چارچوب سیستم حمایت از پژوهش برای وب کاوی

حمیرا عزت پناه*[۱،۰]

۱- گروه مهندسی کامپیوتر، دانشگاه گیلان، رشت، ایران،

چکیده

امروزه طراحی و پیاده سازی یک سیستم پشتیبانی تحقیق و پژوهش، به چالشی برای محققان مایل به استفاده از اطلاعات مفید، تبدیل شده است. این مقاله چارچوبی برای پشتیبانی سیستم های وب کاوی پیشنهاد می کند. این سیستم ها به منظور شناسایی، استخراج، فیلترسازی و تجزیه و تحلیل داده ها از منابع وب طراحی شده اند. همچنین این سیستم ها، بازیابی وب و تکنیک های داده کاوی را با هم جهت ارائه زیر ساختی کارآمد به منظور پشتیبانی اطلاعات وب کاوی برای تحقیق ترکیب می کنند.

کلمات کلیدی: داده کاوی، بازیابی وب، خوشه، طبقه بندی، قوانین انجمن، سیستم پشتیبانی پژوهش.

۱- مقدمه

تکامل شبکه جهانی وب برای ما رشد مقدار زیادی از داده ها و اطلاعات را همراه داده های فراوان ارائه شده توسط وب، به ارمغان آورده است. اینجاست که شبکه جهانی وب به یک منبع مهم برای تحقیق تبدیل می شود. با این حال، اطلاعات مرسوم و تکنیک های استخراج داده ها، به دلیل نیمه ساختار یافته بودن آن و یا حتی به دلیل طبیعت بدون ساختارشان نمی توانند به طور مستقیم به وب اعمال شوند. صفحات وب اسناد ابر متن است که شامل هر دوی متن و لینک به دیگر اسناد می باشد. علاوه بر این، اطلاعات وب ناهمگن و پویا هستند. بنابراین طراحی و پیاده سازی سیستم پشتیبانی پژوهش وب کاوی به یک چالش برای محققان به منظور استفاده از اطلاعات مفید وب، تبدیل شده است. معمولا وب کاوی، به عنوان محتوای وب کاوی وکاربرد وب کاوی طبقه بندی می شود. استخراج محتوای وب، مطالعه محتوای وب کاوی و بازیابی از اطلاعات وب است، در حالی که کاربرد وب کاوی، الگوی دسترسی کاربر را کشف و تجزیه و تحلیل می کند.[۱]

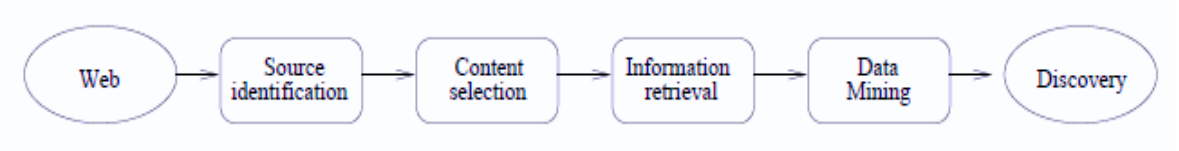
این مقاله چارچوبی برای طراحی وب در سیستم های پژوهش داده کاوی ارائه می کند. این سیستم ها برای شناسایی، استخراج، فیلترسازی و تجزیه و تحلیل داده ها از منابع وب طراحی شده اند و از تکنیک های بازیاب وب و داده کاوی با هم جهت ارائه زیر ساخت کار آمد به منظور حمایت از داده کاوی وب برای تحقیق ترکیب می کنند. این چارچوب از چند مرحله تشکیل شده است. ویژگی های هر مرحله در حال کاوش هستند و تکنیک های پیاده سازی نیز ارائه شده اند. مطالعه موردی روی داده کاوی از یک سایت توسعه یافته نرم افزار بزرگ، به طور مثال از نحوه استفاده ازاین چارچوب، ارائه شده است. کار ما یک راه حل عمومی فراهم می کند که محققان می توانند استفاده از منابع وب در پژوهش خود را دنبال

* کارشناسی نرم افزار از دانشگاه گیلان
h.ezatpanah@yahoo.com

کنند. بقیه این مقاله به شرح زیر است: بخش ۲ یک چارچوب برای حمایت پژوهش وب کاوی سیستم ها و بازبینی مختصری از اجزای آن ارائه می کند. بخش ۳ طراحی و پیاده سازی بازیابی اطلاعات وب را توصیف می کند. بخش ۴ در پردازش و تجزیه و تحلیل داده های وب تمرکز دارد. بخش ۵ یک مطالعه موردی بر اساس وب کاوی سیستم پشتیبانی پژوهش ارائه می کند. نتیجه گیری و کار آینده در بخش ۶ ارائه شده است.

۲- نمای کلی چارچوب

به منظور کشف اطلاعات وب، ما پژوهش چارچوب سیستم پشتیبانی برای داده های وب کاوی را به عنوان نشان داده شده در شکل ۱، متشکل از ۴ مرحله، می سازیم. شناسایی منبع، انتخاب محتوا، بازیابی اطلاعات و داده کاوی. در مرحله اول، وب سایت های مناسب باید با توجه به نیازهای پژوهشی انتخاب شوند. این شامل در دسترس بودن شناسایی، ارتباط و اهمیت وب سایت است. جستجوی واژه های کلیدی با استفاده از موتور جستجو می تواند برای پیدا کردن وب سایت مناسب، مورد استفاده قرار گیرد. بعد از پیدا کردن تمام وب سایت های شناسایی شده در فاز اول، مرحله دوم انتخاب مطالب مناسب در آن وب سایت است. مانند اسناد، گروه های خبری، انجمن ها، لیست پستی و ... . معمولا یک وب سایت حاوی بسیاری از صفحات وب، از جمله اطلاعات مربوط و نامربوط است. این فاز مهم است به این دلیل که تصمیم می گیرد که اطلاعات وب باید استخراج شود. انتخاب صفحات وب بر اساس هدف و تجربه محقق است. در مرحله بازیابی اطلاعات، یک خزنده طراحی شده است. در مرحله بازیابی اطلاعات، خزنده به صورت خودکار اطلاعات انتخاب شده را در طول انتخاب فاز استخراج می کند.



شکل ۱- چارچوب سیستم پشتیبانی پژوهش برای داده های وب کاوی

ابزارها و تکنیک های خاص به طور موثر برای بازیابی مفید دانش/ اطلاعات از منابع وب، به کار گرفته شدند. تلاش اضافی ممکن است برای بازیابی محتوای پویا و منابع داده مشخص مانند گروه خبری، انجمن و ... مورد نیاز باشد. مرحله نهایی، انجام داده کاوی در استخراج اطلاعات وب است. این شامل آماده سازی داده ها برای تجزیه و تحلیل است. استخراج صفحه وب ممکن است حاوی اطلاعات از دست رفته، داده های اضافی، فرمت اشتباه و کاراکترهای غیر ضروری باشد. علاوه بر این، برخی از اطلاعات باید به منظور حفاظت از حریم شخصی پردازش شوند. تکنیک های داده کاوی پیشرفته در اینجا برای کمک به تجزیه و تحلیل اطلاعات به کار گرفته می شوند.

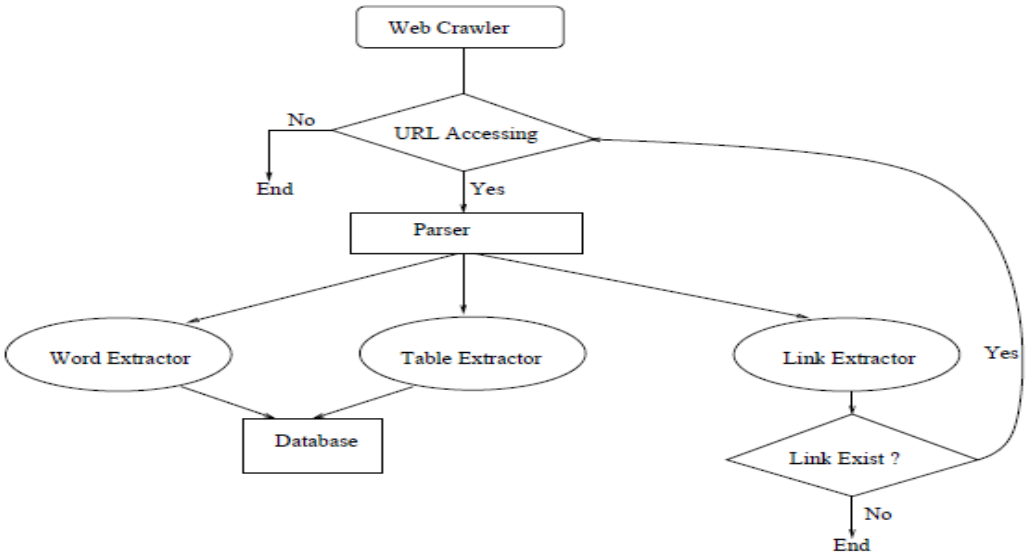
۳- بازیابی اطلاعات

بازیابی اطلاعات برای دسترسی به داده ها در وب استفاده می شود. به این معنا که سیستم پشتیبانی پژوهش وب کاوی باید قادر به جستجو و بازیابی محتویات خاص وب، به طور موثر و کارآمد باشد. دو دسته عمده از ابزار جستجو وجود دارد:

دایرکتوری(یاهو، نت اسکیپ و ...) و موتورهای جستجو(لیکس، گوگل و ...). برای استفاده ازدایرکتوری ها همراه افزایش وب سایت این کار سختی است. موتورهای جستجو نمی توانند هر نیاز مورد جستجویی را بر آورده کنند. در سیستم ما، خزنده وب بر اساس ابزارهای پیشرفته و تکنیک ها برای کمک به پیدا کردن اطلاعات مفید از منابع، توسعه داده شده است. خزنده وب عنکبوت، ربات ، کرم و ... نامیده می شود. خزنده وب شامل برنامه ای است که به طور خودکار از سایت ها عبور می کند، اسناد را دانلود می کند و لینک ها را به صفحات دیگر می فرستد.[۴]

خزنده وب یک کپی از تمام بازدیدهای صفحات برای استفاده های بعدی نگه می دارد. بسیاری از موتورهای جستجوی وب از خزنده وب برای ایجاد مطالب جهت نمایه سازی استفاده می کنند. آن ها همچنین می توانند در دیگر برنامه های کاربردی مانند اعتبار سنجی صفحه، تجزیه و تحلیل ساختاری و تجسم، به روز رسانی اخطار، معکوس سازی و ... استفاده شوند.[۲]

موتورهای جستجو برای وب کاوی یک پروژه تحقیقاتی کافی نیستند. برای طراحی، یک خزنده وب لازم است که شامل متودی برای پیدا کردن و جمع آوری تحقیقات اطلاعات مرتبط از وب باشد. اگرچه پروژه های تحقیقاتی، اطلاعات وب سایت های متفاوتی دارند که منجر به خزنده های وب می شود، این خزنده ها هنوز برخی طرح های مشترکی دارند که در شکل ۲ نشان داده شده است. آن ها می توانند با جاوا، پرل، پایتون و ... پیاده سازی و اجرا شوند. خزنده وب باید همراه یک (URL)[۲] شروع شود، دیگر لینک ها روی صفحه (HTML)[۳] شناسایی شوند، این لینک ها بازدید شوند، اطلاعات از صفحات وب استخراج شوند و اطلاعات در پایگاه داده ذخیره شوند. بنابراین، یک خزنده وب شامل یک روش دسترسی URL، تجزیه کننده صفحه وب همراه با برخی استخراج کننده ها و پایگاه داده ها است.



شکل ۲- خزنده وب

تابع دسترسی خزنده باید بارها و بارها از دسترسی به آدرس یکسان وب جلوگیری کند و باید لینک های مرده را شناسایی کند. تجزیه کننده برچسب ایتدایی، برچسب پایانی، متن و نظرات را به رسمیت می شناسد. پایگاه های داده ذخیره سازی را برای استخراج اطلاعات فراهم می کنند. اجزای کلیدی خزنده وب شامل پارسر، که خود شامل استخراج کلمه، استخراج جدول و استخراج لینک است می باشد. استخراج کننده کلمه برای استخراج اطلاعات کلمه استفاده می شود. آن

[2] **Uniform Resource Locator**
[3] **HyperText Markup Language**

باید توابع چک کردن رشته را فراهم کند. جداول معمولا در صفحات وب به منظور چینش و هم تراز کردن اطلاعات استفاده می شوند. استخراج کننده جدول محل داده ها در یک جدول را شناسایی می کند. استخراج کننده لینک، لینک های موجود در یک صفحه وب را بازیابی می کند. دو نوع از لینک ها- لینک های مطلق و لینک های نسبی وجود دارند. لینک مطلق آدرس کامل یک صفحه وب را می دهد، در حالی که لینک وابسته به اضافه کردن یک آدرس کامل با اضافه کردن یک پیشوند نیاز دارد.

۴-تکنولوژی داده کاوی

به منظور تسهیل در وب کاوی، الگوریتم های داده کاوی زیر می توانند پیدا کردن الگوها و روند هایی در اطلاعات جمع آوری شده از وب سایت اعمال کنند: خوشه بندی، طبقه بندی، قوانین انجمنی. در بخش های زیر، ما هر یک از الگوریتم ها و برنامه های کاربردی خود را توضیح خواهیم داد.

۴-۱- مشاهده قوانین انجمن

قوانین انجمن برای پیدا کردن ارتباط جالب و یا رابطه همبستگی میان یک مجموعه بزرگی از داده ها تلاش می کند. نمونه ای از قوانین انجمن کاوی تجزیه و تحلیل سبد خرید است. قوانین انجمن چیزی است مانند ۸۰٪ افرادی که آبجو، مرغ سرخ شده در روغن، خرید می کنند. قوانین انجمن همچنین می توانند برای پیش بینی الگوهای دسترسی به وب شخصی استفاده شوند. به طور مثال، ما ممکن است که ۸۰٪ افرادی که به صفحه A و صفحه B و همچنین صفحه C دسترسی دارند را کشف کنیم . صفحه C ممکن است یک لینک مستقیم از صفحه A و یا صفحه B نداشته باشد. اطلاعات کشف شده ممکن است برای ایجاد یک لینک به صفحه C و یا صفحه B مورد استفاده قرار گیرد. یک نمونه از این نرم افزار amazon.com است. ما اغلب چیزی شبیه به " مشتریانی که این کتاب و همچنین کتاب A را می خرند" می بینیم. قوانین انجمن می تواند به وب داده برای کشف رفتار کاربران وب و پیدا کردن الگوهای رفتاری خود ارائه شود.

۴-۲-طبقه بندی

هدف طبقه بندی، پیش بینی یک مورد یا چندین کلاس(یا مشاهده) است. هر مورد شامل n صفت است، که یکی از این موجودیت ها هدف است، و بقیه پیش بینی ویژگی هاست. هر یک از اهداف ارزش ویژگی های یک کلاس پیش بینی بر اساس n-1 خواص و ویژگی های پیش بینی کننده است. طبقه بندی یک فرآیند دو مرحله ای است. اول، یک طبقه بندی مدل بر اساس مجموعه ای از آموزش های داده ای ساخته شده است. دوم، مدل به منظور داده های جدید برای طبقه بندی ارائه شده است. در بین این دو مرحله، برخی مراحل دیگر ممکن است به کار گرفته شود، از جمله محاسبات بلند. محاسبات بلند برای تایید این است که آیا یک طبقه بندی، مدل با ارزشی است. مقدار بزرگتر از ۱ به طور معمول خوب است. مدل های طبقه بندی می توانند جهت تصمیم گیری کسب و کار به کار گرفته شوند. کاربردها شامل طبقه بندی عنوان پیام ها ایمیل های ناخواسته، تشخیص تقلب در کارت های اعتباری، تشخیص شبکه نفوذ و ... می باشند.

۴-۳-خوشه بندی

خوشه برای گروهبندی طبیعی از داده ها استفاده می شود. این ها گروهبندی طبیعی خوشه هستند. خوشه مجموعه داده هایی است که شبیه یکدیگرند. یک الگوریتم خوشه بندی خوب، خوشه هایی تولید می کند، به طوری که شباهت بین خوشه کم است و شباهت درون خوشه بالاست. خوشه می تواند جهت مشتریان گروه همراه رفتار مشابه و تصمیم گیری کسب و کار در صنعت استفاده شود.

۵- مطالعه موردی: منبع باز نرم افزار توسعه یافته (OSS)

جامعه (OSS)[4] مقدار قابل توجهی از زیر ساخت های اینترنت را توسعه داده است و دارای چندین دستاورد برجسته فنی، از جمله آپاچی، پرل، لینوکس و ... است. این برنامه ها نوشته شده بودند همچنین توسعه یافته شده بودند، و تا حد زیادی توسط شرکت کنندگان در بخشی از زمان اشکال زدایی شده بودند. که در اغلب موارد برای کار پرداخته نشده بودند، و از هیچ تکنیک مدیریت پروژه ای بهره نگرفته بودند. پژوهش چگونه ممکن است از تابع جامعه OSS به برنامه ریزان (IT)[5] کمک کند و منجر به تصمیم گیری آگاهانه تر و توسعه موثرتر استراتژی برای استفاده از نرم افزار OSS شود. جامعه توسعه داده شده OSS یک انجمن جهانی مجازی است. بنابراین، در حال حاضر ما مزیتی در تعامل دیجیتال داریم که بایگانی شده هستند و داده ها می توانند داده کاوی و استخراج شوند. با حدود ۷۰۰۰۰ پروژه، ۹۰۰۰۰ توسعه دهنده، و ۷۰۰۰۰۰ کاربران ثبت نام شده، SourceForge.net با حمایت مالی نرم افزار VA بزرگترین توسعه و همکاری را با OSS داشته است. این سایت اطلاعات دقیق در مورد پروژه و توسعه دهندگان، از جمله ویژگی های پروژه، فعال ترین پروژه ها و رتبه بالای توسعه دهندگان فراهم می کند.

۵-۱-جمع آوری داده ها

پس از اطلاع رسانی به سورس برنامه های ما و دریافت اجازه، ما ماهنامه داده ها را در سورس جمع کردیم. ابزار سورس مدیریت پروژه، ردیابی اشکال، فهرست خدمات پست الکترونیکی، بحث و تبادل نظر، کنترل نسخه نرم افزار برای پروژه های میزبان را فراهم می کند. اطلاعات در جامعه جمع آوری شده، پروژه و سطح توسعه کننده، مشخص کل پدیده OSS است، در سرشماری از چند پروژه، رفتار و مکانیزم هایی در محل کار و سطح پروژه و توسعه کننده بررسی می شوند. اطلاعات اولیه مورد نیاز برای این تحقیق شامل دو جدول آمار پروژه و توسعه دهندگان می باشد. جدول آمار پروژه در شکل ۱ نشان داده شده است. شامل رکوردهایی با ۹ فیلد: شناسه پروژه، نمایش های طول عمر، رتبه، صفحه، دانلود، اشکالات، پشتیبانی، تکه و (CVS)[6].

جدول توسعه دهندگان دو فیلد دارد: شناسه پروژه و شناسه توسعه دهنده. به این دلیل که پروژه ها می توانند بسیاری توسعه دهنده داشته باشند و توسعه دهندگان نیز می توانند بسیاری پروژه داشته باشند، هیچ یک از فیلدها منحصربه فرد برای کلید اصلی نیستند. بنابراین کلید کامپوزیت متشکل از هر دو ویژگی به عنوان یک کلید اصلی در خدمت است. هر پروژه در سورس دارای یک ID(شناسه) منحصر به فرد در سورس خواهد بود. خزنده وب توسط پرل و

[4] **Office of Strategic Services**
[5] **Information technology**
[6] **Concurrent Versions System**

(CPAN)[7] (جامع بایگانی پرل- مخزن ماژول پرل/ کتابخانه)، گذشته از سورس برای وب سرور جهت جمع آوری اطلاعات لازم، ماژول سازی شده است. همه صفحات اصلی پروژه ها در سورس، یک طراحی سطح بالای مشابه دارند. بسیاری از این صفحات به صورت پویا از یک پایگاه داده ایجاد شدند. خزنده وب از کتابخانه-libwww- پرل، به بهانه هر پروژه صفحه خانگی استفاده می کند. CPAN دارای یک تجزیه گرHTML عمومی جهت شناسایی و به رسمیت شناختن شروع برچسب ها، پایان برچسب ها، متن و نظرات و ... است. به این دلیل که هر دو اطلاعات آماری و عضو در جداول ذخیره شده اند، خزنده وب از ماژول پرل موجود به نام HTML:TableExtract و رشته مقایسه باارائه پرل برای استخراج اطلاعات استفاده می کند. اگر این ها بیشتر از اعضای یک صفحه باشند، استخراج لینک ها استفاده می شوند.

۵-۲-داده کاوی

چندین الگوریتم داده کاوی وجود دارد که می تواند برای وب داده استفاده شود. که در میان آن ها ساده و بی تکلف ، طبقه بندی شده، و دارای رگرسیون درخت(سبد خرید) وجود دارد. بیایید ابتدا روی الگوریتم های طبقه بندی ساده و بی تکلف بیز و سبد خرید تمرکز کنیم.

جدول ۱- آمار پروژه

project ID	lifespan	rank	page views	downloads	bugs	support	patches	all trackers	tasks	cvs
1	1355 days	31	12,163,712	71,478	4,160	46,811	277	52,732	44	0
2	1355 days	226	4,399,238	662,961	0	53	0	62	0	14,979
3	1355 days	301	1,656,020	1,064,236	364	35	15	421	0	12,263
7	1355 days	3322	50,257	20,091	0	0	0	12	0	0
8	1355 days	2849	6,541,480	0	17	1	1	26	0	13,896

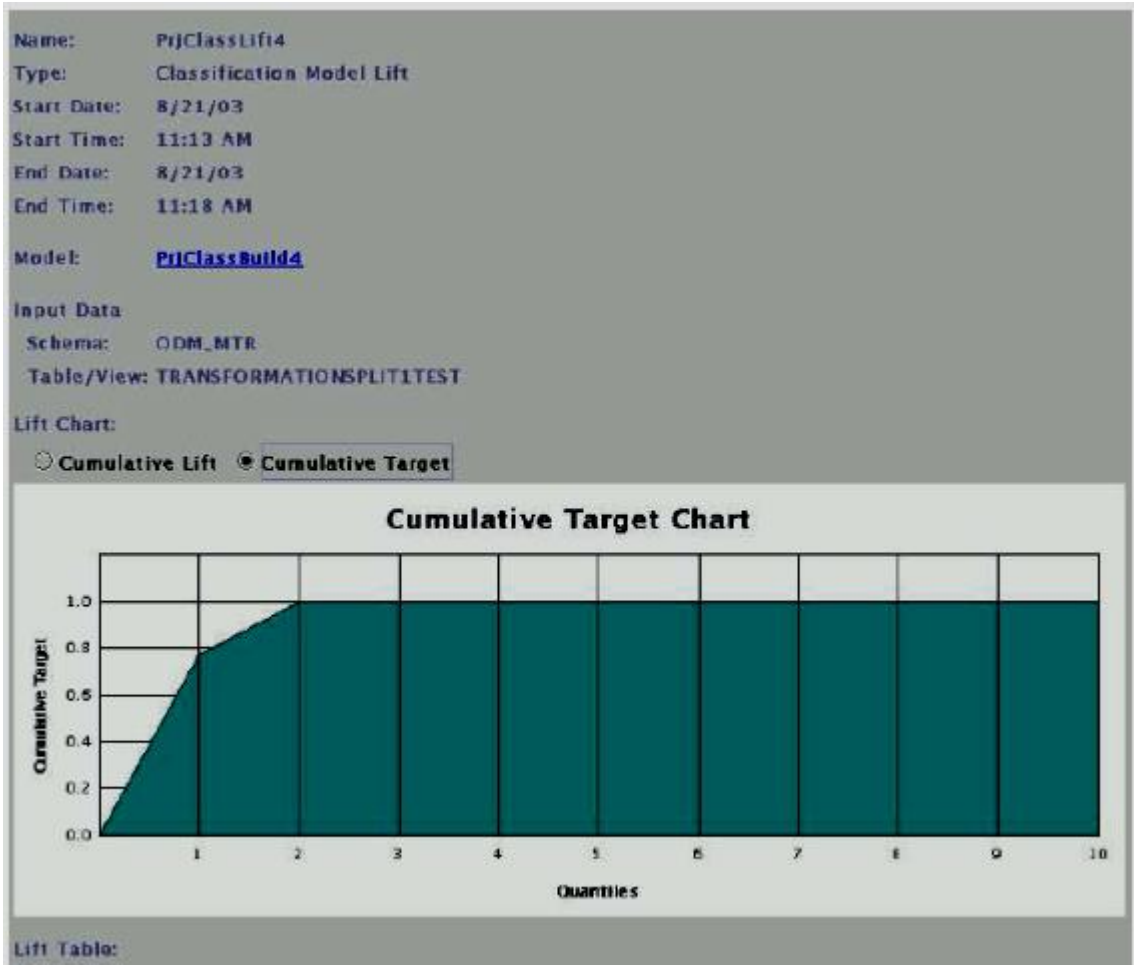
۵-۲-۱ طبقه بندی

الگوریتم ساده و بی تکلف بیز با استفاده از قضیه بیز پیش بینی می کند. ساده و بی تکلف بیز فرض می کند که هر صفتی مستقل از دیگران است. در این مطالعه، که موردی نیست، به عنوان مثال، " دانلود" ویژگی ای نزدیک به ویژگی "CVS" است، ویژگی "رتبه" از نزدیک به ویژگی های دیگر مرتبط است و از دیگر ویژگی ها محاسبه می شود. سبد خرید ایجاد طبقه بندی و درخت رگرسیون برای پیش بینی متغیر وابسته (رگرسیون) و طبقه بندی متغیرهای پیش بینی کننده(طبقه بندی) است. ما تنها زمانی علاقمند به نوع طبقه بندی از مشکلات هستیم که ما تنها از دسته این نوع مشکلات در حال استفاده باشیم. اجرا و پیاده سازی اوراکل از سبد خرید به نام تطبیقی شبکه بیز (ABN)[8] نامیده می شود. ABN پیش بینی اهداف باینری و طبقه بندی چندگانه است. بنابراین جداسازی موجودیت هدف مطلوب است. در مطالعه این مورد، ما سعی می کنیم دانلودها (بارگیری) را از ویژگی های دیگر پیش بینی کنیم.

همانطور که قبلا گفته شد، ویژگی "دانلود" نه برابر داخل باکت است. ما پیش بینی می کنیم که دریافت دانلودهای باکت بر اساس ارزش دیگر ویژگی هاست. به عنوان انتظار، الگوریتم بیز ساده و بی تکلف است و مناسب برای پیش بینی "دانلود" و آن مربوط به دیگر ویژگی ها مانند "CVS " است. دقت در بیز 10% است. در حالی که بیز روی پیش بینی

[7] **Comprehensive Perl Archive Network**
[8] **Australian Business Number**

"دانلود" به بدی عمل می کند. الگوریتم ABN می تواند " دانلود" را با کمال دقت ۶۳٪ پیش بینی کند. در ابتدای دید، دقت ۶۳٪ جذاب نیست، اما یک پیش بینی خوب است زمانی که ما می توانیم ۱۰ درصد پیش بینی درست و بدون طبقه بندی داشته باشیم. در نتیجه محاسبات بلند تایید می کند که طبقه بندی کاملا خوب است، همانطور که در شکل ۳ نشان داده شده است.



شکل ۳- نمودار بلند

این ارقام نشان می دهد که ما همه رکوردهای مربوط به ویژگی "دانلود" رامی یابیم، ویژگی ای که فقط در ۲۰٪ برای اولین بار از سوابق ۱ است. قوانین ساخته شده توسط طبقه بندی ABN نشان می دهد که " دانلود" نزدیک و مرتبط"CVS" است. ما نتیجه می گیریم که الگوریتم ABN، مناسب برای پیش بینی دانلود در مطالعه ماست. جدول زیر مقایسه دو الگوریتم یعنی ABN و Navis bayes است. از جدول، ما می بینیم که ABN جهت ساخت یک مدل طبقه بندی، بسیار طولانی طول می کشد، اما در نتیجه مدلی بسیار دقیق تر است.

جدول ۲. مقایسه بین ABN وNaive Bayes

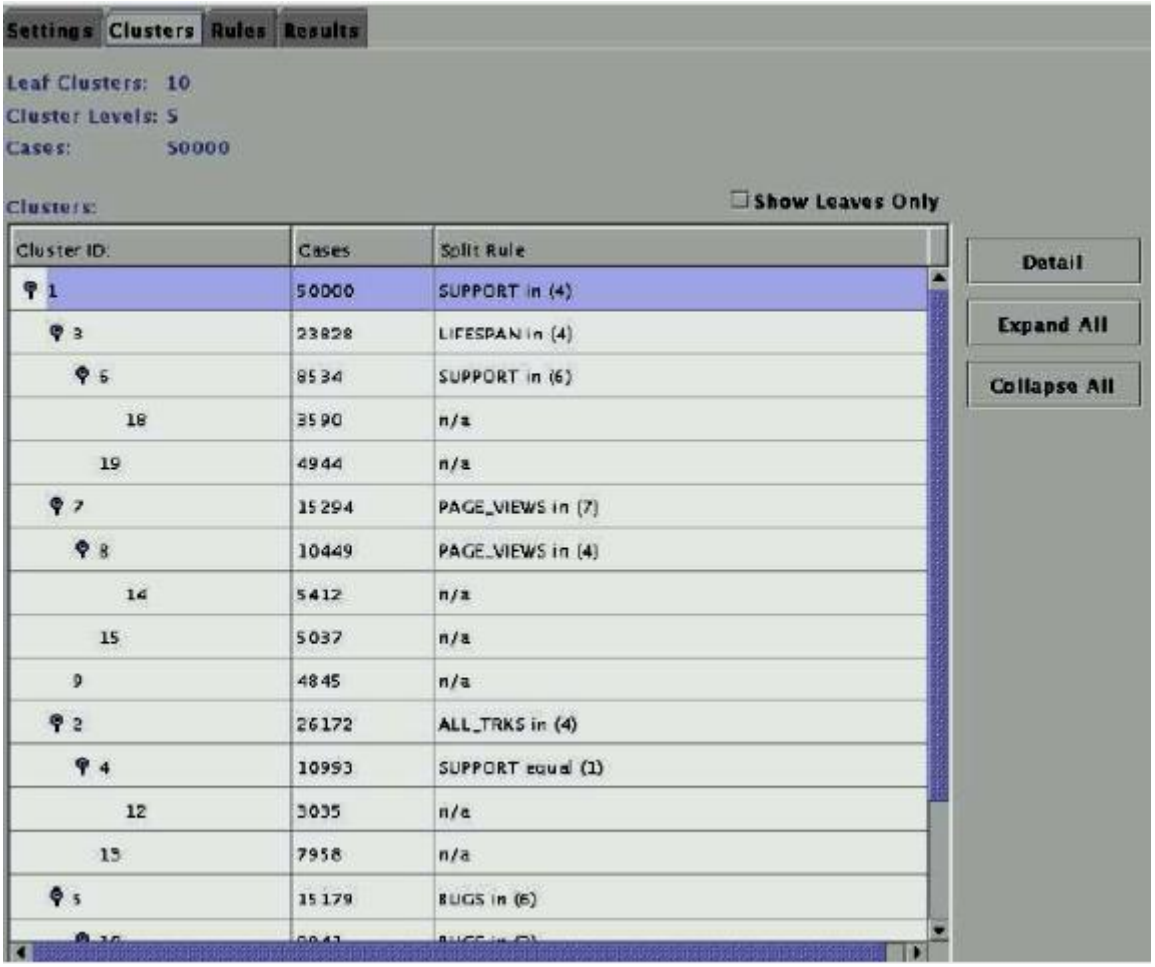
Name	Build Time	Accuracy
ABN	0:19:56	63%
Naive Bayes	0:0:30	9%

۵-۲-۲ مشاهده قوانین انجمن

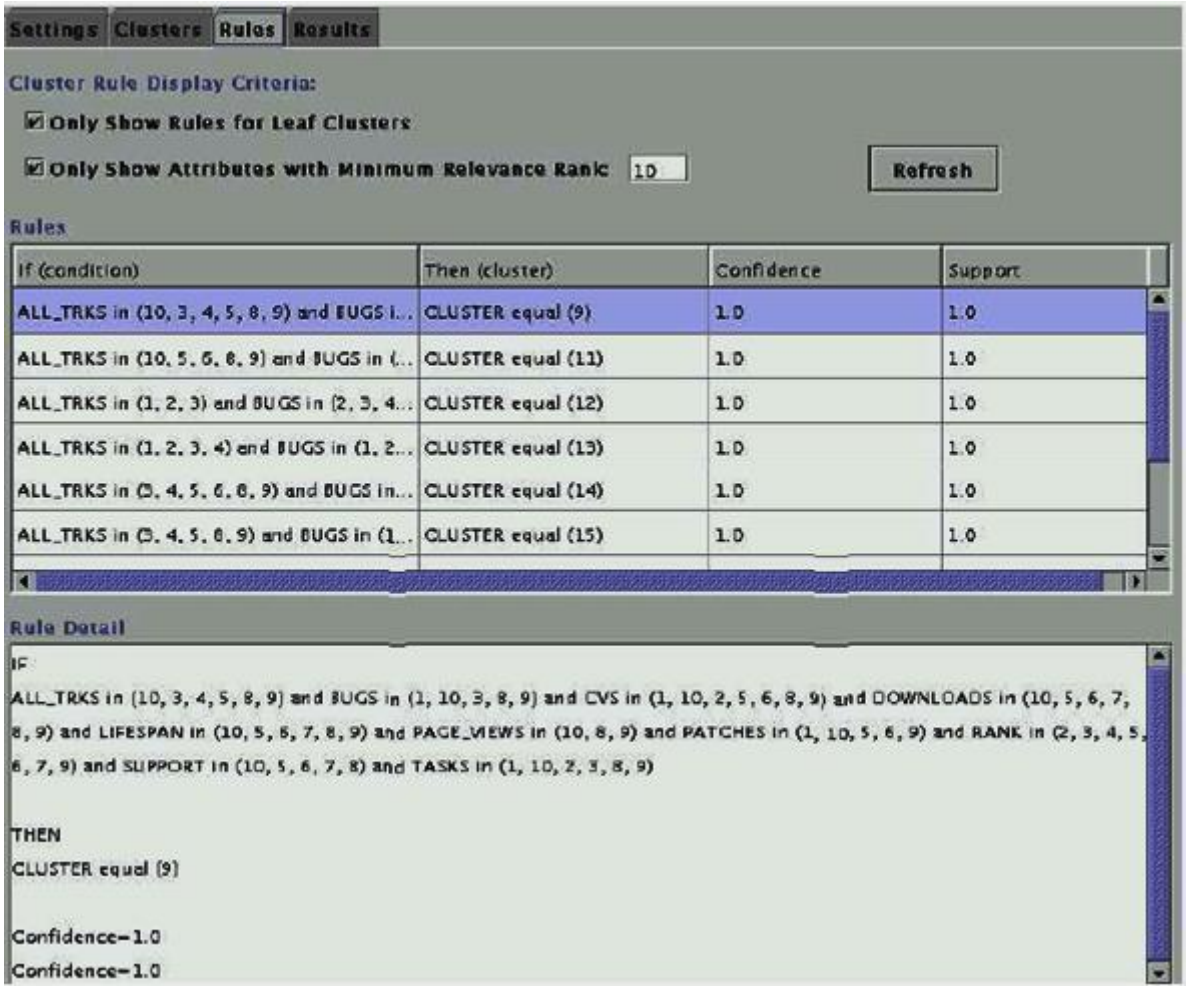
مشکل قوانین انجمن را می توان به دو مسئله زیر تجزیه کرد:
یافتن همه ترکیبات ازآیتم ها، نام مجموعه اقلام مکرر، که پشتیبانی پس از آن حداقل حمایت است. استفاده از مجموعه اقلام مکرر برای تولید انجمن، برای مثال، اگر AB مجموعه اقلام مکرر هستند، پس از آن نقش $A \rightarrow B$ را حفظ می کند اگر نسبت پشتیبانی (AB) برای حمایت از A بیشتراز حداقل اعتماد باشد. یکی از معروف ترین قوانین الگوریتم های کاوش به نام Priori است. اوراکل این الگوریتم را با استفاده از SQL پیاده سازی می کند. ما سعی می کنیم از این الگوریتم برای پیدا کردن ارتباط بین ویژگی های پروژه استفاده کنیم. الگوریتم ۲ ورودی دارد، یعنی حداقل حمایت و حداقل اعتماد به نفس. ما (۰.۰۱) را برای حداقل حمایت و (۰.۵) را برای حداقل اعتماد به نفس انتخاب می کنیم. ما همه ویژگی های CVS و دانلود را می یابیم. قوانین بیشتر را می توان در شکل ۵ دید. هیچ یک از قوانین بر اساس علاقه کشف نشده اند.

۵-۲-۳ خوشه بندی

ما علاقمند به قرار دادن پروژه با ویژگی های مشابه به صورت خوشه هستیم. دو الگوریتم را به این منظور می توان مورد استفاده قرار داد: k-means و o-cluster.
k-means یک الگوریتم مبتنی بر راه دور است، که داده ها را به تعدادی خوشه از پیش تعریف شده پارتیشن بندی می کند. o-cluster الگوریتمی سلسله مراتبی و مبتنی بر شبکه است. خوشه منجر به تعریف مناطق متراکم در ویژگی فضااست. بعد از فضای ویژگی، شماری از تعداد صفات، درگیر در الگوریتم خوشه بندی هستند. ما دو الگوریتم خوشه بندی پروژه را در مطالعه این مورد به کار می بریم. شکل ۴ و ۵ نتایج خوشه بندی و قوانینی که خوشه را تعریف می کند نشان می دهد.



شکل۴- خوشه بندی



شکل۵- قوانینی که خوشه را تعریف می کند

۶- نتیجه گیری و کار آینده

این مقاله یک چارچوب برای سیستم پشتیبانی وب کاوی پژوهش ارائه می کند و روش های آن را توصیف می کند. سپس تکنیک های پیاده سازی در استخراج و تجزیه و تحلیل داده های وب بحث می شود. سورس وب کاوی موردی است که ارائه شده است، به طور مثال، چگونگی کاربرد این چارچوب، یک مطالعه اکتشافی از بازیابی و داده کاوی داده های وب است. ما سعی می کنیم اطلاعات استخراج شده و نرم افزار داده کاوی که می تواند برای کشف دانش در وب داده استفاده شود را بررسی کنیم. نو آوری های واقعی و جالب هنوز در حال پیشرفت هستند. ما انتظار به کشف الگوهای جالب از داده ها داریم. همچنین لازم به ذکر است که این پژوهش بنیاد علوم ایالت متحده شد.

۷- مراجع

1. R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Informationand pattern discovery on the world wide web. InInternational Conference on Tools with Artificial Intelligence,pages 558–567, Newpot Beach, 1997.

2. Francis Crimmins. Web crawler review.http://dev.funnelback.com/crawler review.html, 2001.

3. Yao J.T. and Yao Y.Y.Web-based information retrieval support systems: building research tools for scientists in the new information age. In Proceedings of the IEEE/WIC International Conference on Web Intelligence, Halifax, Canada, 2003.

4. M. Koster. The web robots pages.http://info.webcrawler.com/mak/projects/robots/robots.html, 1999.

5. Bamshad Mobasher, Honghua Dai, Tao Luo, Yuqing Sun, and Jiang Zhu. Integrating web usage and content mining for more effective personalization. In EC-Web, pages 165–176, 2000.

6. Yao Y.Y. Information retrieval support systems. In FUZZIEEE' 02 in The 2002 IEEEWorld Congress on Computational Intelligence, Honolulu, Hawaii, USA, 2002.

7. Yao Y.Y. A framework for web-based research support systems. In Computer Software and Application Conference, (COMPOSAC 2003), Dallas, Texas, 2003.

8. Osmar R. Za¨iane, Man Xin, and Jiawei Han. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In Advances in Digital Libraries, pages 19–29, 1998.