وزارت علوم، تحقیقات و فناوری
دانشگاه تحصیلات تکمیلی علوم پایه
گاوازنگ - زنجان

استخراج خودکار اطلاعات از تالارهای گفتمان

کارشناسی ارشد

سعید سارنچه

استاد راهنما: دکتر بهرام صادقی بی غم

اساتید مشاور: دکتر ویدی آساگار پوتدار

تیرماه ۱۳۹۰

چکیده

با همه‌گیرشدن استفاده از اینترنت و افزایش توان تولید اطلاعات و نرخ بارگزاری آن به شبکه جهانی اطلاعات، دسترسی به اطلاعات در این شبکه با مشکلاتی مواجه کرده است که نیاز به استفاده از ابزارهای کمکی جهت دسترسی سریع به اطلاعات مفید می‌باشد. امروزه موتورهای جستجو از علم داده کاوی برای ارائه سرویس‌های مناسب و مفید برای کاربران خود استفاده می‌نمایند. از آنجایی که اطلاعات وارد شده اطلاعاتی هستند که به نوعی در زندگی روزمره کاربران تولید می‌شوند. اطلاعاتی در مورد سلیقه‌ها، تفریحات، عقاید و سوالات کاربران که توسط ابزارهای دیجتال کاربر بارگزاری می‌گردند، پس با پردازش این اطلاعات می‌توان از داده‌های خام، اطلاعاتی را استخراج کرد که در نگاه اول دور از دسترس می باشند.

امروزه بیشتر کاربران از نرم افزارهای جدید به عنوان web 2.0 برای ارائه اطلاعات بر روی وب استفاده می‌کنند. این نوع نرم افزارها برای ایجاد یک محیط برای اشتراک گذاری اطلاعات ایجاد شده اند و همچنین امکان مدیریت اطلاعات را به کاربران می‌دهند. یکی از این نرم افزارها که مورد استقبال کاربران قرار گرفته است نرم افزا تالار گفتمان (Forum) می‌باشد. این نرم افزار تلاش می‌کند تا محیطی برای بحث و گفتگوی کاربران ایجاد نماید و کاربران نیازهای خود، سوالات، نظرات خود را در مورد بحث‌های مختلف را ثبت نمایند. امروزه افراد، شرکت‌ها، سازمانهای دولتی و موسسات آموزشی از این سیستم به عنوان کانال ارتباطی بین خودشان و کاربران استفاده می‌نمایند. نوکیا، سیستم عامل Ubuntu و دانشگاه IASBS از تالار گفتمان برای ایجاد ارتباط مفید و دو سویه از این نرم افزار استفاده می‌کنند. تالار گفتمان مربوط به Ubuntu دارای نزدیک به یک میلیون کاربر می‌باشد که این کاربران توانسته‌اند نزدیک به ۹.۵ میلیون مطلب را در تالار پست نمایند. که به مطالب، سوالات و پاسخ‌های متفاوتی در باره این سیستم عامل اشاره شده است بطوریکه روزانه تعداد زیادی از کاربران با مراجعه به این سایت مشکل خودشان را حل می‌کنند. این شهرت و استقبال کاربران مشکلاتی را نیز به همراه داشته است. از جمله این مشکلات افزایش کاربرانی است که داده‌های هجو در این تالارها وارد می‌کنند. امروزه Spammer ها با استفاده از ابزارهای مربوط به یادگیری ماشین توانسته اند تا از سدهای مختلف عبور کنند و این مطالب را وارد تالارهای گفتمان کنند. در عین حال تشخیص این نوع کاربران از کاربران دیگر مشکل می‌باشد. تنها راه حل بررسی رفتاری این نوع از کاربران می‌باشد که این هم نیاز به پردازش اطلاعات پست شده در تالارگفتمان را دارد که این کار را میتوان از طریق استخراج اطلاعات انجام داد.

الگوریتم‌های مختلفی برای استخراج اطلاعات مطرح گردیده است. الگوریتم‌های اتوماتیک، دستی و یا نیمه اتوماتیک. هر کدام در تلاش اند که بیشترین خروجی را دارند. ولی امروزه بیشتر اطلاعات د نرم افزارهای زیر مجموعه web 2.0 ثبت، ذخیره و نمایش داده می‌شود لذا در این پروژه هدف ایجاد نرم افزار می‌باشد که بتواند با استفاده از ویژگیهای این گونه نرم افزارهای تحت وب بهینه ترین عملکرد را

داشته باشد. در این پروژه با توجه به اهمیت تالارگفتمان ، نرم افزار برای استخراج اطلاعات از این نوع نرم افزار طراحی گردیده است. که در فازها بعدی بدنبال گسترش آن برای نرم افزارهای مشابه نیز هستیم.

در این پروژه ۲ نوع الگوریتم یا برنامه پیاده سازی و تست شد. در اولین مرحله یک برنامه نیمه اتوماتیک برای استخراج اطلاعات پیاده سازی شد که در آن در فاز یادگیری نرم افزار الگوهای مربط به مکان اطلاعات به نرم افزار داده می‌شد و این اطلاعات و الگوها در سیستم ذخیره شده تا در صورتی که اگر نوع فایل ورودی با یکی از فایل‌های بررسی شده یکسان باشند از این الگو برای استخراج اطلاعات استفاده می‌شود.

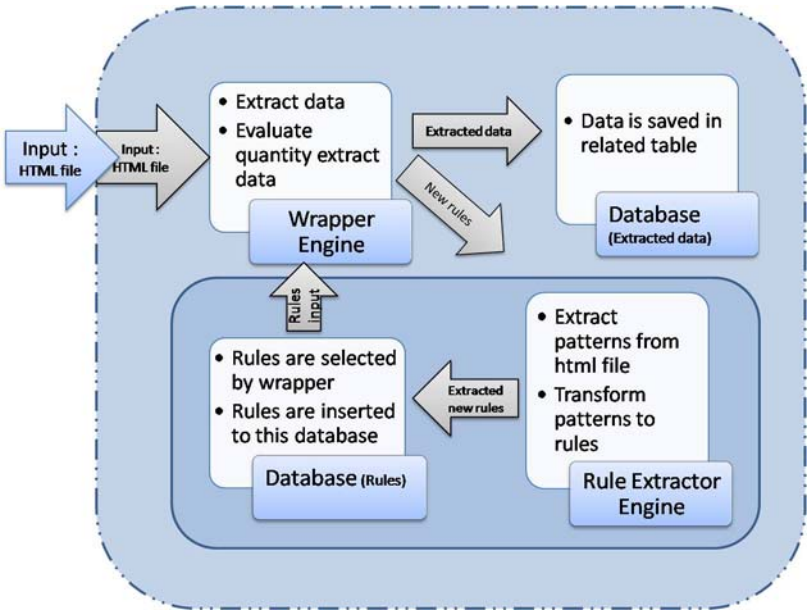
در الگوریتم اتوماتیک با اضافه کردن هسته یادگیری اتوامتیک به این الگوریتم توانسته ایم که عملکرد آن را بهینه نموده و باعث افزایش کارآیی، پایداری و همچنین کاهش زمان اجرار شده است. این هسته یادگیری، زمانی که الگوی مناسبی برای فایل ورودی نداشته فعال شده و الگوها اطلاعاتی فایل را استخراج می‌نماید و در پایگاه داده برای این نوع از فایل‌ها ثبت می‌نماید و سپس هسته استخراج را بر اساس الگوی جدید فراخوانی می‌نماید.

در این پروژه هدف بهینه سازی الگوریتم‌های استخراج جهت استخراج اطلاعات از نرم افزراهای تحت وب با عنوان web 2.0 تعریف گردیده است. الگوریتم باید توانایی استخراج اطلاعات از وب سایت‌ها با قالب‌های مختلف را داشته باشد و همچنین اطلاعات استخراج شده باید از کیفیت خوبی برخوردار باشد. در اینجا منظور از کیفیت در استخراج اطلاعات، استخراج اطلاعات به همراه تمام ویژگی‌هاییش می‌باشد. بطور مثال اگر بخواهیم از یک تالار گفتمان اطلاعات استخراج کنیم این اطلاعات باید شامل پیام‌های ثبت شده، نویسنده پیام، تاریخ پیام و کلیه اطلاعات مربوط به آن پیام باشد.

به طور کلی این الگوریتم باید کارکردی به شرح زیر داشته باشد:

«الگوریتم فایل ورودی را از ورودی گرفته، سپس به بررسی جهت پیدا کردن نشانه ای خاص که به عنوان امضاء در فایل ها ثبت می‌شود می پردازد. الگوریتم با یافتن این امضاء الگوی مخصوص این فایل را بارگذاری کرده و اطلاعات استخراج شده را به پایگاه داده ارسال می کند». این ایده اولیه برای طراحی چنین الگوریتمی‌می‌باشد. که این ایده در مرحله تکمل پیدا کرده و به الگوریتم بهینه تری رسیده ایم. در این بخش، جزئیات الگوریتم، عملکرد الگوریتم، ورودی و خروجی الگوریتم و در آخر نتیجه بدست آمده از اجرا و پیاده سازی الگوریتم مورد بررسی قرار گفته است.

این الگوریتم از چند هسته تشکیل شده است که در این قسمت به توصیف انها می‌پردازیم. همچنین زیر بخش‌های الگوریتم و ورودی و خروجی بخشها مختلف در شکل ۱.۱ به صورت شماتیک نمایش داده شده است. همانطور که در شکل ۱.۱ مشخص است، بعضی از بخش‌های الگوریتم در ارای پس زمینه پر رنگتری نسبت به بقیه دارند. این بخش‌ها در مرحله دوم به سیستم اضافه شده اند.



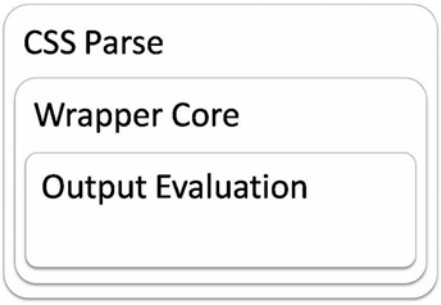
شکل ۱.۱ . این تصویر الگوریتم بهینه استخراج را به صورت شماتیک نمایش می‌دهد.

اولین بخش Wrapper Engine می‌باشد که کار استخراج اطلاعات از فایل ورودی را برعهده دارد. همانطور در فصل ۳ به معرفی این چنین الگوریتم‌های پرداختیم، این نوع الگوریتم‌ها به ۲ دسته supervised و unsupervised تقسیم می‌شوند. استراتژی‌های مختلفی برای این الگوریتم تعریف شده است تا کارآیی‌های بهتری داشته باشند.الگوریتمی‌که در این قسمت تعریف شده است عبارت است از «شناسایی فایل ورودی و بارگذاری الگوی مناسب از پایگاه داده و تغییر شکل داده از یک فایل HTML که از رشته‌های شامل تگ‌ها و اطلاعات به یک ساختار درخت مانند به نام DOM که دسترسی به هر قسمت از اطلاعات موجد در صفحه را آسان نموده و همچنین قابلیت پیدا کردن اطلاعات بر اساس الگو را فرآهم می‌آورد. پس ایجاد ساختار جدید برای نمایش فایل اطلاعات اضافی و غیر ضروری بر اساس تعریف‌های موجود در الگو حذف شده و بقیه اطلاعات در پایگاه داده ذخیره می‌شوند».

این بخش نیز از ۳ زیر بخش زیر تشکیل شده است:

- **CSS Parser**
- **Wrapper Core**
- **Evaluation Module**

هر یک از این زیربخش‌ها به طور کامل در زیر مورد بحث قرار گرفته اند و ارتباط بین این سه بخش در شکل ۲.۱ نشان داده شده است.

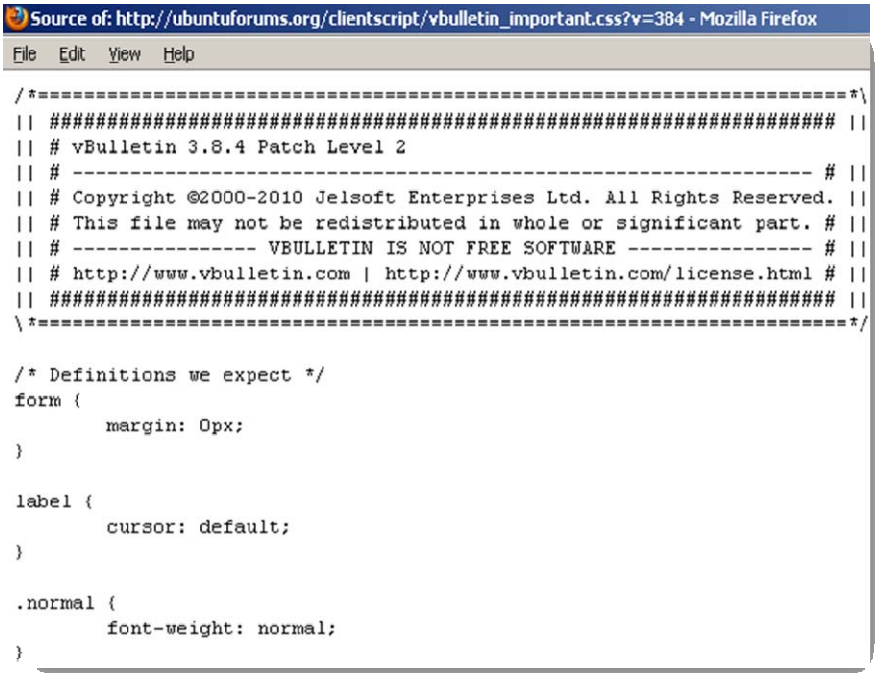


شکل ۳.۱. بخشی از فایل CSS از وب سایت ubuntuforums.org .

در ابتدا فایل تعریفی از فایل CSS را مطرح می‌کنیم. فایل CSS که مخفف عبارت Cascading Style Sheets می‌باشد برای مدیریت نحوه نمایش اطلاعات بر روی صفحه استفاده می‌شود. در این فایل قالب نمایش اطلاعات برای جاهای مختلف صفحه تعریف می‌شود و در فایلHTML فقط نام قالب تعریف شده فرآخوانی می‌گردد.

از آنجایی که امروزه نرم افزارهای تحت وب توسط تیم‌های مختلفی نوشته می‌شود معمولا مشخصات نرم افزار و نسخه آن را نیز به عنوان امضاء د فایل ذکر می‌کنند. که این اتفاق در هنگام تهیه فایل CSS نیز می‌افتد. نرم افزارهای تالارگفتمان نیز از این امر مستثنی نیستن و می‌توان نوع نرم افزار و نسخه آن را در فایل CSS و یا در پایین صفحه در قسمت امضاء یا حق کپی برداری یافت. به طور مثال، تیم پشتیبانی سیتم عامل Ubuntu از نزم افزار تالار گفتمان جهت ارائه سرویس برای کاربران خود استفاده می‌کند. بخشی از متن فایل CSS آن در شکل ۳.۱ نشان داده شده است.

در این بخش تلاش می‌شود تا این قسمت از اطلاعات از داخل فایل CSS یا از قسمت حق کپی رایت استخراج کرده و الگوی ثبت شده برای این قسمت را بارگزاری کرده و فایل و الگو را برای قسمت wrapper ارسال می‌نماید.



شکل ۳.۱. بخشی از فایل CSS از وب سایت ubuntuforums.org .

دومین قسمت از wrapper Engine هسته مرکزی آن می‌باشد که کار اصلی استخراج اطلاعات را بر عهده گرفته است. این قسمت با استفاده از الگوی که از بخش قبلی دریافت می‌کند و همچنین ساختار درخت گونه DOM مربوط به فایل،عناصر اطلاعاتی تعریف شده در الگو را استخراج کرده و بعداز انجام پردازش و نتیجه گیری اطلاعات در پایگاه داده ذخیره می‌شوند.

در این بخش، ابتدا ساختار درختی DOM از فایل ایجاد می‌شود. چرا اه این ساختار اجازه درسترسی ساده به هر قسمت از فایل را ایجاد می‌کند و برنامه می‌تواند به راحتی و با کمک الگو اطلاعات ناخواسته را حذف نماید و اطلاعات مفید و مورد نیاز را استخراج نماید. پس الگوها آدرس اطلاعات مفید را بر اساس نودهای پدر و پدرشان نشان می‌دهد. با یک الگو ممکن است چند زیردرخت

استخراج شوند و این به خاطر این است که این‌ها همه از یک نوع اطلاعات هستند. به طور مثال اگـر در یک صفحه تالار گفتمان ۱۰ عدد پست وجود داشته باشد پس با هر الگو مربوط به یک عنصر اطلاعـات ۱۰ داده باید استخراج شود یا به عبارت دیگر برای هر عنصر از اطلاعات ۱۰ خروجی باد ثبت گردد..

از آنجایی که یکی از اهداف ما از طراحی و پیاده سازی این پروژه استخراج اطلاعات از داخـل تالارهـا می‌باشد. لذا پس از استخراج اطلاعات نیاز به پردازش اولیه اطلاعات می‌باشـد. چـرا کـه در یـک تـالار پستی که اول انجام می‌شود به عنوان موضوع و بقیه پست‌ها جواب‌های آن ثبت می‌شود. بـه نـوعی ایـن ساختار مرتبط باید حفظ شود. پس بعداز استخراج اطلاعات،پردازش اطلاعـات بـر روی آنهـا صـورت گرفته و به صورت رابطه ای در پایگاه داده ذخیره می‌شوند.

در این قسمت، که آخرین قسمت از استخراج اطلاعات می‌شود نتیجه کار ارزیابی شده تا از عملکرد استخراج که شامل انتخاب الگو و استخراج هست اطمینان حاصل شود. ارزیابی شامل دو قسمت می‌باشد. یک بخش به ارزیابی کیفیت اطلاعات استخراج شده می‌پردازد و در قسمت دیگر بحث کیمیت عناصر اطلاعتی که باید استخراج می‌شد پرداخته می‌شود.

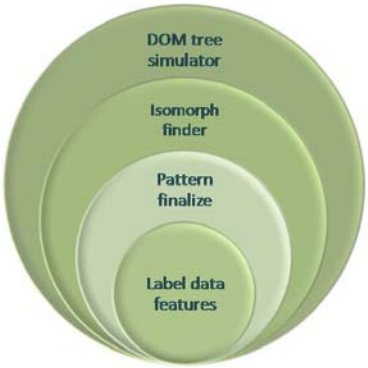
در ارزیابی کیفیت استخراج، خروجی نرم افزار توسط یک سیستم خوشه بندی که برای این نرمافزار ساخته شده دسته بندی شده و تفاوت خروجی سیستم خوشه بندی مبنی بر اینکه این خروجی در کدام دسته از اطلاعات تعریف شده برای استخراج دسته بندی شده نسبت به مولفه ای اطلاعلتی کهای الگو برای آن تعریف شده انحراف از جواب را نمایش می‌دهد. به طور مثال در تالار گفتمان موجودیت اطلاعات عبارتند از پست و اطلاعات وابسته به آن. فرض کنید الگوریتم کلمه ۵ حرفی را به عنوان نام کاربری استخرا ج کرده است. این خروجی در سیستم خوشه بندی دسته بندی می‌شود. اگر خروجی سیستم خوشه بندی نیز همان نام کاربری بشد پس این الگوتوانایی استخراج داده را دارا می‌باشد.

در ارزیابی کمیتی، تعداد مولفه‌های استخراج شده با تعداد مولفه‌های که باید استخراج شوند باید یکسان باشند و این دهل بر این است که الگو منطبق بر فایل می‌باشد و برنامه می‌تواند به کار خود برای استخراج ادامه دهد.

قابل ذکر است که این قسمت در فاز گسترش به این الگوریتم اضافه شده است و در هر شرایطی که جواب مناسبی دریافت نکند بخش دوم این برنامه که کار استخراج الگو می‌باشد را فرآهم می‌نماید.

این قسمت برای افزایش پایداری نرم افزار استخراج اطلاعات طراحی و افزوده شده است. این قسمت در زمان‌های که الگوی مناسبی برای استخراج وجود ندارد شروع به کار می‌کند و الگوی مربوط به این

نوع فایل‌ها را وارد پایگاه داده می‌نماید. این بخش شامل ٤ زیر بخش می‌باشد که در شکل٤.١ نمایش داده شده است. هر بخش به صوت کامل در زیربخش‌های زیر توضیح داده شده است.



شکل ٤.١ . زیر بخش‌های موتور استخراج الگو و ارتباط بین بخشها.

اولین بخش از این قسمت DOM tree Simulatore می‌باشد. در این قسمت شاختار درخت مانند فایل HTML ساخته شده و برای استفاده در بخش‌های دیگر در پایگاه داده ذخیره می‌شود. در این ساختار درختی هر تگ HTML به عنوان یک نود محسوب شده و در پایگاه داده ذخیره می‌شوند و همچنین اطلاعاتی چون شناسه پدر، سطح تگ در درخت، تعداد فرزندان و کلاس مربوط قالب نمایشی را ذخیره می‌گردد. تگ‌های HTML را می‌توان به ٢ دسته تقسیم کرد، تگ‌های قالب و تگ‌های ساختاری. تگ‌های قالبی، تگ‌های هستند که برای تنظیمات گرافیکی و نمایشی اطلاعات استفاده می‌شوند. تگ‌هایی مانند <B> ، <P> ، <I> و بسیاری تگ‌های دیگراز این دسته اند. تگ‌های ساختاری که به مدیریت ساختار اطلاعات در صفحه می‌پردازد. تگ‌هایی چون <Table>، <Tr>، وتگ‌های مشابه دیگر جزء این دسته محسوب می‌گردند. برای سادگی کار و کاهش پیچیدگی برنامه تگ‌های قالبی بعد از وارد شدن به پایگاه داده حذف شده و فرزندانشان جای آن‌ها وارد می‌شوند. این کار بسیار مفید بوده به طوریکه در یک از این فایل‌ها با حذف این نوع تگ‌ها سطح درخت نهایی از ١٧ به ٨ کاهش پیدا کرد و این موضوع باعث کاهش زمان اجرا می‌شود.

پس از ذخیره سازی ساختار درختی در پایگاه داده، فرآیند پیدا کردن زیر درخت‌های مشابه آغاز می‌شود. از آنجایی که برنامه نویسان وب از یک قالب برای نمایش پیام‌ها نشان می‌دهند. پس قالب استفاده شده برای نمایش پیام میتواند تکرار شده باشد در صورتی که ما این قالب‌های تکراری را پیدا کنیم می‌توانیم الگوی استخراج اطلاعات را تدوین کرده و مورد استفاده قرار دهیم. برای استخراج

اطلاعات الگو می‌توان از الگوریتم پیدا کردن زیر درخت‌های مشابه استفاده نمود که برا ی اینکار طراحی شده است.

در این الگوریتم ابتدا نودهای درخت در هر سطح بر اساس نام و ویژگی‌هایشان که در پایگاه داده ثبت شده اند گروه بندی می‌شوند. به طور مثال در سطح ۵ درخت تمام نودهای با نام <td> که دارای کلاس «post-detail» در یک گروه جای می‌گیرند.

پس از این دسته بندی نوبت به دست بندی دیگر بر روی داده‌های دسته بندی شده می‌رسد. در این دسته بندی ۲ نود در یک گروه قرار می‌گیرند در صورتی که خودشلن در یک گروه باشند و تعداد فرزندانشان و گروه فرزندانشان یکسان باشند. در سطح آخر تغییری دسته بندی قبلی رخ نمی‌دهد ولی برای سطوح بالاتر ممکن است دسته‌ها فرق کنند. چرا که فرزندانشان در سطوح پایینتر ممکن است در گروهای متفاوتی قرار گیرند.

پس از انجام این دسته بندی از سطوح پایین به سطوح بالاتر نودهای موجود در یک گروه ریشه زیر درخت‌هایی هستند که شبیه ام می‌باشند. از ویژگی‌های این الگوریتم می‌توان قابلیت یافتن زیر ریشه‌هایی که فقط در بازه خاصی از سطوح مشابه اند را نام برد.پس از اتمام الگورتیم با پردازشی برروی گروهای ایجاد شده در دراخت می‌توان زیر درخت‌های مشابه را استخراج کرد. البته برای حذف نویز از داخل جواب‌ها محدودیت‌های چون تعداد تکرار این الگو، عمق این زیر درخت و تعداد نودهای موجوداستفاده می‌کردد.پس بعداز اجرای این الگوریتم الگوهای استخراج اطلاعات بدست می‌آیند.

در آخرین مرحله، باید عناصر اطلاعات نیز در این زیر دخت مشخص گردد.به طور مثال کدامین برگ مربوط به نام کاربری می‌شود و کامیک متن پست شده ا در بر میگیرد. برای این کار از یک سیستم خوشه بندی استفاده می‌شود. تا برگهای زیر درخت را بررسی و هر کدام را به عنوان یکی از ویژگی‌های اطلاعاتی تشخیص داده به صورت دقیق الگوی استخراج اطلاعات را تهیه و در پایگاه داده وارد نماید.

بری ارزیابی عملکرد الگوریتم، ۷۲ عدد سایت که از تالار گفتمان استفاده کرده بودند انتخاب شده و از هر کدام فقط یک صفحه به عنوان تست ذخیره گردید تا بتوان پایداری الگوریتم برای فایل‌ها با قالب‌های متفاوت را تست کرد. هر کدام از این تالارهای گفتمان شامل صفحه‌های متفاوت بودند به طوریکه هر کدام از آنها شامل ۱۰ها هزار صفحه و پست می‌باشند که اگر برنامه بتواند اطلاعات فقط یکی از این صفحه را استخراج نماید می‌تواند بقیه فایل‌ها را با موفقیت تجزیه و تحلیل نماید.

در جدول ۱.۱ مشخصات مربوط به فایل که شامل نوع نرم افزار، نسخه آن و همچنین ویرایش نرم افزار مشخص شده است.

جدول ۱.۱ . لیست مشخصات فایل‌ها که در این پیاده سازی و تست مورد استفاده قرار گرفته اند.

Forum Name	Forum Software	Version	Template	Number of files
The Astro Post	SMF	1.1.10	Standard	10
SoloIngegneria FORUM	SMF	1.1.2	Standard	8
www.pietraligure.net	SMF	1.1.4	Standard	2
Parentguideclub	SMF	1.1.6	Standard	1
MERZ ALUMNI EINGETRAGENER VEREIN	SMF	1.1.7	Standard	2
STRABILIARDO	SMF	1.1.8	Standard	3
b a b o n g a	SMF	2.0	Standard	4
other	SMF	?	?	3
IREM Houston Member Forum	phpBB	2	subSilver	7
Entertainment comes in blueberry	phpBB	3	Prosilver	17
nukeCops Forums	phpBB	2	Fisubsilver	2
Family / Supporter Support and Resources Forum	phpBB	3	subsilver2	4
other	phpBB	?	?	2

ما در این فایل بدنبال استخراج اطلاعاتی با عناصر یا ویزگی‌های لیست شده در جدول ۲.۱ می‌باشیم. همانطور که در این جدول مشخص شده است ما بدنبال استخراج اطلاعات مربوط به پست‌های صورت گرفته در تالار می‌باشیم.

جدول ۲.۱ . لیست عناصر اطلاعاتی که باید استخراج گردد

List of Information Extracted	SMF	phpBB
Author	✓	✓
Number of author post	✓	✓
Author category	✓	✓
Post Content	✓	✓

الگوریتم خود را در دو حالت supervised و unsupervised مورد تست و آزمایش قرار داده ایم و نتایج زیر حاصل گردیده است. که در جدول ۳.۱ لیست شده است.

جدول ۳.۱. خروجی الگوریتم در دو حالت supervised و unsupervised

Wrapper Engine Ver.	Number of File	Correct Extraction	Errors
Semi-automatic	72	57	**15**
Automatic	72	64	**6**

زمانی که الگوریتم به صورت supervised مورد تست و آزمایش قرار می‌گرفت الگوریتم فاقد بخش استخراج الگو می‌باشد و همچنین هیچ گونه ارزیابی پس از استخراج بر روی اطلاعات ان صورت نمی‌گیرد. ولی زمانی که به صورت unsupervised مورد تستت قرار می‌گرفت شامل موتور استخراج الگو و همچنین بخش ارزیابی خروجی می‌باشد.

همانطور که از نتایج مشخص است الگوریتم unsupervised کارکرد بهتری نسبت به قبلی دارد و بیشتر خطاها نیز بدلیل وجود مشکلات فنی در فایل HTML میباشد. به طور مثال، تگی در این فایل باز بوده که باعث می‌شود موتورهای استخراج نتوانند شماتیک مناسبی از فایل را طراحی نمایند و این باعث می‌شود تا الگویی قابل استخراج نباشد.

بیشتر اشکالات در الگوریتم supervised بدلیل عدم تطابق الگو با فایل ورودی می‌باشد به عبارت دیگر در بخش یادگیری نمونه از این فایل وجود نداشته پس الگوی نیز برای این فایل در پایگاه داده موجود نیست پس استخراج اطلاعات از این فایل غیر ممکن می‌گردد.

Abstract

Nowadays, users use web 2.0 software to manage data in the Internet network. Users can upload digital data which is captured by digital devices. Lots of knowledge are existed in sets of data that saved in web pages. Information Extraction(IE) methods are used to extract these knowledge from online data sources.

In this study, we developed two new Information extraction algorithms that have ability to extract information form web 2.0 software. The aim of these algorithms is to extract data from web 2.0 software more accurately than other algorithms. In these algorithms, a tree representation is used to explore data. In this structure each part of web page and data can be accessed by extractor engine. In the second phase of thesis, a rule extractor engine is added to analyze the tree representation to extract pattern of data in the web pages with unfamiliar format. Also a decision unit is designed in wrapper extraction engine to evaluate the output of engine. This unit decides that the rule was proper for file extraction or not, then the rule extractor engine should be called or not.

These algorithms were tested by a set of 72 files from different versions of forums softwares. In this stage, the semi-automatic and automatic versions of our algorithms were tested and the results are shown that the automatic version produced better performance than semi-automatic. It could extract with accuracy of more than 90% of the input files. In the automatic algorithm, new engine is used as rule extractor to detect data patterns and extract patterns.

The semi-automatic algorithms is done by one of the Phd student in Curtin university for gathering necessary information from forums and also it will use by

Antoine Blanchard who currently works on social scientists who would like to analyze a forum where autism patients discuss, powered by MyBB.

Acknowledgements

I would like to express my sincere gratitude to Dr. Bahram Sadeghi Bigham for being an out- standing advisor, excellent professor and a great friend. His constant encouragement, support, and invaluable suggestions made this work successful. I am also grateful to my committee members Bahman Ghandchi for their time and effort in reviewing this work.

My sincere thanks go to Dr. Vidyasagar Potadar for giving me the opportunity to get into Data Mining Technologies and for his encouragement. I would also like to acknowledge my best friends and classmates for their help and valuable advice.

I am deeply and forever indebted to my parents, my brother Amir Houssein and my sister Maryam for their love, support, patience and encouragement throughout my entire life.

Table of Contents

List of Tables

List of Figures

Abbreviations

WWW	World Wide Web
DBMS	Database Management System
ML	Machine Learning
IR	Information Retrieve
NLP	Natural Language Processing
WI	Wrapper Induction
HITS	Hypertext-Induced Topic Selection
IE	Information Extraction
URL	Universal Resource Locater
DOM	Document Object Model
CSS	Cascading Style Sheet
KE	Knowledge Engineering

Chapter 1

Web Mining

Abstract. By increasing input rate of data on Internet, finding relevant and useful information is going to be a difficult and time consuming task. New tools or methods are needed to filtering unwanted data and separated the useful from unwanted data. So, Web mining is a concept define to cluster or classify data in World Wide Web(WWW) accurately which cause to update hierarchy search engine and data repository. Web mining consist of three categories, *Web Content Mining*, *Web Structure Mining* and *Web Usage Mining*. In this chapter, Web mining will be described and related algorithms of mining in each category will be discussed.

1.1 Introduction

As technology improvement and new devices invention cause to growth amount of digital data sharply against other types of data. These days, people use electronic device to save a wide range of data like their shopping list, the path of shopping center, daily report of their job, images of memorable times and more and more data that saved by digital devices as digital data to share in personal web, social networks or forums. So, numbers of web pages have exponentially growth obviously. The other important feature is the content, theme and also style of web pages dynamically have been changed. Since, the growth of data on Internet and also its characteristics, some problems are be inevitable. The first one is "finding relevant information"; The web is now known as a big resource of information that users can search on it to find what they want. Users usually use search engines to find specific data like