یک راه حل موقت است چرا که رونوشـتها هرگز کامل نیستند مثل این است که از یک صحنه متحرک یک عکس بگیریم در این صـورت نمیتوانیـد حرکت آن صـحنه را در عکس ایجـاد کنیـد. همچنین نخواهیـد تـوانست امتیـاز بی درنـگ بودن را هنگـام مرور در گزارشـهای برخط یا جنب و جوش مبادله اطلاعات را داشـته باشـید ولی این از حیث امکان اجرا و کیفیت در بیشتر موارد ناگزیر پذیرفتنی است. مثل عکس ایستایی(9) از یک وبگاه است که باید در یک آلبوم عکس نگهداری شود؛ عکسی که بارها و بارها بشود آن را ،دید حتی بـدون نگرانی از این که اصل آن دیگر وجود نـدارد. رونوشت برداری از وبگاهها با این روش امری بسـیار شـهودی است روش دقیقاً همان است که گویی خودتان بخواهید با یک مرورگر از یک وبگاه رونوشت برداری کنید در این صورت از نخستین صـفحه آغاز میکردید صفحه و تصاویر آن را ذخیره و سپس بر هر پیوند در صفحه کلیک میکردید تا

1- Clock Counter

2- View

3- Session

4- Logical

5- Server – side

6- Schemas

7- Transaction

8- Client-side archiving

آنها را ببینید و صفحات مرتبط را در یک دیسک ذخیره میکردید و آنقدر این کار را ادامه میدادید تا از همه صفحاتی که میخواهید رونوشت بردارید. بعد تغییراتی در صفحات HTML می دادید و همهٔ تگهای مربوط را وارسی میکردید تا با مرورگر سیستم خودتان قابل دیدن شوند. ولی رونوشت برداری از بیش از یکی دو صفحه به صورت دستی خسته کننده است و داشتن ابزاری برای خودکار کردن آن می تواند راه حل مؤثری باشد. شمارنده خودکار پیوندها را معمولا تجزیه کننده(1) و نرم افزاری را که داده را به طور خودکار و از راه دور بارگذاری میکند خزش گر مینامند. این دو جزء اصلی نقشهای دیگری هم دارند. تجزیه کننده عهده دار بررسی و تضمین این است که پیوندها در یک رونوشت محلی (روی سیستم) نیز کار کنند، و این کار با تغییر نحو(2) یوآرال آن به نحوی صحیح و کاملا مرتبط انجام پذیر می شود و خزش گر مسئول حافظهٔ دم دستی(3) و روزآمد کردنهاست.

دلایل متعددی برای رونوشت برداری از وبگاهها وجود دارد در مدرسه ملی مهندسی کن(4) ما میخواستیم وبگاههایی با اندازه کوچک و متوسط را آرشیو کنیم، نه برای آرشیو سازی رایج، بلکه برای گردآوری سایتهای فنی که اشخاص راه میانداختند و سریع هم دستخوش تغییر می شدند. همچنین میخواستیم وبگاههای بزرگ را که دارای محتوای چند رسانه ای بودند و با استفاده از خطهای dial-up خانگی قابل دسترسی نبودند، گردآوری و آنها را در رسانه های دائمی مانند CD ذخیره کنیم تا بعد بتوانیم آنها را به صورت غیر برخط ببینیم. در مجموع به ابزاری برای گردآوری اطلاعات بسیار ویژه از وب برای کاربران نهایی نیاز داشتیم.

طرح HTTrack برای پاسخ به این نیازها پدید آمد این نرم افزار ابزاری سهلال الاستفاده است که به کاربران معمولی اجازه میدهد از بخشهای کوچک – ولی مهم وب رونوشت برداری کند طراحی این نرم افزار بیشتر به طور تجربی صورت گرفته .است اینترنت و معماری شبکهٔ مربوط به آن عرصه های نسبتاً جدیدی بودند که ما در آنها به اکتشاف می پرداختیم و رونوشت برداری از وبگاه به ویژه و به طور خاص موضوع کاملاً جدیدی برای ما بود تجربیاتی که از تدوین و توسعه این نرم افزار به دست آمد روش – «فناوری(5)» رونوشت برداری از وبگاهها و راه حلهای برطرف کردن نقاط ضعف موجود را تشریح می کند.

2- تجزیه کننده

2-1- تجزیه کننده هسته HTML(6)

تجزیه کننده HTML یکی از دو جزء هسته در یک ابزار رونوشت برداری(7) از وب است. اگر یک صفحهٔ

1- Parser

2- Syntax

3- Cache

4- National School of Engineering of Caen

5- Art

6- Core parser.

7- Web copying tool

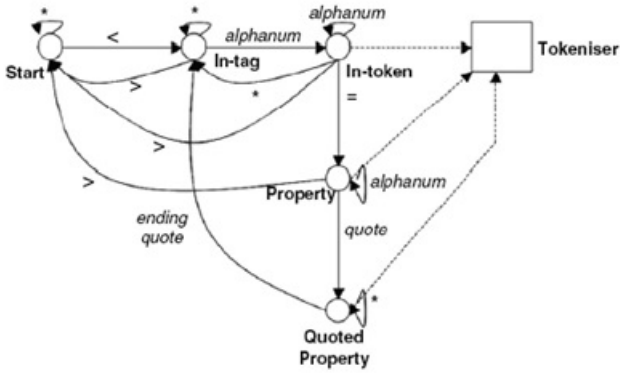
HTML(1) را در نظر بگیریم که یک فایل متنی 8 بیتی(2) متشکل از متن ساده (بدون رمز)(3) و تگهای نشانه گذاری(4) است- و اطلاعات همراه آن مانند URL اولیه(5) هدف تجزیه کننده HTML است تا با پویش(6) صفحه پیوندها را گردآوری و تجزیه و تحلیل کند و آنها را به خزشگر بسپارد. ساختار HTML برای گردآوری پیوندها مناسب نیست ما در ابتدا به تعداد محدودی از تگ ها نظیر عناصر «a» یا «img»، علاقه مندیم که بالقوه ممکن است حاوی ابرپیوندهایی به سایر منابع (تصاویر style sheets، صفحه های HTML ، و مانند آن) باشند جای قرار گرفتن آنها در صفحه معمولاً مهم نیست، مگر در دامنه های خاصی که هیورستیک پیشرفته(7) میتواند اطلاعات اضافی (مانند موضوعات «حول و حوش») به آن ضمیمه کند. این تگها معین میکنند کدام صفحهها تعقیب شوند و کدام صفحه های نامرتبط تعقیب نشوند. برای یک تجزیه کننده معمولی تنها اطلاعات مهم تگها و متعلقات جاسازی شده آنها هستند.

شکل ساده شده اتوماتون(8) هسته، به خوبی و سادگی قابل درک است پویش خطی(9) دادههای صفحه HTML، شروع از اول کشف تگهای آغازین (<) و بازشناسی(10) عناصر مختلف HTML با نامشان(نگاه کنید به شکل 1).

عکس

HTML [۱] را در نظر بگیریم- که یک فایل متنی ۸ بیتی [۲] متشـــکل از متن ســـاده (بدون رمز)[۳] و تگ‌های نشـــانه گذاری[۴] اســـت- و اطلاعات همراه آن مانند URL اولیه[۵] هدف تجزیه‌کنندهٔ HTML اســـت تا با پویش[۶] صفحه پیوندها را گردآوری و تجزیه و تحلیل کند و آنها را به خزشگر بسپارد. ساختار HTML برای گردآوری پیوندها مناسب نیست. ما در ابتدا به تعداد محدودی از تگ‌ها نظیر عناصر «a» یا «img»، علاقه مندیم که بالقوه ممکن است حاوی ابرپیوندهایی به سایر منابع (تصاویر، style sheets، صفحه‌های HTML ، و مانند آن) باشند. جای قرارگرفتن آنها در صفحه معمولاً مهم نیست، مگر در دامنه‌های خاصی که هیورستیک پیشرفته[۷] می‌تواند اطلاعات اضافی (مانند موضوعات «حول وحوش») به آن ضمیمه کند. این تگ‌ها، معین می‌کنند کدام صفحه‌ها تعقیب شوند و کدام صفحه‌های نامرتبط تعقیب نشوند. برای یک تجزیه‌کنندهٔ معمولی تنها اطلاعات مهم، تگ‌ها و متعلقات جاسازی شدهٔ آنها هستند.

شکل ساده شدهٔ اتوماتون [۸] هسته، به‌خوبی و سادگی قابل درک است: پویش خطی [۹] داده‌های صفحهٔ HTML ، شـــروع از اول، کشف تگ‌های آغازین (<) و بازشناسی[۱۰] عناصر مختلف HTML با نامشان (نگاه کنید به شکل ۱).



شکل ۱. هستهٔ تجزیه کننده

۱. نگاه کنید به [۱۸۶۶]

۲ . به یاد داشته باشید که رمزگذاری کاراکتر صفحه برای نامگذاری (naming) به خصوص در فایل سیستم‌های UCS۲ (ازجمله ویندوز) اهمیت خواهد داشت.

3. Plain text
4. Markup tags

۵ . نگاه کنید به [۱۷۳۸]

6. Scan
7. Advanced heuristic

۸. Automaton : ماشین خودکار، کامپیوتر

9. Linear Scan
10. Recognizing

شکل 1. هسته تجزیه کننده

ص: 188

1- نگاه کنید به [1866]

2- به یاد داشته باشید که رمزگذاری کاراکتر صفحه برای نامگذاری (naming) به خصوص در فایل سیستمهای UCS2 از جمله ویندوز اهمیت خواهد داشت.

3- Plain text

4- Markup tags

5- نگاه کنید به [1738]

6- Scan

7- Advanced heuristic

8- Automaton : ماشین خودکار کامپیوتر

9- Linear Scan

10- Recognizing

دو دسته از محتویات درون تگ های HTML قابل بازشناسی است نام تگها(1) مانند img" یا "a" و ویژگیهای تگ مانند herf یا "src" این تگها را میتوان به دو گروه اصلی تقسیم کرد تگهایی که امکان جاسازی(2) منابع (نظیر ،تصویر style sheetsهایی که در صفحه وجود دارند) و تگهایی که ناوبری به سایر منابع (ابرپیوند) را میسر می سازند در یک صفحه میتوانید از پیوندهای نامرتبط در گروه دوم صرف نظر کنید (برای مثال پیوندهایی که خارج از دامنه رونوشت آیینه ای(3) هستند). در این صورت، این یک محیط برون خطی غیر قابل دسترسی خواهند بود ولی این مسئله باعث تغییر صفحه نمی شود در مورد تگهای گروه نخست باید دقت بیشتری به خرج دهید چرا که در غیر این صورت صفحه در حالت غیر برخط به درستی قابل دیدن نخواهد بود ممکن است تصاویری را از دست بدهید یا چینش(4) صفحات به علت اجزای از دست رفته مانند style sheets یا فایلهای پردازه نویسی(5) جاسازی شده به هم ریخته باشد بنابراین یوآرالهای پیوند تنها اطلاعاتی نیستند که باید به خزشگر داده شوند. بافت تگ(6) ، مثلاً اینکه آیا منبع یک منبع جاسازی شده است یا نه؟ هنگام تصمیم گیری در مورد این که این پیوند را بگیر یا خیر اهمیت خواهند داشت.

Tokenizer پیوندها را با تجزیه و تحلیل ویژگیهای شناخته شده شان برداشت میکند. پیوندها با استفاده از URL صفحهٔ اصلی به شکل مطلق(7) تبدیل میشوند که این قسمتهاست پروتکل http»، میزبان: "//www.example.com" و مسیر مربوط index html . به عنوان مثال پیوندهای نسبی(8) (/top.html» در درون صفحهٔ [مثال]

http:/www.-example.com/foo/index.html»

تبدیل می شوند به پیوند «http: / www. example.com /foo/top.html». سپس جای پیوند بررسی میشود تا از منطبق بودن با دامنهٔ پیش فرض رونوشت آیینهای اطمینان حاصل آید؛ بررسی شامل عبارت(9) معمولی است که مقدار آن به طور پیش فرض پیشوند(10) اصلی URL .است اگر گرفتن رونوشت آیینه ای را از «http://www.example.com/foo/index.html» آغاز کرده باشیم دامنهٔ پیش فرض در نحو شبه معمول عبارت خواهد بود:

*/http://www.exampel.com/foo

غیر از این پیوندهایی نظیر

,http: / www. example.com /foo/top.html

1- Tag names

2- Embed

3- Mirror

4- Layout

5- Scripting files

6- Tag context

7- Absolute Form: نگاه کنید به:Sect [2396]4.4Hierarchical URL aud Relative Forms

8- : نگاه کنید [1808]8. Relative links.

9- Expression

10- Prefix

به طور پیش فرض در رونوشت آیینهای داخل خواهنـد شـد. البته بسـته به سایتی که قرار است رونوشت برداری شود، قواعد بیشتری ممکن است لازم باشـند بنابراین عبارت پیش فرض – بسـته به نیاز – باید قابل تغییر باشد. سرانجام پیوندهای تکراری نباید به خزشگر منتقل شوند تجزیه کننده باید وضعیت همه URL های شناخته شده را بداند و از چندبار گرفتن پیوندها بپرهیزد.

تجزیه کننده همچنین باید بتواند با نحوهای متعددی کار کند که ممکن است مرکب از اشـکال نسبی یا مطلق URL،ها گریز (1)HTML) (مانندnbsp)گریز یو آرال(2) (مانند3aدرصـد) و به طور کلی هرگونه نحو یله ای باشـند مرورگرها تا حد زیادی با این نحوها مدارا میکنند حتی وقتی صـفحه دچار از هم گسیختگی است (از جمله خطا در نحو تگها) مرورگرها اغلب نهایت تلاش میکنند تا آن را تجزیه و تحلیل کنند و به شکلی که قابل درک باشد بالا بیاورند.

به عنوان مثال پیونـد URL مطلق http:www.Example.com /page 2 html " را میتوان با نحوهای متعـددی از جمله نحوهای غلط ارجاع داد به هر شکل URL را باید بتوان بازشناخت و آن را به حساب آورد.

در پایان، پیونـدها بایـد بازشناسـی شونـد تا با ساختار رونوشت آیینهای وبگاه هماهنگ شود. لازم است پیوندهایی که شـکل مطلق دارند مانند

" http:www.example.com /index.html به شـکل نسبی مبدل شوند مانند index.html» پیوندهایی که خارج از دامنه رونوشـت آینه ای هسـتند پیوندهایی که با دامنه عبارت پیش فرض جور نیسـتند باید به شکل مطلق تبدیل شوند بنابراین در صفحه های رونوشت آیینه ای باید تغییراتی داد تا در یک ساختار محلی قابل استفاده شوند.

2-2- تجزیه کننده پردازه(3)

چند ماه پس از آغاز تدوین و توسعه HTTrack و علیرغم بهبودی که در تجزیه کننده های HTML رخ داده بود ملاحظه میشد که بعضی وبگاهها به درستی رونوشت برداری نشدهاند و تعداد زیادی از تصاویر و فایلهای صفحات در رونوشتها وجود نداشت و این باعث خطا در ناوبری میشد فقط به این علت که تجزیه کننده پیوندهای مربوط به آن تصاویر و فایلها را ندیده بود.

درون صفحات HTML باید مناطق(4) پردازه نویسی(5) خاصـی مانند جاوا (کد فعال داخل شده درصفحات) در نظر گرفته میشدند که عملکرد تجزیه ویژه ای برایشان لازم بود پرداخت کامل کد پردازه تقریباً غیر ممکن است کد با تگهای HTML متفاوت است. آنها مشابه تگهای HTML نیستند که بسیار آسانتر میتوان تجزیه و تحلیل شان کرد منطق نهفته در متغیرها توابع و عبارات ممکن است

1- [HTML escaping: "Proposed Eutites".14.sect [7666]

2- نگاه کنید: [1630]URL escaping .

3- Script arser

4- Zones

5- : نگاه کنید به [ECMA-262 Scripting ECMA Script scripting generalization]

بالقوه غیر قابل دریافت باشد نخست اینکه حتی با تفسیر کننده(1) جاوای کامل، عملیاتی که یکباره به علت وضعیت قرار گرفتن موشواره انجام میشوند کلیک بر عناصر و اجزای صفحه، یا محیط (زمان، متغیرهای کاربری و به طور کلی آنتروپی محیط و مانند آن) را نمیتوان به دست آورد. دوم اینکه گرفتن پیوندها با استفاده از تفسیر کننده مسئلهٔ دیگر را حل نخواهد کرد که آن تغییر منطق کدها برای انطباق با سایت رونوشت برداری شده است و اگر چه این کار درون تگ HTML ساده است انجام آن درون یک کدپردازه نویسی پیچیده تقریباً غیر ممکن است.

خوشبختانه در بیشتر موارد کدهای جاوای به کار گرفته شده آنقدر ساده هستند که با توان محدود برنامه بتواند جور در بیایند برای بارگیری خودکار تصاویر یا برای گرفتن آنها در تصویر زمینه طراحان نرم افزار معمولا از ارجاع مستقیم به ویژگی شیء با استفاده از رشته های ایستایی(2) نظیر foo.src bar grf استفاده میکنند یا برای گشودن یک پنجره جدید عبارتی مانند window. Open (foo.html) را به کار میبرند. حدود 80 درصد از پیوندهای پنهان درون مناطق پردازه را با به کار بردن این نمونه های ساده میتوان کشف کرد و تغییر داد سایر موارد با استفاده از عبارتها یا روشهای ناشناخته همانطور که هستند باقی میمانند نتیجه عالی نخواهد بود و در مورد HTTrack ، از ابتدا میدانستیم که همه چیز آنطور که میخواهیم پیش نخواهد رفت هدف رسیدن به سطح قابل قبولی از کیفیت برای بیشتر سایت ها بود.

مناطق CSS(3) را میتوان با الگوریتمهای مشابهی تجزیه کرد.

اتوماتون ساده شده زیر برداشت رشته های(4) متنی درون ناحیه های پردازه نظیر بخشهای تگ