مدیریت منابع اطلاعاتی وب

جلد دوم

دیدگاه‌های فناورانه، اخلاقی و مدیریتی

به کوشش

دکتر غلامعلی منتظر

فرزانه شاه‌طالبی 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

مدیریت منابع اطلاعاتی وب 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 [10]، فن خسار [11] و نجدی[12] (۱۲) و پاسترناک [13] و راث[14] [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

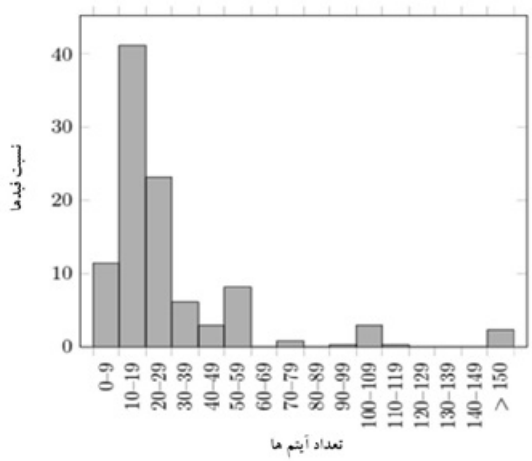
4- RDF

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

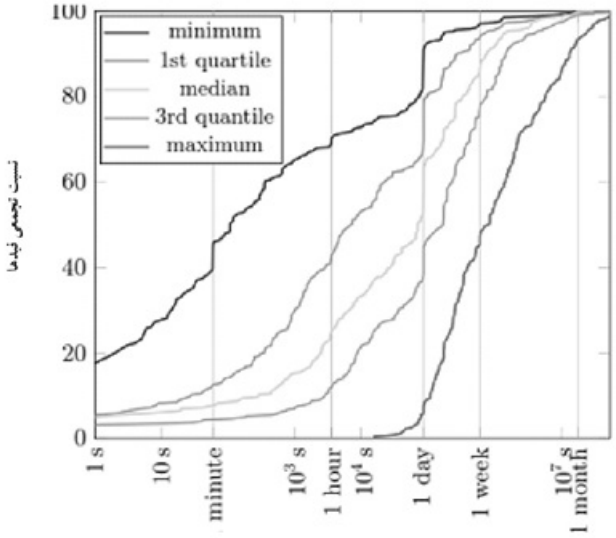
ص: 13

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.

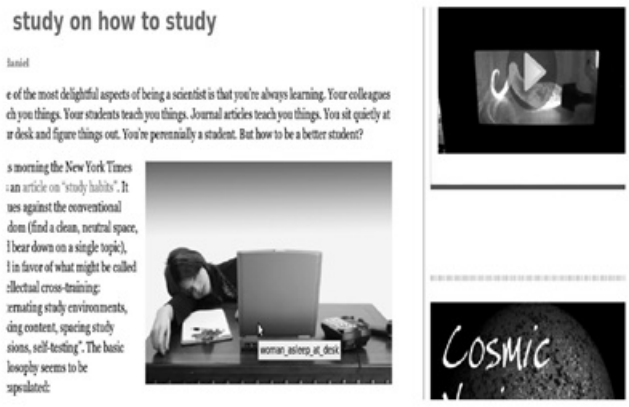
عکس

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.

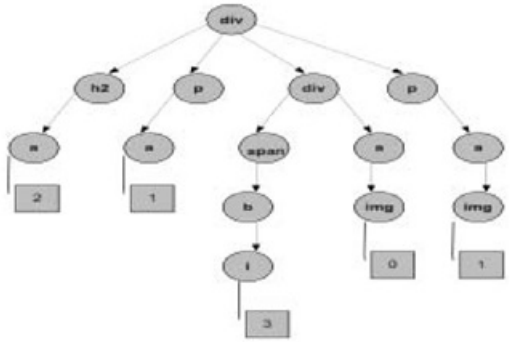
1- Delightful Aspects

2- Being a Scientist

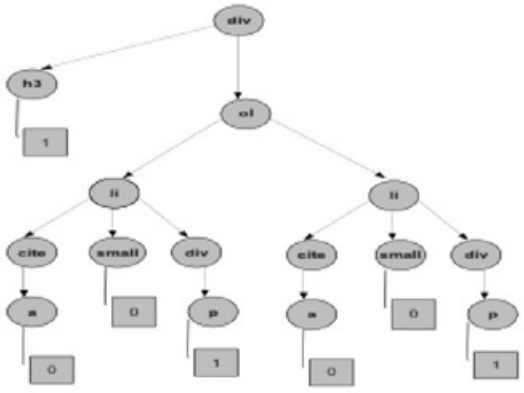
3- This morning the New

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.

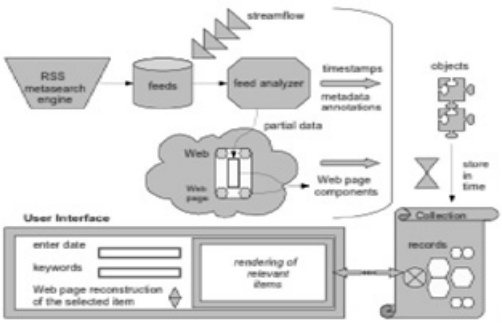
علاـوه بر این، بـا اسـتفاده از الگوریتم (احتمالاًـ برخی از فناوریهای هوشـمند)، میتوانیم نه تنها گره DOM را – که شامـل مقـاله است – شناسـایی کنیم بلکه سـایر گره هـای معنایی صـفحه را که شامل نظرها مقوله ها یا تگ هاست نیز شـناسایی کنیم این مناطق میتواننـد برای محتوای متنی و منابع به طور مسـتقل از تمپلت 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

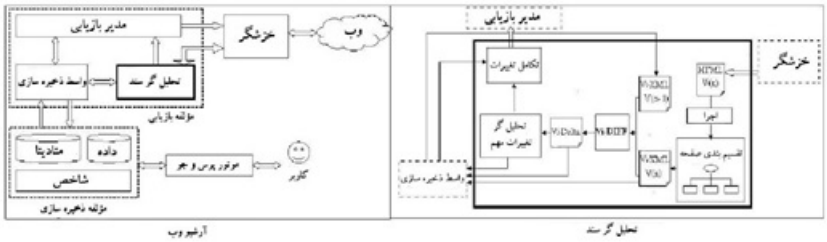
یکی از وظایف INA ایجاد ذخایر قانونی[1] است که صفحات وب رادیو وتلویزیون فرانسه و صفحات مرتبط آن را نگهداری می‌کند. یکی از نیازهای این پروژه، نگهداری جنبه دیداری صفحات است. بنابراین ایده ما به‌کارگیری تحلیل دیداری صفحه برای نسبت دادن اهمیت به بخش‌های مختلف صفحه وب بر مبنای موقعیت نسبی آنهاست. مفاهیم تحلیل 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

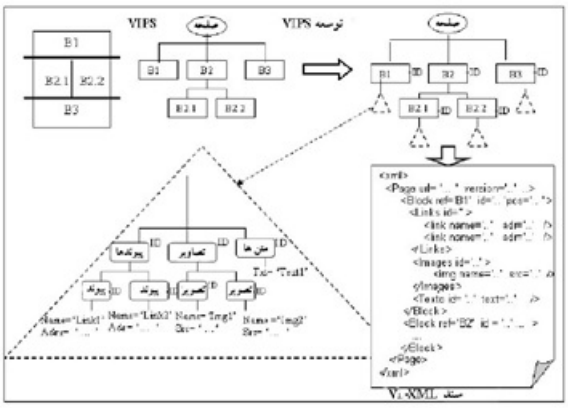
همانطور که قبلا ذکر شـد ، [2] VIPS برای تقسـیم بنـدی صـفحه وب به بلاکهای معنایی تو در تو بر مبنای گره های مناسب در درخت HTML DOM در صـفحه به کار میرود و جداکننـده های(1) عمودی و افقی در صـفحه را مشـخص میکنـد بر مبنای این جداکننـده ها درخت معنایی صـفحه و بی را که به بلاکهای متعدد تقسـیم شده است تولید 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:

$$\sum_{i=1}^{N_{Bk}} I(Bk_i) = 1; 0 \leq I(Op) \leq 1$$

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 نسـخه ها بر مبنای زمان بندی خزش است. مهمترین وظیفه زمان بند عبارت است از خزش به مهمترین نسـخه اسـناد برمبنای تاریخچه تغییرات زمان بند، فهرستی از اسـناد مرتب شده توسط یک تابع ضـرورت بازیابی(1) را مدیریت می.کند این تابع برای هر صفحه

Freshness urgency function -1

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.

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

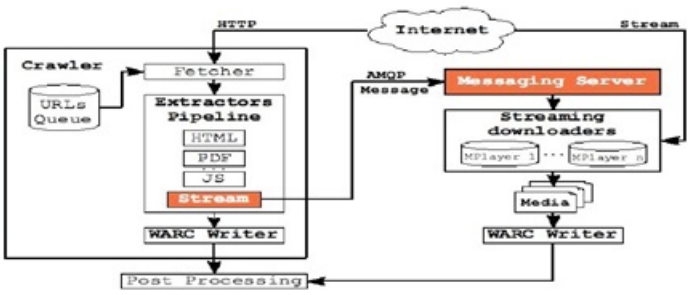
تصویر3 – ماژول ضبط داده در جریان در تعامل با خزشگر

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.

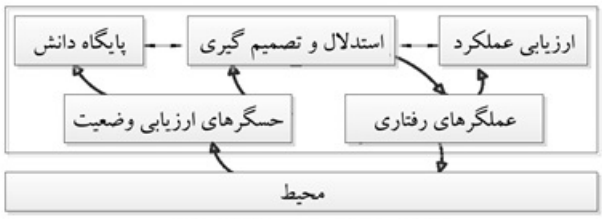
2. عاملهای هوشمند

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.

2-3 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.

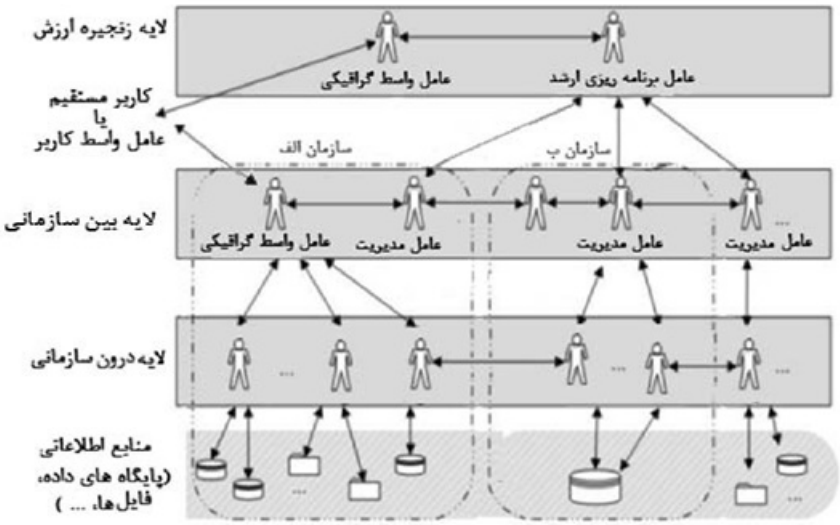
4. پیشنهاد چارچوبی مبتنی بر عامل در یکپارچه سازی مدیریت اطلاعات سازمانها ایجاد یکپارچگی و امکان تعامل پذیری در سازمانها نیازمند فراهم کردن چارچوبی منطقی برای جریان یافتن اطلاعات در میان بخشـهای مختلف و نیز تصمیم گیریهای مشارکتی است به این منظور لازم

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 شـــود. استفاده از شـــیوه‌های 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.

بـــا در نظر گرفتن این موضوع که تعاملات ســـازمان‌ها علاوه بر اطلاعـــات، لایه فرایند را نیز در بر می‌گیرد،معماری‌های سلسله مراتبی نمی‌تواند جوابگوی نیاز آنها باشد. از طرفی برای برقراری امنیت در تعاملات ســـازمان 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 شـــبکه‌ای 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 بایـــد بتواند در فرایندهای برنامه ریزی، زمان‌بندی و کنترل با پیشـــامدهای 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:

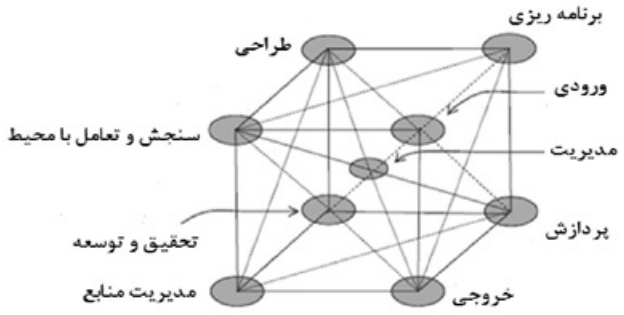
- لایه زنجیره ارزش: وظیفه این لایه ایجاد یکپارچگی در طول زنجیره ارزش و برقراری ارتباط با کاربر نهایی است. با توجه به درخواست کاربر و موقعیت سازمانها این لایه وظایف سطوح پایینتر را مشخص کرده و به آنها اعلام میکند عامل برنامه ریزی ارشد که وظیفه اصلی برنامه ریزی را بر عهده دارد در طول زنجیره ارزش یکتا است. 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:

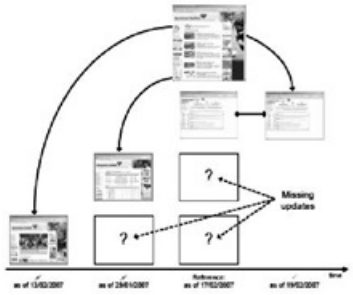
1)تصویر سـازی تغییرات در درخت Spanning خزش با2 ; visone)تصویر سازی Scatterplot در تحلیـل محتوای تغییر؛ 3)منطقه مکانی تصویر سازی سری زمانی تغییرات؛ و 4)تصویر سازی Scatterplot سری زمانی تغییرات.

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- وFrom_Site_id,To_Site_id,Crawl_id,From_url_id,To_url_id,Visited_TimeStamp) 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.

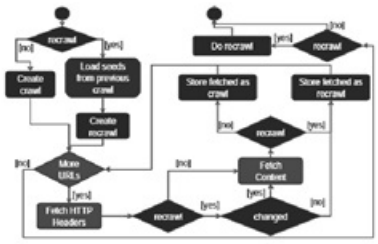
عکس

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

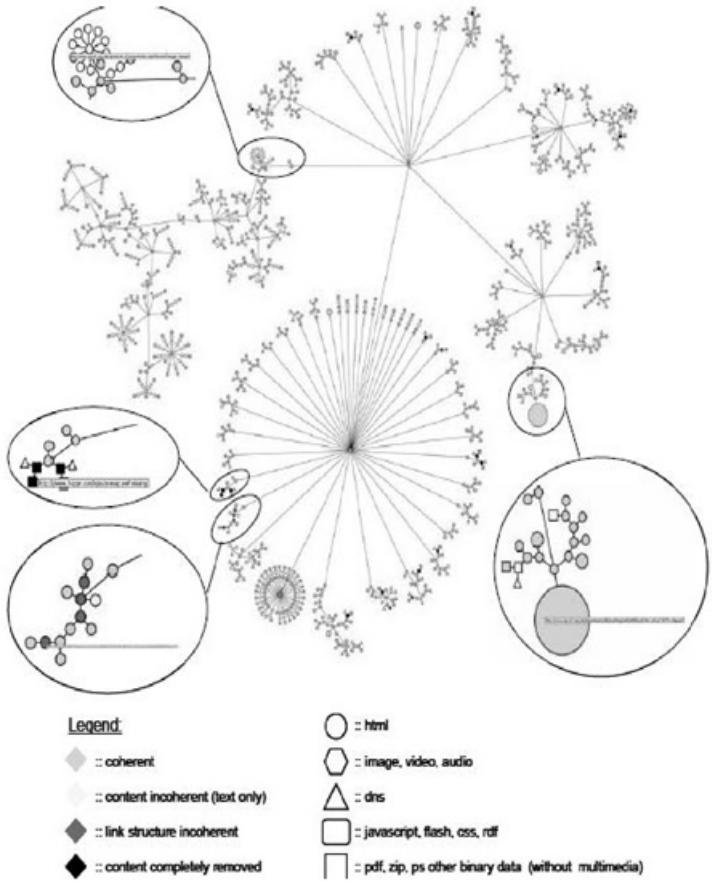
مرتبط قادر خواهند بود درخت پوشا و ظاهر نقص در انسـجام را نمایش دهند. این تصویر سازیها به عنوان وسیله ای اضافی برای آمارهای خودکار جهت کشف مشـکلی که در حین ثبت رخ میدهد در نظر گرفته می.شود. قسـمت اصلی این برنامه تحلیل با کیفیت از خزش یک وبگاه است. شکل 3 تصویر ساده ای از یک خزش از نشانی mpi-inf.mpg.de 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.

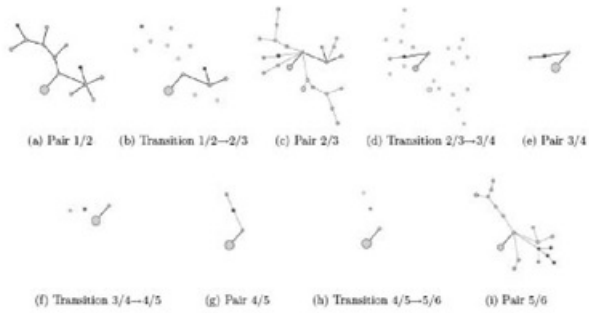


Figure 4: Tracing of coherence defects in crawl-recrawl pairs of the dmoz.org/news subdomain over time

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 وroyal-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(b)-٦(a) 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

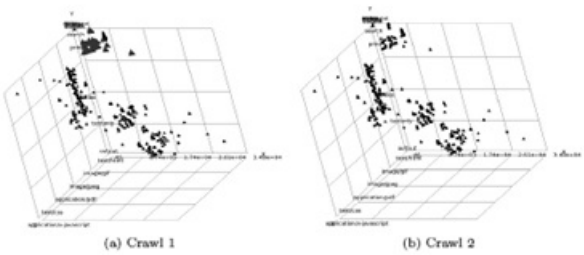
بالای تصویر) تغییرات صـفحه ها (چهار نقطه (قرمز) وابسـتگیهای بین صفحه ها را نمایش میدهد. اگر صفحه ای در مسیر خروجی تغییر کند صـفحه منطبق با آن در آن مسیر نیز تغییر خواهد کرد. صفحه – هایی که تازه اضافه شده اند نشان میدهند که ضوابط ساختاری سایت متحمل تغییرات می شوند. اگرچه محتویات سایت خیلی تغییر نکرده 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

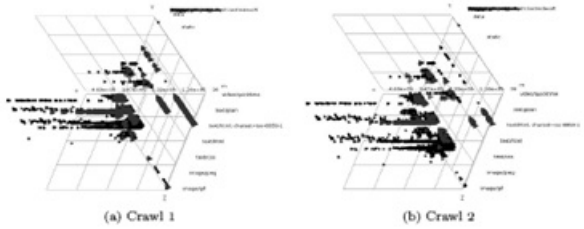
ساختار وبگاه را نشان میدهد.

عکس

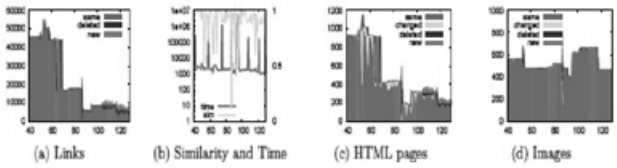
ساختار وبگاه را نشان می‌دهد.



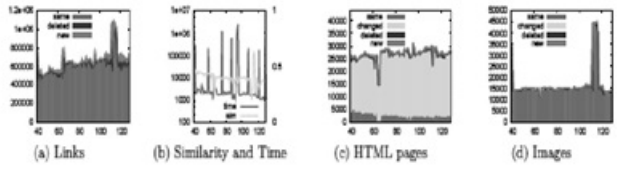
شکل ۵. تحلیل توالی خزش www.sabre.m.d.uksits



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

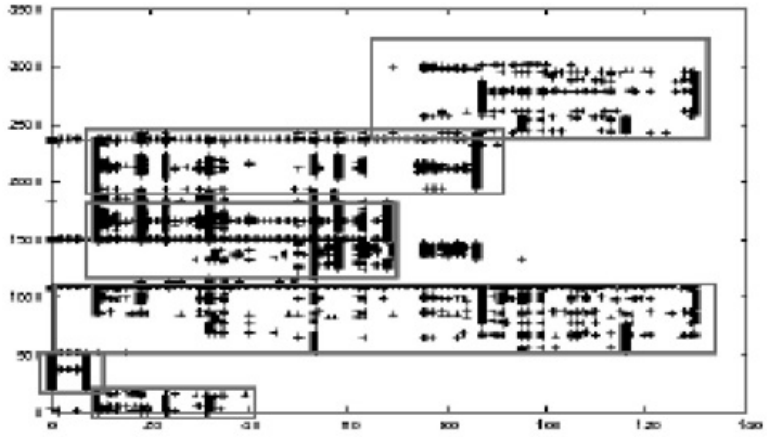


Figure 9: Scatterplot of lines for www.sabre.mod.

شکل ۹. الگوی تغییرات صفحه های یک سایت

```
create table t_pages_dd as
select t_pages.* from (
  select crawl_id, url_id, max(visited_timestamp)
         as latest_timestamp
  from t_pages
  group by crawl_id, url_id) as x, t_pages
where t_pages.crawl_id = x.crawl_id and
      t_pages.url_id = x.url_id and
      t_pages.visited_timestamp
           = x.latest_timestamp

create table t_links_dd as
select t_links.*
from t_pages_dd, t_links
where t_pages_dd.url_id = t_links.from_url_id and
      t_pages_dd.visited_timestamp
        = t_links.visited_timestamp
```

فهرست ۲

فهرست3 . SQL به کار رفته برای پاکسازی URLهای پایین تر

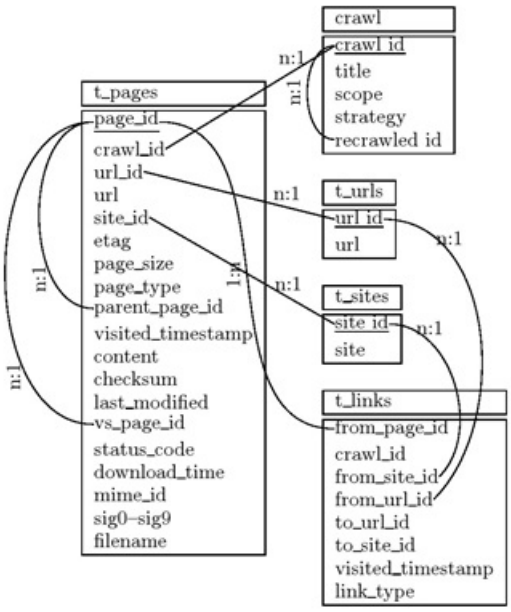
نمودار 10 شمای DB

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.درسـهای آموخته شده و کارهای آینده از نظر یک آرشـیویست آرشیو کردن مطلوب ،وب جلوگیری از تغییرات محتویات در حین عمل خزش است. البته این یـک توهم و عملاـ نشـدنی است در نتیجه ممکن است یـک نفر هرگز مطمئن نشود که محتویـات که تا کنون جمع کرده است هنوز با محتویاتی که بعـدا جمع خواهد شد منطبق است. به هر حال انسـجام در آرشـیو وب یک موضوع کلیدی برای انسـجام خزش جهت دادههـای رقومی، در یـک حالت قابل تکثیر و تفسـیر است به این منظور ما گستره ای از Heritrix که با ارتباطات صحیح و همچنین تاریخ نامناسب محتویات سازگار است را توسـعه داده ایم. به علاوه ما قادر هستیم شـکل انسجام را مؤثرتر بدون توجه به تکیه بر وب سـرور تحلیـل کنیم به همین ترتیب توسـعه تحلیل و تصویرسازی ویژگیها در کمک به مهنـدسان خزش برای درک بهتر ذات نقص در انسـجام درون و بین وبگاهها و سازگاری راهبردهای Crawling برای خزشـهای آینده مفید است در نتیجه مقاله به افزایش انسـجام 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 پیشـرفت به عمل خزش ناتمام و افزایش پوشـش آرشـیو کمک میکند حتی ترکیب بخشی از Recrawl در ترکیب بـا یـک بخش افزایشـی خزش ممکن است جـذاب و مؤثر شود. به علاـوه نتیجه به دست آمـده از تحلیل خزشـهای 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 کـار توسـط هفتمین برنـامه FrameworkIST از E توسـط تمرکزی کوچـک یـا متوسـط پروژه هـای 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 پیایندها 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

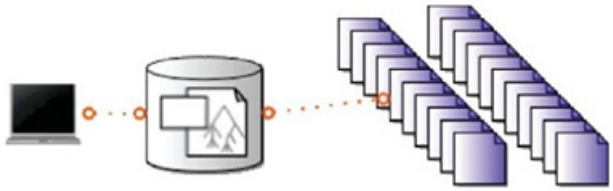
از ترکیب نحوی XML[1] به آشکارسازی ابرداده بر روی HTTP می‌پردازد. پیاده‌سازی در سطح پیمان‌نامه، باعث ارتباط واقعی از طریق درخواست و پاسخ سرورمی‌شود، و از این طریق احتمالات افزایش می‌یابد. به عنوان مثال، امکان پرس‌وجو از سند به کمک تاریخ و نوع اسناد فراهم می‌شود. خزشگر می‌تواند به‌طور مستقیم با سرور OAI 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

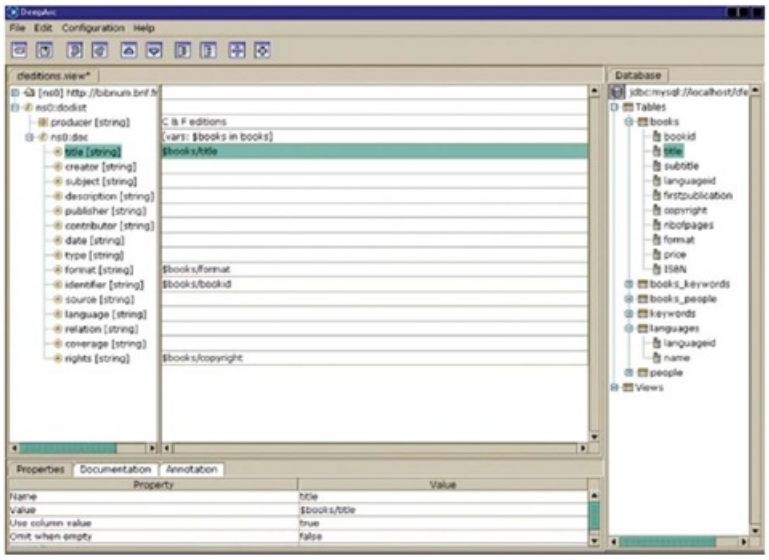
Liu -2

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- MARCXML (Machine-readable Cataloguing in XML)

(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

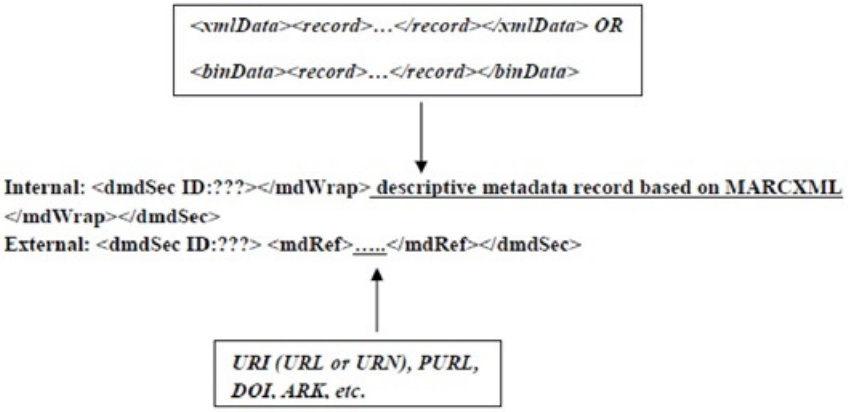
3- Network Development and MARC Standards Office

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

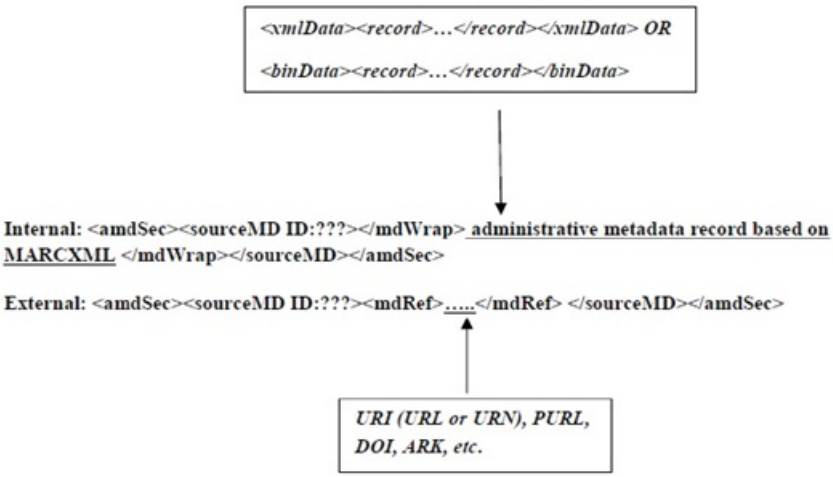
متس تعامل پیدا می کنند.

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

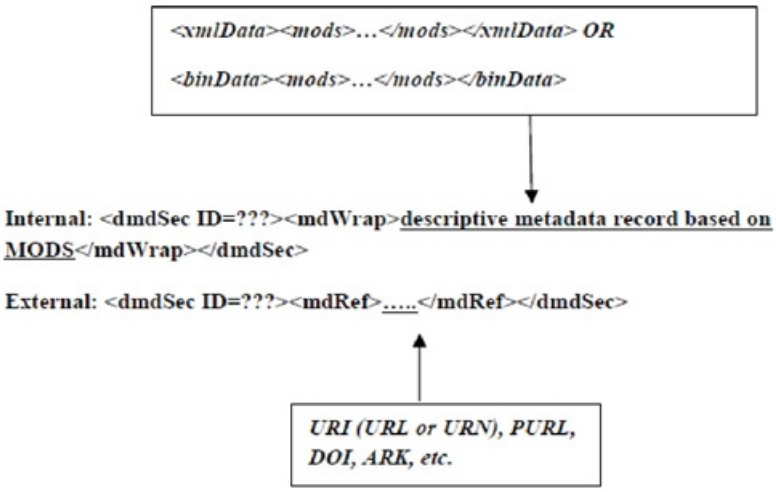
این طرح برای مجموعه عناصر کتابشناختی که با اهداف گوناگون به خصوص کاربردهای کتابخانه- ای استفاده میشوند در بستر زبان نشانه گذاری گسترش پذیر تهیه شده .است مودس امکان انتقال داده های کتابشناختی گزیده از پیشینههای موجود مارک و ایجاد پیشینههای توصیفی برای اشیای محتوایی جدید را فراهم می آورد طرح فراداده ای 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

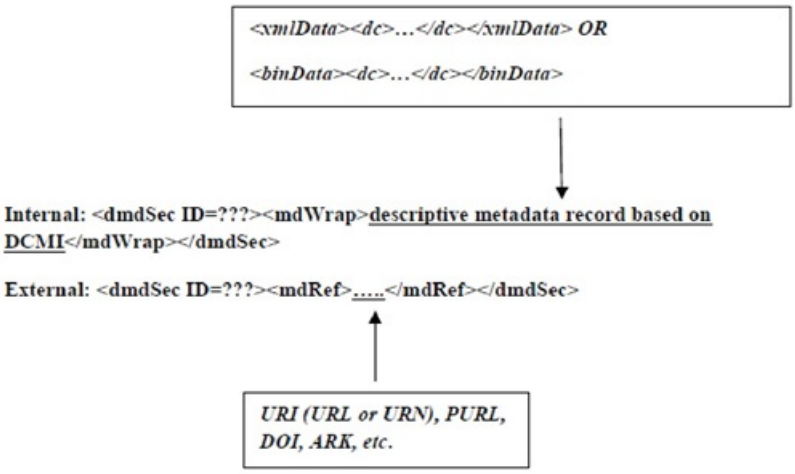
طرحی بین المللی و میان رشته ای که مجموعه عناصری ساده و کارآمد برای توصیف طیف گسترده ای از اشیای محتوایی شبکه ای ارائه

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

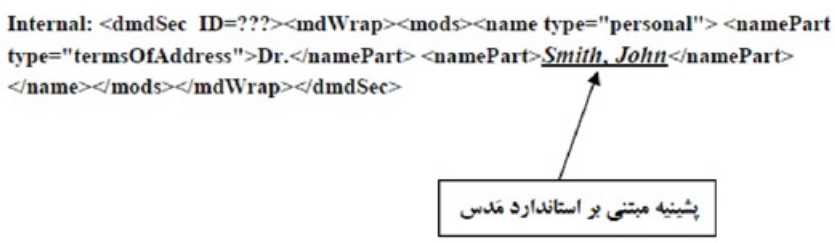
عکس

</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



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)

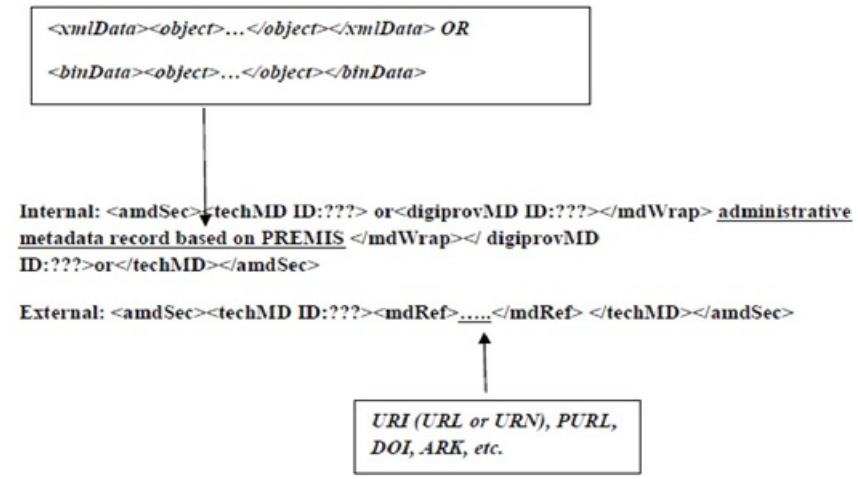
مجموعه عناصر مبتنی بر ایکس ام ال که با هدف ثبت فرادادههای مربوط به نگهداری شیء دیجیتالی در کتابخانه یا دیگر مجموعه های دیجیتالی گسترش یافته است بنابراین کارکرد اصلی استاندارد پریمیس نگهداری اشیای دیجیتالی است (هابینگ(2)، 2008)پیشینههای مبتنی بر این استاندارد بر اساس نوع

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

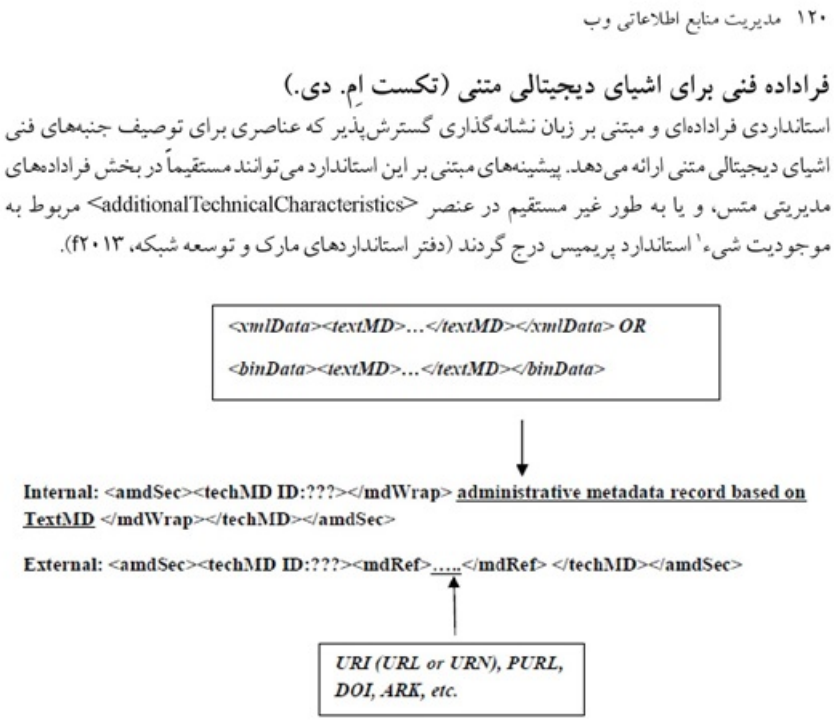
creation

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 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).

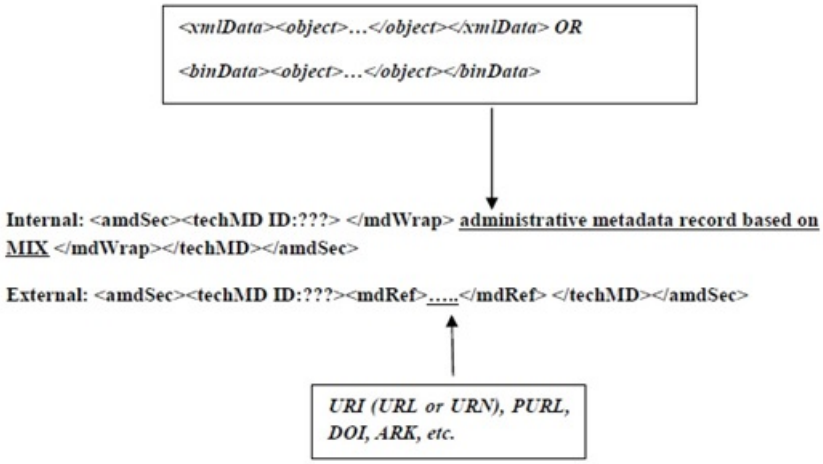
الگوی 7 شیوه درج پیشینه تکست .ام دی در بخش فرادادههای مدیریتی پیشینه متس

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.

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

همـان طور که در الگوهـای طراحی شـده مشاهـده میشود بستر نحوی اسـتانداردهای فراداده ای مورد امکان میانکنش پـذیری 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 T W:

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

---

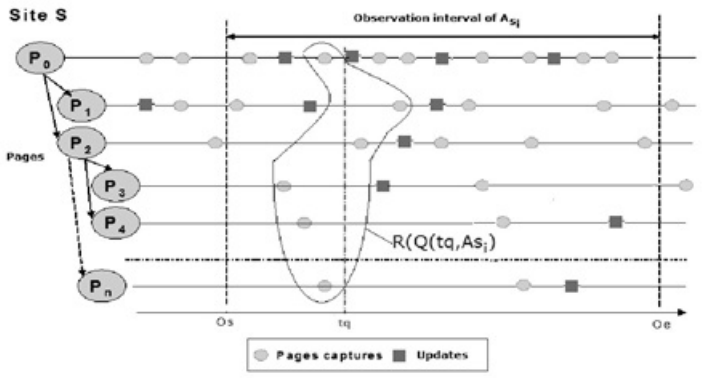
off-line -1

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 حاضر 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 صفحه دیداری شناسایی تغییر، و اهمیت بلاکهای صفحه های وب برای تشخیص بهتر تغییرات مهم میان نسخه ها رویکردهای دیگر آرشیو وب، فقط براساس فراوانی تغییرات هستند. نخستین قدم رویکرد ما ایجاد ساختار دیداری سند بر اساس بلاکهای معنایی خاصی است که الگوریتم VIPS را بسط میدهند [9] قدم دوم، شناسایی تغییرات بین آخرین نسخه صفحه آرشیو شده و نسخه ماقبل آن است الگوریتم Vi-Diff که برای این بخش از کار طراحی شده از روشهای کلی موجود برای ساختار دیداری اسناد مناسبتر .است سپس به موضوع ارزیابی اهمیت تغییرات پرداختیم برای اینکار راهبرد زمان بندی خزشگر را طراحی و با استفاده از آنالیز اهمیت تغییرات به بهینه سازی خزش وب پرداخته شد. آزمایشهای اولیه بر روی بخش بندیها و فازهای مختلف نشان دادند که زمان اجرا امیدوار کننده است. به هر حال زمان بخش بندی بسیار بالاتر از زمان مقایسه است برای بهینه سازی بهتر ،سیستم باید بر روی کاهش مدت زمان بخش بندی توجه میکردیم. 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. بدین منظور، نوعی راهبرد خزش مبتنی بر الگو را پیشنهاد کردیم. یک الگو رفتار تغییرات صفحه و اهمیت آنها را با گذشت زمان در طول روز طراحی میکند دو سنجه کیفیت جامعیت و انسجام نیز برای ارزیابی راهبرد پیشنهادی تعریف شد سودمندی راهبرد الگو محور خود را با مقایسه آن با راهبردهای مرتبط نشان دادیم در شرایط شبیه سازی شده کامل جامعیت و انسجام کلی هر راهبرد در آرشیوسازی مقایسه شد نتایج نشان دادند که راهبرد ما بهتر از راهبردهای موجود، جامعیت و انسجام 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 صفحه های وب واقعی بررسی میکنیم. ما میخواهیم که الگوهایی را از صفحه 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

.The Web archive bibliography, http://www.ifs.tuwien.ac.at/aola/links/webarchiving.html [1]

S. Abiteboul, G. Cobena, J. Masanes, and G. Sedrati. A First Experience in Archiving the [2]

French Web. In ECDL '02: Proceedings of the 6th European Con- ference on Research

.and Advanced Technology for Digital Libraries, 2002

H. Artail and K. Fawaz. A fast HTML web page change detection approach based [3]

,.on hashing and reducing the number of similarity computations. Data Knowl. Eng

.2008 ,326–337:(2)66

.M. Ben Saad and S. Gançarski. Using visual pages analysis for optimizing web archiving [4]

.In EDBT/ICDT PhD Workshops, Lausanne, Switzerland, 2010

M. Ben Saad, S. Gançarski, and Z. Pehlivan. Archiving web pages based on visual [5]

(analysis and diff. In 25è me Journé es des Bases de Donné es Avancé es (BDA

.Demonstration, Poster), Namur, Belgium, 2009)

M. Ben Saad, S. Gançarski, and Z. Pehlivan. A novel web archiving approach based on [6]

,visual pages analysis. In the 9th International Web Archiving Workshop (IWAW), Corfu

.Greece, 2009

B. Brewington and G. Cybenko. How dynamic is the web? In In World Wide Web [7]

.conference (WWW'2000), pages 257–276, 2000

D. J. C. Lampos, M. Eirinaki and M. Vazirgiannis. Archiving the greek web. In 4th [8]

.International Web Archiving Workshop (IWAW04), Bath, UK, 2004

D. Cai, S. Yu, J.-R.Wen, andW.-Y.Ma. VIPS: a Vision-based Page Segmentation [9]

.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 انتشـار اطلاعـات بحث دربارهٔ موضوعات ،سیاسـی به اشتراک گذاشـتن محتوای ویـدئویی ارسال نظرات مـدیریت وب نوشت و بیان نظرات(6) شخصـی خود دربارهٔ مباحث روز اسـتفاده می کنند فقط کاربران عادی از وب اجتماعی اسـتفاده نمی کنند، بلکه هر روز توجه رهبران سیاسـی بیش از پیش به این پدیده جلب میشود. امروزه در امریکا و انگلسـتان پاسخ به پرسشهای پارلمانی با استفاده از توئیتر امری عادی شـده است. به تازگی در 6 جولای ،2011 بـاراک اوبامـا نام خود را به عنوان اولین رئیس جمهوری که از توئیتر به عنوان ابزاری برای ارتباط جمعی اسـتفاده کرده است، ثبت نمود [13]. ،بنابراین وب اجتماعی به صورت بخشـی از مبارزات سیاسـی 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

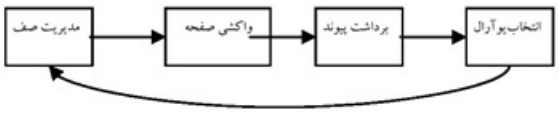
6- Comments

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

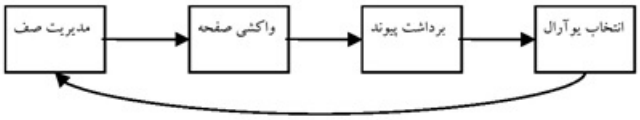
7- Music networks

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

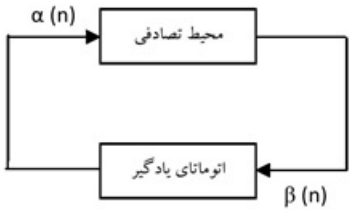
(Selamat et al.۲۰۰٤) (Anagnos et al.۲۰۰٤) ، استفاده از تئوری‌های ناهنجار و نامعلوم و سیستم های فازی برای کاهش داده های زائد (Jensen et al.۲۰۰٤) ، استفاده از مدل SVM دستگاه حفاظتی بردار (Chen et al.۲۰۰٦) (Wakaki et al.۲۰۰٦)، استفاده از وزن کلمات با الگوریتم TFIDF در صفحات وابسته به یک دیکشنری (Liang et 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 اتوماتاهـا اگر عمل در مرحله nام انتخاب شود و پاسـخ مطلوب از محیط دریافت نماید ، احتمال (n) افزایش یافته و سایر احتمالها کـاهش می یابنـد و برای پاسـخ نا مطلوب احتمال (n) کـاهش یـافته و سـایر الگوریتم زیر یـک نمونه از الگوریتمهـای 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 شـبکه ای از اتوماتاهای یادگیر است که برای حل یک مسأله خاص با یکدیگر همکاری دارند در این شبکه اتوماتاهـای یـادگیر همکـار در هر زمـان تنهـا یـک اتوماتـا فعـال است تعـداد اعمال قابل انجام توسط یک اتوماتا در DIA برابر با تعـداد اتوماتاهایی است که به این اتوماتا متصل شدهاند. انتخاب یک عمل توسط اتوماتای یادگیر در این شبکه باعث فعال شدن اتوماتای یادگیر متصـل شـده به این اتوماتـای یـادگیر متناظر با این عمل گردد . به عبارت دیگر انتخاب یک عمل توسط یک اتوماتای یادگیر در این شـبکه 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، نشـان داده میشود وجود یال (LALA در این گراف بـدین معنـاست که انتخاب عمل توسط LA باعث فعال شدن LA می گردد . تعداد اعمال قابل انتخاب توسط LA بصورت pk = {p, p. .... palpk = { pt pt. ... pf نمایش داده شود. در این مجموعه عـدد نشان دهنـده احتمال مربوط به عمل aman است انتخـاب عمل توسط LA باعث فعال شـدن LA میشـود . r تعداد اعمال قابل انجام توسـط اتوماتـای LA را نشـان میدهـد . برای کسـب اطلاعـات بیشـتر راجـع بـه اتوماتاهـای یـادگیر توزیـع شـده و کاربردهـای آن میتـوان 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:

$$Tf(T_{t,d}) = \frac{count(t,d)}{\sum_{v \in d} count(v,d)} \frac{count(t,d)}{\sum_{v \in d} count(v,d)} \quad (۳)$$

تکرار سند معکوس (idf)،نسبت اسناد d راکه tدر آنها وجود دارد درمجموعه D ارزیابی می کند.

$$idf(t) = \frac{log|D|}{|\{d \in D : t \in d\}|} idf(t) = \frac{log|D|}{|\{d \in D : t \in d\}|} \quad (٤)$$

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.

$$Tfidf(t_{t,d}) = tf(t_{t,d}).idf(t) \quad (٥)$$

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

$$sim(d_{t,d}, a) = \frac{va.va}{||vd||\ ||va||} \frac{va.va}{||vd||\ ||va||} \quad (٧)$$

یا به طور هم ارز

$$sim(d,a) \frac{\sum_{t \in d \cap a} tfidf(t,d).tfidf(t,a)}{\sqrt{\sum_{t \in a} tfidf(t,d)^2}\ \sqrt{\sum_{t \in a} tfidf(t,a)^2}} \quad (٨)$$

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

ص: 164

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

$$Corr(p,p') = \frac{\sum pp' - (\sum p \sum p')/n}{\sqrt{(\sum p^2 - (\sum p)^2/n)(\sum p'^2 - (\sum p')^2/n)}}$$

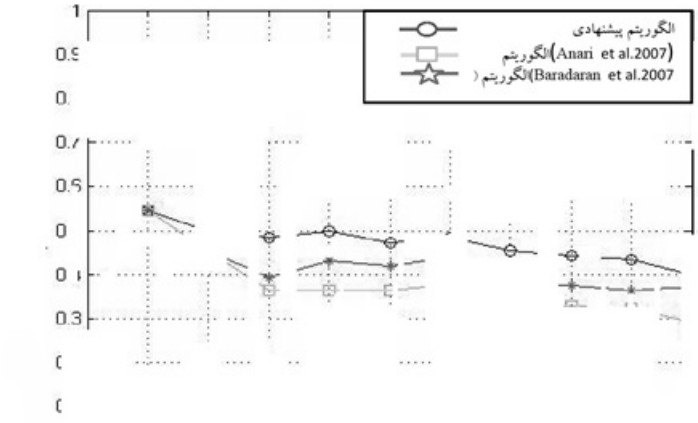
$p=\{p_{ij} \mid i,j=1,2,3,\ldots,n,\ i\neq j\}$

$$p_{ij}=\frac{d_{ij}^{-2}}{\sum_{k=1}^{n} d_{ik}^{-2}} \quad (9)$$

$$d_{ij} = \sqrt{\sum_{k=1}^{m}\left(cw_i^k - cw_j^k\right)}\sqrt{\sum_{k=1}^{m}\left(cw_i^k - cw_j^k\right)}$$

$p' = \{p'_{ij} \mid i,j=1,2,3,\ldots,n,\ i\neq j\}$

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

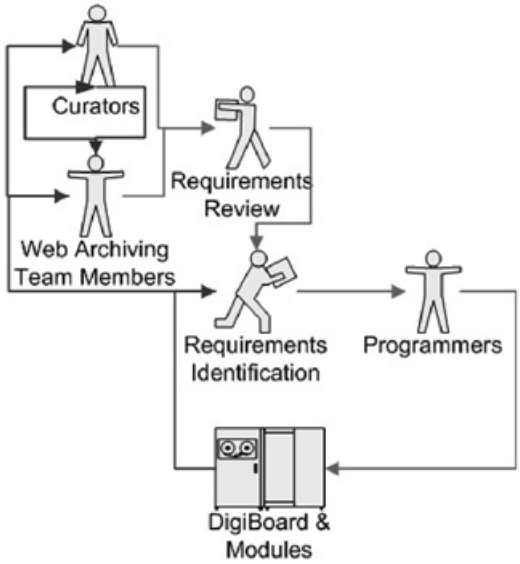
توسعه DigiBoard به صورت چابک(1) و فرآیند توسعه تکراری(2) (شکل 1) میباشد که نیازمندیها و مسائل شناسایی شده توسط کتابداران و اعضای تیم را پوشش میدهد نیازمندیهای هر مرحله از فرآیند تیم آرشیو وب به صورت ترتیبی و افزایشی به سیستم اضافه شده است؛ بنابراین کارآیی مورد اضافه شده توسعه سیستم را تحت تأثیر قرار 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 مشاهده 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

---

frequency -1

post-crawl -2

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

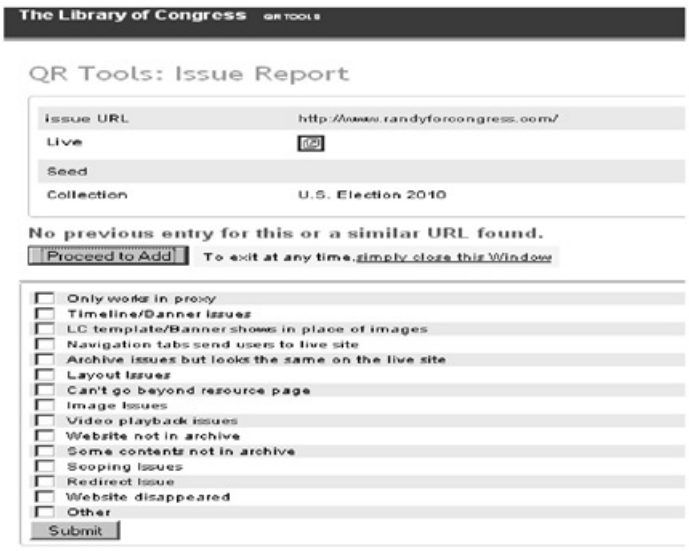
2- issue

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

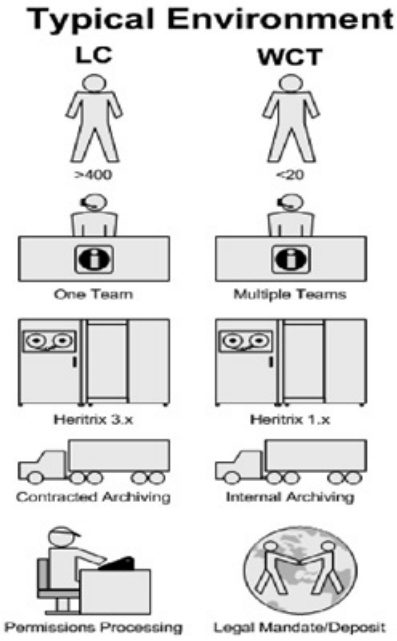
کتابخانه متقاعد شد که WCT به علت برخی تفاوتهای عملیاتی نیازهای آرشیو وب کتابخانه کنگره را مرتفع نخواهد کرد (شکل 8)

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

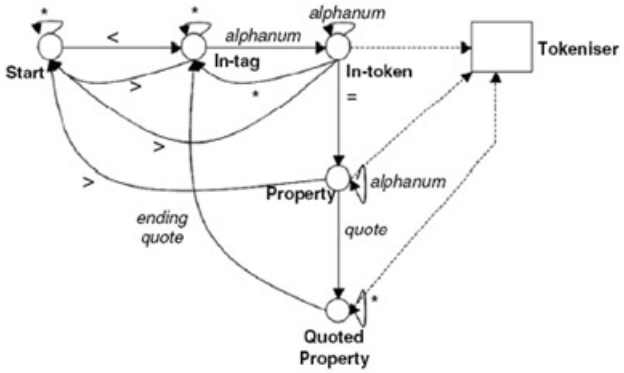
HTML(1) را در نظر بگیریم که یک فایل متنی 8 بیتی(2) متشکل از متن ساده (بدون رمز)(3) و تگهای نشانه گذاری(4) است- و اطلاعات همراه آن مانند URL اولیه(5) هدف تجزیه کننده HTML است تا با پویش(6) صفحه پیوندها را گردآوری و تجزیه و تحلیل کند و آنها را به خزشگر بسپارد. ساختار HTML برای گردآوری پیوندها مناسب نیست 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 4.4Hierarchical URL aud Relative Forms[2396]

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