کشف مسیر حرکت کاربران اطلاعات الکترونیکی با استفاده از الگوریتم قوانین وابستگی در داده کاوی: مطالعه موردی وب‌سایت کتابخانه دانشگاه یو تی اس، استرالیا

دکتر زهیر حیاتی[۱]
مرجان صادقی مجرد[۲]
نیما جعفری[۳]

چکیده

هدف اصلی این تحقیق، جستجوی روشهایی برای مطالعه رفتار کاربران در ارتباط با هدفهای آموزشی آنها در یک وب‌سایت مشخص است. در حال حاضر، داده‌کاوی، مهم‌ترین فناوری برای بهره‌برداری مؤثر، صحیح و سریع از داده‌های حجیم است. موضوع داده‌کاوی، شناخت دانش جدید و مفید، رابطه‌های منطقی و الگوهای موجود در داده‌هاست و پل ارتباطی بین علم آمار، رایانه، هوش مصنوعی، الگوشناسی، فراگیری ماشین و بازنمائی بصری داده‌ها می‌باشد. پژوهش حاضر با استفاده از تکنیک داده‌کاوی و بهره‌گیری از الگوریتم «قوانین وابستگی» روی داده‌های جمع‌آوری شده در قالب فایل ثبت وقایع وب‌سایت کتابخانه دانشگاه UTS استرالیا، به کشف الگوی مسیر حرکت کاربران در سایت پرداخته است. نتایج حاصل از این پژوهش، بینش وسیعی از رفتار کاربران و عملکرد آنها در وب‌سایت را در اختیار مدیران و طراحان آن کتابخانه قرار می‌دهد.

کلیدواژه‌ها: داده کاوی، قوانین وابستگی، تجارت الکترونیکی، کاوش کاربردی وب، وب‌سایت، کتابخانه، دانشگاه یو تی اس، استرالیا

---

۱. دانشیار بخش علوم کتابداری و اطلاع رسانی دانشگاه شیراز.
۲. کارشناس ارشد مهندسی فناوری اطلاعات، دانشگاه شیراز.
۳. کارشناس ارشد مهندسی فناوری اطلاعات، دانشگاه شیراز.

مقدمه

داده کاوی در سالهای اخیر، به دلیل در دسترس بودن حجم انبوهی از داده‌ها، توجه بسیار زیادی را در جوامع علمی و صنعت اطلاعات، به خود جلب کرده است و به عنوان یکی از پیشرفتهای اخیر در راستای فناوریهای مدیریت داده‌ها به شمار می‌رود. فناوری بر پایه وب، به دلیل فراهم نمودن امکانات مفید از جمله در دسترس بودن منابع، سادگی گسترش و به روز کردن و نگهداری آنها روی وب، به عنوان یک فناوری مناسب معرفی شده است و در بسیاری از محیطهای آموزشی توسعه‌یافته تحت وب در سرتاسر دنیا در حال استفاده از آن هستند. اگرچه ابزارهای هوشمندی برای درک رفتارهای کاربران برخط به منظور افزایش فروش و سود، توسعه یافته است، اما کارهای اندکی بر روی کشف و دسترسی به الگوهای کاربران برخط برای درک رفتارهای آموزشی آنها صورت گرفته است. مربیانی که از ابزارها و محیطهای الکترونیکی برای آموزش استفاده می‌کنند، به منظور ارزیابی فعالیتها و تمایز بین رفتارهای مختلف یادگیرنده‌های برخط با مشکلاتی مواجهند (زیان[1]، ۲۰۰۱).

با افزایش محبوبیت شبکه جهانی وب، مقدار حجیمی از داده‌ها توسط وب سرورها در قالب فایلهای ثبت وقایع وب[2] جمع‌آوری می‌شوند. این فایلها که در آنها تمامی فعالیتهای و رخ‌داده در سیستم وب‌سرور ثبت می‌شود، می‌توانند به عنوان منابع بسیار غنی از اطلاعات برای درک و تشخیص رفتار کاربران وب، استفاده شوند. کاوش کاربردی وب[3] که آن را کاوش فایل ثبت وقایع در وب[4] نیز می‌نامند، در واقع استفاده از الگوریتمهای داده کاوی بر روی فایلهای ثبت وقایع وب به منظور پیدا کردن مسیر حرکت و نظم موجود در الگوهای جستجوی کاربران وب است (سن[5]، ۲۰۰۵).

---

1. Osmar Zaiane.
2. Web Access logs.
3. Web Usage Mining.
4. Web Log Mining.
5. Sen Zhang.

هدف از انجام این پژوهش، دست یافتن به رفتار کاربران با استفاده از فناوری داده کاوی در وب‌سایت کتابخانه دانشگاه یو تی اس UTS استرالیا و کشف قوانین موجود در داده‌های جمع‌آوری شده در طول ۷ ماه در قالب فایل ثبت وقایع است. این قوانین می‌توانند مدیران کتابخانه و بخش فناوری اطلاعات این دانشگاه را در تصمیم‌گیریهای مهم توسعهٔ مجموعه اطلاعاتی و طراحی کارآمد وب‌سایت به منظور افزایش رضایت کاربران یاری دهد.

پیشینهٔ پژوهش

داده‌کاوی، فرایندی است که در آغاز دهه ۹۰ پا به عرصه ظهور گذاشته است و با نگرشی نو به مسئله استخراج اطلاعات از پایگاه داده‌ها می‌پردازد. در سالهای ۱۹۸۹ و ۱۹۹۱، کارگاه‌های کشف دانش از پایگاه داده‌ها توسط «پیاتتسکی و همکارانش» و در فاصله سالهای ۱۹۹۱ تا ۱۹۹۴ کارگاه‌های فوق، توسط «فایاد و پیاتتسکی» برگزار شد. به طور رسمی، اصطلاح داده‌کاوی برای اولین بار توسط «فیاض» در اولین کنفرانس بین‌المللی کشف دانش و داده‌کاوی در سال ۱۹۹۵ مطرح شد. از سال ۱۹۹۵ داده‌کاوی به صورت جدی وارد مباحث آمار گردید (فیاد، پیاتتسکی و اسمیت، ۱۹۹۶). در سال ۱۹۹۶ اولین شمارهٔ مجله «کشف دانش» از پایگاه داده‌ها منتشر شد.

امروزه کنفرانسهای مختلفی در این زمینه در سراسر دنیا برگزار می‌شود. داده‌کاوی با همه‌گیر شدن استفاده از پایگاه‌های داده‌ای به عنوان یک علم مطرح شده است (کوئین لن[1]، ۱۹۹۲). «راسل» (۱۹۹۸) معتقد است افزایش رشد شبکه جهانی وب، یک منبع جدید گسترده و بزرگ از اطلاعات قابل دسترس به وجود آورده است که بسیاری از وب‌سایتها تمایل دارند هدفهای آموزشی خود را از طریق آن انجام دهند. سرعت توسعه و رشد وب، از میزان توسعهٔ روشهای مطالعه کارآمد وب‌سایتها به عنوان ابزاری برای پشتیبانی آموزش و یادگیری پیشی گرفته است.

---

1. Ross Quinlan.

هدف اصلی این تحقیق، جستجوی روشهایی برای مطالعهٔ رفتار کاربران در ارتباط با هدفهای آموزشی آنها در یک وب‌سایت مشخص بود. هدف این مطالعه، کمک و یاری به توسعه‌دهندگان وب‌سایتها به منظور انتخاب تکنیکهای کارآمد برای ارزیابی سایت است. پژوهشگران دیگری چون «نیکولاس»، «هانتینگتون» و «جمالی» (۲۰۰۶) نیز به مطالعهٔ رفتار اطلاع‌یابی کاربران با به‌کارگیری فنون داده‌کاوی وب پرداخته‌اند. آنها دریافتند که بسیاری از کاربران وب برای زمانهای طولانی صفحات وب را مطالعه نکرده و قبل از ترک منابع وبی، تنها به بررسی اجمالی اقلام و صفحات وبی محدودی مشغول بوده‌اند. «بریدینگ» (۲۰۰۵) با به‌کارگیری گروه ویژه از کاربران وب‌سایتها و نرم‌افزارهای تجزیه و تحلیل وب‌لاگ‌ها، رفتار اطلاع‌یابی کابران را در سطوح عمیق‌تر مطالعه کرده است. او به مطالعهٔ رفتار اطلاع‌یابی کاربران انفرادی اکتفا نکرده، بلکه به مطالعهٔ گروهی از آنها از طریق جلساتی که برگزار کرده‌اند نیز پرداخته است.

«هانتینگتون، نیکولاس و جمالی» (۲۰۰۷) با مطالعهٔ تراکنشهای جستجوی وب‌لاگ‌ها بیان می‌دارند که ابزارهای اندازه‌گیری که از این منابع کشف می‌شود، وسایل سودمندی برای بررسی میزان کارایی و همچنین میزان رضایت و عدم رضایت از موتورهای جستجو می‌باشند. آنها دو معیار اندازه‌گیری زمان سپری شده میان جلسات جستجو و تعداد جستجوهای انجام شده در هر جلسه را برای مطالعهٔ رفتار اطلاع‌یابی کاربران موتورهای جستجو به‌کار گرفتند. مطالعهٔ دیگری در همین سطح توسط «نیکولاس، هانتینگتون و واتکینسون» (۲۰۰۵) در مورد رفتار اطلاع‌یابی کاربران کتابخانه‌های مجله‌های دیجیتالی انجام گردید. تمرکز آنها بر روی کاربران پایگاه اطلاعاتی Blackwell Synergy بود و معیارهای تعداد جلسات برگزار شده و اقلام مورد مشاهده و مورد تقاضا را برای بررسی رفتار اطلاع‌یابی اعضای هیئت علمی مجله‌های دیجیتالی پایگاه مذکور به کار گرفتند. این پژوهشگران بیان می‌دارند چنانچه این نوع مطالعات با مطالعات کیفی رفتار اطلاع‌یابی کاربران تکمیل گردد، به نتایج بهتر و واقعی‌تری می‌توان دست یافت.

منبع داده در این پژوهشها، تمامی صفحات رؤیت شده توسط مشتریان سایت در یک فایلِ ثبت وقایع روی وبِ سرور بوده است. تحلیل این فایلهای داده، به ارزیابی‌کنندگان سایت کمک می‌کند تا نقاط اصلی مسیر حرکت سطوح پرترافیک در سایت را تشخیص دهند. جستجوکنندگان می‌توانند در مورد هویت دیدارکنندگان سایت، صفحات و بخشهایی را که در یک سایت توسط دیدارکنندگان دیده شده است، استخراج کنند.

کاوش کاربردی وب، به عنوان یکی از کاربردهای تکنیک داده‌کاوی به منظور استفاده از فایلهای ثبت وقایع برای بهبود طراحی وب‌سایتهاست (کولی[1]، مباشر[2] و سریواستاوا[3]، ۱۹۹۹). فایلهای ثبت وقایع وب سرورها به صورت بالقوه شامل داده‌های تجربی مفیدی برای بهبود کارایی وب‌سایتها هستند و منافعی را برای بعضی از کاربردها، بخصوص موارد تجاری، در بردارند. با تحلیل این فایلها می‌توان به پیش‌بینی لینکهایی پرداخت که در افزایش کارایی وب‌سایت تأثیر مثبت دارند و برای طراحان وب‌سایت بسیار مفیدند (یانگ[4]، ۲۰۰۵). به عنوان مثال، پیش‌بینی لینک‌ها می‌تواند برای بارگذاری اسنادی که ممکن است دیدارکننده از آنها دیدن کند، در زمانی که وی در حال خواندن صفحهٔ جاری است، تأثیری بسیار مثبت در کار وی داشته باشد. با استفاده از فایل ثبت وقایع جریان کاری می‌توان راهکارهایی را برای حلّ مشکلات موجود در بهبود فرایندهای کسب و کار ارائه نمود (سابرامینام[5]، ۲۰۰۶).

روش شناسی پژوهش

در این پژوهش، از روش تجزیه و تحلیل الگوریتمها و اجرای الگوریتم قوانین وابستگی روی داده‌های موجود در بانک اطلاعات به منظور کشف وابستگی بین اطلاعات

1. Robert Cooley.
2. Bamshad Mobasher.
3. Jaideep Srivastava.
4. Zhijian Yang.
5. Sharmila Subramaniam.

و اقلام موجود در بانک اطلاعات و پیش‌بینی قوانین وابستگی به منظور بهبود طراحی وب‌سایت، استفاده شده است.

مجموعه داده‌ها

از فایلهای ثبت وقایع کتابخانه دانشگاه UTS استرالیا به منظور کشف قوانین وابستگی در این مجموعه اطلاعات استفاده شده است. این اطلاعات مجموعه داده‌های عمومی جمع‌آوری شده حاصل از تمامی فعالیتها و وقایع مربوط به کاربران دانشجو در مقاطع کارشناسی، کارشناسی ارشد و دکتری است که از سراسر جهان به پایگاه‌های اطلاعاتی موجود در کتابخانه دانشگاه UTS مراجعه کرده و سپس مجوز استفاده از این پایگاه‌ها و اطلاعات موجود در آنها را دریافت کرده‌اند. عموماً، تولیدات علمی در این مجموعه، اطلاعاتی در قالب فایلهایی با انواع مختلف مانند Pdf ، Doc، Zip، Rtf ، exe ، Txt و PPt به کاربران عرضه شده است.

سپس با استفاده از فناوری OLAM[1] و به‌کارگیری الگوریتم قوانین وابستگی بر روی اطلاعات جمع‌آوری شده از مراجعات کاربران به این مرکز در مدت ۷ ماه (سپتامبر ۲۰۰۶،ژانویه ـ ژوئیه ۲۰۰۷) اقدام به کشف مسیر و الگوی حرکت کاربران می‌شود. اطلاعات کاربران این مرکز با توجه به اطلاعات حاصل از فایل ثبت وقایع، شامل دانشکدهٔ محل تحصیل، محل اشتغال و همچنین مکان جغرافیایی استقرار کاربران می‌باشد.

شناسایی قالب اطلاعات ذخیره شده

فایلهای ثبت وقایع، اغلب برای کاوش کاربردی وب استفاده و در سه فُرمت عمومی[2]، توسعه‌یافته[3] و اختصاصی دسته‌بندی می‌شوند. در این پژوهش، فرمت فایل ثبت وقایع استفاده شده جهت ذخیرهٔ اطلاعات از نوع قالب عمومی بوده و دارای فیلدهای زیر است:

1. Online Analytical Mining.
2. Common Log Format (CLF).
3. Extended Common Log Format (ECLF).

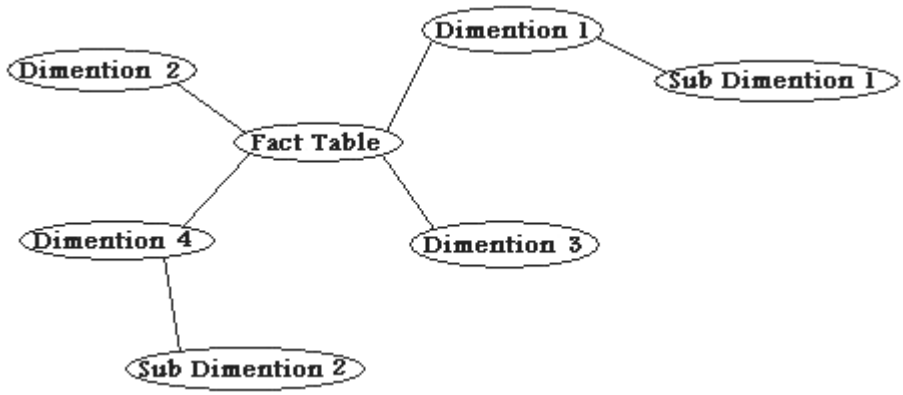
شکل ۱. فیلدهای موجود در فایل ثبت وقایع عمومی

فایل ثبت وقایع در دسترس دارای حجم حدود ۱۶ گیگا بایت بود که پس از عمل پاکسازی، یکپارچه‌سازی و انتقال به حجم ۵۰۰ مگابایت، در قالب بانک اطلاعات در محیط sqlserver 2005 کاهش یافت. در این مجموعه، ۱۶۹ وب‌سایت منحصر به فرد موجود است (وب‌سایتها دربرگیرندهٔ بانکهای اطلاعاتی هستند که کاربران به آنها مراجعه می‌کنند). همچنین، حدود ۲۱۳۰۰۰ آدرس اینترنتی و ۲۰۰۰۰۰ فایل در این مجموعه دانلود شده است. مجموعه فایلهای استفاده شده در این مجموعه پس از انجام پاکسازی اطلاعات، ۱۰ نوع فایل می‌باشد که شامل txt، doc، zip، rtf، ppt، pdf، xls، html، asp، do، php، htm، aspx هستند. پس از انجام مراحل پیش پردازش ۲۰۰۰۰۰۰ درخواست حاصل و در بانک اطلاعات ثبت شد.

ساختار انباره داده[1]

پس از بررسی ساختار فایل ثبت وقایع و اطلاعات موجود در آن، انباره داده ایجاد گردید. انباره داده به عنوان یک منبع نگهداری مجموعه‌ای از داده‌های جمع آوری شده از چندین مرجع یا منبع داده است که معمولاً ناهمگن و هدف آن ایجاد مجموعه‌ای تحت یک طرح و ساختار به صورت یکپارچه است. در پژوهش حاضر، از ساختار برف‌دانه‌ای به منظور ایجاد انباره داده استفاده شده است که دارای یک مرکز با عنوان جدول اصلی و چندین بُعد می‌باشد و ابعاد از طریق کلید اصلی با مرکز ارتباط دارند.

1. Data Warehousing.



شکل۲. ساختار مدل برف‌دانه‌ای در ایجاد انباره داده

پیش پردازش[1]

در مرحله پیش پردازش، سه عمل اصلی بر روی داده‌های موجود در فایل ثبت وقایع انجام می‌شود که شامل پاکسازی و یکپارچه‌سازی، تبدیل داده‌ها و در نهایت بارگذاری در انباره داده است. در مرحلهٔ پاکسازی، اطلاعاتی که در انجام و اجرای مراحل داده‌کاوی ضرورتی به وجودشان نبود، حذف شدند تا در محاسبات شرکت داده نشوند. این اطلاعات شامل تراکنشهای موجود همراه با آدرسها و فایلهایی که صرفاً جهت ساخت یک صفحه وب استفاده می‌شوند، بود. فایلهای تصویری و کدهای جاوا اسکریپت و فایلهای مربوط به قالب و شکل ظاهری صفحات وب‌سایت در طی این مرحله حذف شدند. در این پژوهش تنها منبع اطلاعات، فایل ثبت وقایع بوده، بنابراین مرحله یکپارچه‌سازی در طول فرایند پیش‌پردازش حذف گردید. بعضی از فیلدهای موجود در فایل مانند تاریخ و زمان که دارای مقادیر ترکیبی بودند، به منظور کاوش عمیق‌تر در داده‌ها تجزیه شدند. پس از انجام مراحل فوق، داده‌های حاصل از سه مرحله قبل به درون انباره داده، انتقال یافت.

1. Preprocessing.

شناسایی و معرفی قوانین وابستگی در داده

منظور از قوانین وابستگی، کشف وابستگی بین اقلامی است که رخداد آنها در یک زمان است؛ برای مثال، اجناسی که در یک فروشگاه احتمال خرید آنها با هم در یک تراکنش خرید زیاد است. این اقلام دارای وابستگی هستند که این وابستگی‌ها به صورت A->B نمایش داده می‌شود. به A مقدم و به B مؤخر یا نتیجه گفته می‌شود. کشف مجموعه عناصر تکرار شونده، به کشف وابستگی بین عناصر در مجموعه داده‌ها با حجم زیاد منجر می‌شود. بسیاری از صنایع مشتاقند تا با داشتن حجم عظیمی از داده‌هایی که به طور پیوسته جمع‌آوری و ذخیره می‌شوند، چنین الگوهایی را از بانکهای اطلاعاتی خود استخراج کنند. کشف روابط وابستگی قابل توجه در بین حجم عظیمی از تراکنشهای کسب و کار ثبت شده، در بسیاری از فرایندهای تصمیم‌گیری کسب و کار مانند طراحی کاتالوگ، بازاریابی عرضی و تحلیل رفتار خرید مشتریان کمک می‌کند.

از مقیاسهای مهم در قوانین وابستگی که به منظور ارزیابی قوانین کشف شده مورد استفاده قرار می‌گیرد، Support و confidence هستند که به ترتیب سودمندی و قطعیت قوانین کشف شده را نتیجه می‌دهند.

- **Confidence:** زمانی که خرید یک قلم به خرید اقلام دیگری منجر می‌شود، احتمال رخداد با استفاده از این معیار اندازه‌گیری می‌شود.
- **Support:** اگر خرید دو کالا با هم انجام شود، میزان احتمال رخداد آن با این معیار، اندازه‌گیری و میزان درصد خرید آنها با هم، با عدد support نشان داده می‌شود.

قوانین وابستگی دارای یک آستانه حداقل support و یک آستانه حداقل confidence هستند که با توجه به این مقدار آستانه، معناداری قوانین تشخیص داده می‌شود. این آستانه می‌تواند توسط کارشناسان و یا نرم‌افزار، تنظیم شود. تحلیلهای بیشتر می‌تواند برای کشف وابستگی‌های قابل توجه بین عناصر وابسته به‌کار گرفته شود.

کشف قوانین وابستگی، دارای دو مرحلهٔ تکرارپذیر است:

۱. کشف تمامی مجموعه یا itemset های تکرارپذیر

۲. تولید قوانین محکم از itemset های تکرارشونده

در این پژوهش، به منظور کشف مجموعه‌های تکرارپذیر در قوانین وابستگی منطقی، از الگوریتم ای‌پریوری[1] استفاده شده است. «ای پریوری» یک روش تکرارپذیر به کار می‌گیرد که k-itemset ها برای یافتن (K+1)-itemset ها مورد استفاده قرار می‌گیرند و از دو بخش الحاق[2] و هرس[3] تشکیل شده‌اند. زمانی که itemset های تکرارپذیر از بین تراکنشها در بانک اطلاعات به دست آمدند، ایجاد قوانین وابستگی محکم از آنها به راحتی امکان‌پذیر است که با استفاده از معادله زیر انجام می‌شود:

Confidence (A⇒B) = P (B|A) = support_count (A∪B)/ support _ count (A)

احتمال شرطی فوق بیان می‌کند که support_count(A∪B) تعداد تراکنشهای شامل A∪B و support_count(A) تعداد تراکنشهای شامل itemset ، A می‌باشند. بر اساس این معادله، قوانین وابستگی می‌تواند به صورت زیر تولید شود:

- برای هر itemset تکرارپذیر L، همه زیر مجموعه‌های غیر تهی آن ایجاد شود.
- برای هر زیر مجموعه غیر تهی S از L قانون:

S ⇒(l-s): if (support_count (l)/support_count(s)) ≥min_conf, where min_conf is the minimum confidence threshold.

یعنی اگر تعداد تکرارهای Itemset انتخابی l بر تعداد تکرارهای زیرمجموعه انتخاب شده از آن، از مقدار min_conf بزرگتر باشد آنگاه s⇒(l-s) یک قانون وابستگی است. چون قوانین از itemset های تکرارپذیر تولید می‌شوند، هر کدام به صورت خودکار، minimum support مورد نظر را دارند.

تجزیه و تحلیل داده‌ها

در این بخش یافته‌ها بر اساس الگوریتم «قوانین وابستگی» و اجرای این الگوریتم بر روی داده‌های موجود در انباره داده حاصل از عمل پیش‌پردازش فایلهای ثبت وقایع،

1. Aprior Algorithm.
2. Join.
3. Prune.

تجزیه و تحلیل شده است. در ابتدا، مدلهای داده کاوی روی داده‌های موجود در جدولهای مختلف موجود در انباره داده، طراحی شده و با اجرای الگوریتم مورد نظر، قوانین وابستگی و رابطه‌های موجود بین اقلام اطلاعاتی، کشف و بر اساس این قوانین، پیش‌بینی‌هایی صورت گرفته که هر کدام به صورت مدل جداگانه در این بخش آورده شده است. در زمان طراحی مدل، تعریف متغیرهای ورودی و متغیرهایی که پیش بینی روی آنها انجام می‌شود، ضروری است. تعیین این متغیرها و انتخاب آنها به عنوان ورودی و متغیر پیش‌بینی شونده، اهمیت بسیاری دارد و می‌تواند در روند ایجاد مدل و نتایج خروجی و همچنین معناداری قوانین کشف شده، تأثیر بسیاری بگذارد. بنابراین، این مرحله به دانش و مطالعه روی فیلدهای تعریف شده در انباره داده و تسلط کافی بر درک داده‌ها نیاز دارد.

الگوریتم «قوانین وابستگی» در داده‌های موجود در انباره داده به دنبال مجموعه‌های تکرارپذیر معنادار که معناداری آنها بر اساس معیار minimum_support ارزیابی می‌شود، جستجو کرده و به فهرست Itemset های معنادار تکرارپذیر دست می‌یابد. سپس در این مجموعه‌ها به دنبال کشف روابط وابستگی نهفته بین اقلام هر مجموعه و مجموعه‌ها با یکدیگر، قوانینی را با ضرایب معناداری مختلف که بر اساس معیار minimum_probabilty ارزیابی می‌شود، استخراج می‌کند. قوانین کشف شده دارای مقادیر مختلف Confidence (که در نرم‌افزار Sql Server2005 با عنوان Probability نام برده شده است) بوده و بیانگر احتمال رخداد آن قانون است. در تمام مدلهای ارائه شده، از مقدار پیشنهادی نرم‌افزار برای minimum_probabilty و minimum_support استفاده شده است. برای استفاده از الگوریتم، نرم‌افزار Sql Server نسخه ۲۰۰۵ و Sql Server Analysis Services نرم‌افزار Microsoft Visual Studio.net نسخه ۲۰۰۵ استفاده شده است. مدلهای طراحی شده روی سه مقطع تحصیلی کارشناسی، کارشناسی ارشد و دکتری اجرا شده است.در مدلهای استفاده شده، کلیهٔ اطلاعات علمی و اطلاعاتی با پسوندهای pdf ، doc، txt، zip، xls، ppt و rtf

و پسوندهای htm, html, asp, aspx ,php, do به عنوان صفحات ملاقات شده توسط مشتریان در نظر گرفته شده است.

مدل کاوش شماره ۱

این مدل رفتار مشتریان در استفاده از پایگاه‌ها و صفحات پر استفاده توسط آنها پیش‌بینی شده است. در جدول ۱، بخشی از عناصر تکرارپذیر که رخداد آنها با هم بوده، آورده شده است:

جدول۱. بخشی از مجموعه‌های تکرارپذیر پس از اجرای الگوریتم Association Rules با minimum support=1

Row	Support	Size	ItemSet
۱	۷	۲	help/whgdata/ = Existing, w Name = csa.com
۲	۷	۱	rpsv/cw/vhosts/oecdthemes/99980037/v1998n1/ = Existing
۳	۷	۳	ids70/ = Existing, w Name = csa.com, help/ = Existing
۴	۷	۲	ids70/ = Existing, help/ = Existing
۵	۷	۳	csaillumina/ = Existing, w Name = csa.com, help/ = Existing

ردیف ۲ نشان می‌دهد که مسیر rpsv/cw/vhosts/oecdthemes/99980037/v1998n1 چندین مرتبه و به تکرار ملاقات شده است. همچنین، ردیف ۳ نشان می‌دهد مسیر help و وب‌سایت csa.com به تکرار با هم دیده شده‌اند.

قوانین کشف شدهٔ مدل کاوش ۱

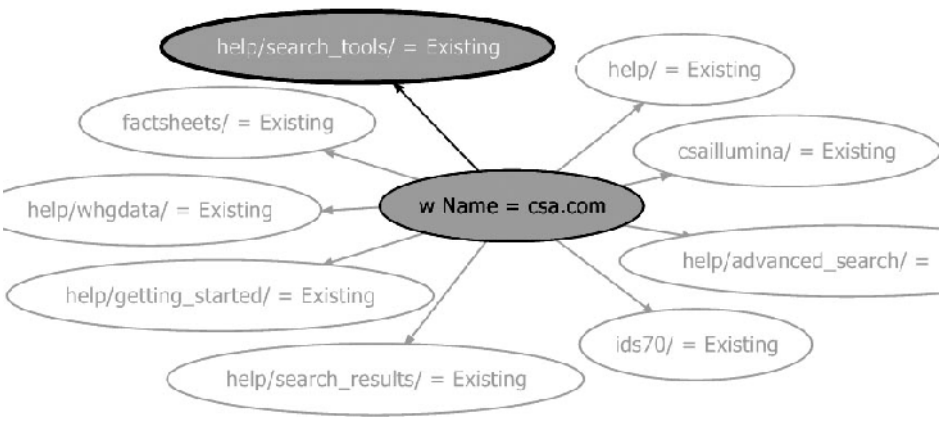
از مجموعه‌های تکرارپذیر حاصل شده در مرحلهٔ قبل، روابط وابستگی بین اقلام یک مجموعه و مجموعه‌ها با یکدیگر کشف و قوانین وابستگی ارائه شده، در جدول ۲ ارائه شده است:

جدول۲. قوانین کشف شده پس از تعیین مجموعه های تکرارپذیر با minimum probability=0.4

Row	Probability	Rule
1	0.667	w Name = sourceoecd.org -> rpsv/cw/vhosts/oecdthemes/99980126/v2003n19/ = Existing
2	0.7	w Name = csa.com -> csaillumina/ = Existing
3	0.7	w Name = csa.com -> help/whgdata/ = Existing
4	0.75	w Name = igi-online.com -> content/ = Existing
5	0.778	w Name = sourceoecd.org -> rpsv/cw/vhosts/oecdthemes/99980037/v1998n1/ = Existing
6	0.8	w Name = csa.com -> ids70/ = Existing
7	0.8	w Name = csa.com -> help/ = Existing
8	0.8	w Name = lib.uts.edu.au -> / = Existing
9	0.889	w Name = sourceoecd.org -> rpsv/cgi-bin/fastforward/ =

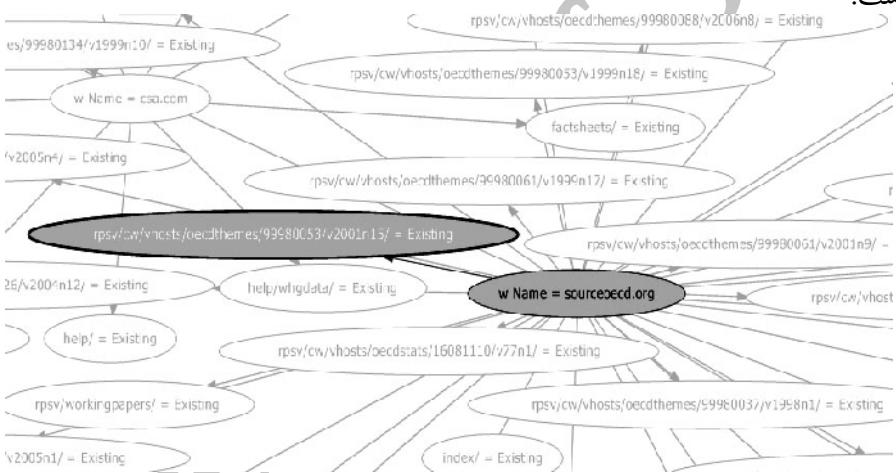
برای مثال، قانون شماره ۷ بیان می کند ۸۰٪ مشتریانی که از پایگاه‌های موجود روی csa.com استفاده می کنند، وارد صفحه راهنما می‌شوند. همچنین، قانون شماره ۹ بیان می کند ۹۰٪ مشتریانی که از پایگاه‌های sourcecode.com استفاده می کنند، وارد صفحه rpsv/cw/cgi-bin/fastforward می‌شوند.

همچنانکه در شبکهٔ وابستگی مربوط به این مدل نشان داده شده، مسیر rpsv/cw توسط مشتریان بسیار استفاده شده است.



شکل۳. شبکهٔ وابستگی پایگاه csa.com و رفتار حرکت مشتریان مراجعه‌کننده به این وب‌سایت

در این شبکه، به صورت واضح نشان داده شده است که مشتریان در زمان ورود به سایت csa.com، به دفعات وارد صفحهٔ راهنما می‌شوند. دلیل این امر می‌تواند ناآشنا بودن مشتریان با این سایت و پایگاه‌های موجود در آن باشد که در زمان ورود به دلیل ناآگاهی از هدفهای سایت دچار مشکل شده‌اند و این از دلایل طراحی ضعیف سایت است. همچنین، این احتمال وجود دارد که مشتریان در یافتن اطلاعات مورد نیاز خود با مشکلاتی روبه‌رو می‌شوند؛ لذا در صدد رفع نیاز خود، به راهنمای سایت مراجعه می‌کنند. در قانون دیگری که توسط این الگوریتم کشف شده، مسیر پر استفاده در پایگاه‌های موجود در sourceoecd.com می‌باشد. در زیر، شبکهٔ وابستگی این قانون نشان داده شده است.



شکل٤. شبکهٔ وابستگی پایگاه sourceoecd.com و رفتار مصرف مشتریان مراجعه کننده به این وب‌سایت

شبکهٔ فوق، میزان استفادهٔ مشتریان مراجعه‌کننده به وب‌سایت sourceoecd.com از مسیر rpsv/cw و صفحات پرمصرف را نشان می‌دهد. مشتریان در زمان ورود به سایت sourcecode.com به منظور استفاده از پایگاه‌های موجود روی آن، به تکرار وارد صفحه rpsv/cw/cgi-bin/fastforward می‌شوند و این نشان می‌دهد در این مسیر اطلاعات مفیدی وجود دارد. همچنین، مسیر rpsv/cw از دیگر مسیرهای پر استفاده

توسط مشتریان است. با توجه به این قوانین، می‌توان با بررسی بیشتر اطلاعات موجود در مسیرهای پرمصرف، دسته‌بندی بهتری را برای چینش اطلاعات در نظر گرفت و آنها را در مسیرهای کوتاه‌تر قرار داد. بدین ترتیب، ترافیک شبکه کنترل شده و مشتریان در یافتن نیازهای خود به رضایت بیشتری دست خواهند یافت.

مدل داده‌کاوی شماره۲

در این مدل، رفتار اطلاع‌یابی مشتریان مقیم در کشورهای مختلف در مقطع لیسانس، پیش‌بینی شده است. در جدول ۳ عناصر تکرارپذیری که رخداد وقوع آنها با هم بوده، آورده شده است.

جدول۳. بخشی از مجموعه‌های تکرارپذیر کشف شده پس از اجرای الگوریتم Association Rules با minimum support=16

Row	Support	Size	Item Set
1	212	2	umi.com = Existing, lib.uts.edu.au = Existing
2	212	1	umi.com = Existing
3	210	2	lexisnexis.com = Existing, umi.com = Existing
4	209	2	ebsco.com = Existing, umi.com = Existing
5	207	3	ebsco.com = Existing, lexisnexis.com = Existing, umi.com = Existing

در این جدول، بخشی از عناصر تکرار شونده آورده شده است. برای مثال، ردیف ۴ نشان می‌دهد رخداد ملاقات پایگاه‌های موجود در ebsco.com و umi.com به تکرار با هم بوده و تعداد مرتبه این رخداد ۲۰۹ است و این با توجه به حداقل مقدار Support ، مقدارقابل توجهی است.

قوانین کشف شده مدل کاوش ۲

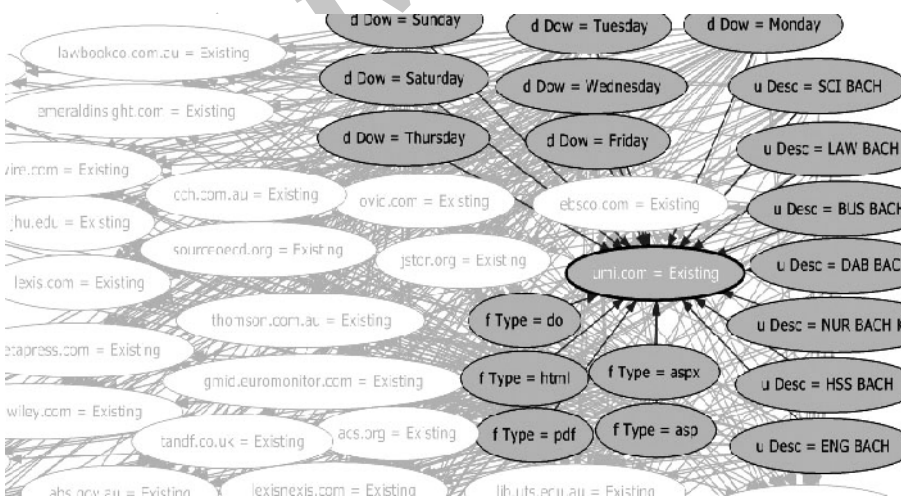
پس از ایجاد مجموعه‌ها، رابطه‌های وابستگی موجود بین این مجموعه‌ها و اقلام هر مجموعه کشف شد:

جدول٤. قوانین کشف شده پس از تعیین مجموعه‌های تکرارپذیر ارائه شده در مرحلهٔ قبل با minimum probability=0.43

Row	Probability	Rule
1	0.974	u Desc = DAB BACH -> umi.com = Existing
2	0.967	f Type = do -> umi.com = Existing
3	0.967	u Desc = NUR BACH KC -> umi.com = Existing
4	0.967	u Desc = HSS BACH -> umi.com = Existing
5	0.933	d Dow = Tuesday -> umi.com = Existing
6	0.874	u Desc = SCI BACH -> umi.com = Existing
7	0.874	u Desc = LAW BACH -> umi.com = Existing
8	0.874	f Type = asp -> umi.com = Existing
9	0.874	f Type = pdf -> umi.com = Existing

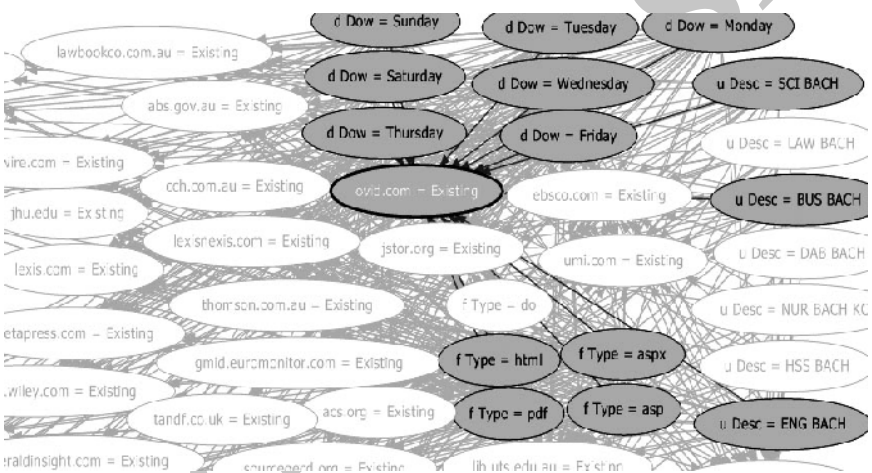
برای مثال، قانون ۷ نشان می‌دهد رشتهٔ حقوق در مقطع لیسانس با احتمال ۸۷٪ از پایگاه umi.com استفاده می کند. همچنین قانون شماره ۹ نشان می‌دهد مشتریان از این پایگاه با احتمال ۸۷٪ فایلهایی از نوع pdf دانلود می کنند.

پس از کشف قوانین وابستگی برای درک بهتر، ارتباطهای کشف شده در قالب شبکه وابستگی در شکل ۵ نمایش داده شده است.



شکل۵. شبکهٔ وابستگی پایگاه umi.com و رفتار مصرف مشتریان مراجعه‌کننده از رشته‌های مختلف تحصیلی به این پایگاه در روزهای هفته و وضعیت ناوبری آنها در سایت

در شبکهٔ فوق، همان‌طور که نمایش داده شده است، تمامی رشته‌ها در مقطع لیسانس از پایگاه‌های موجود در Umi.com در طول روزهای هفته استفاده کرده و اطلاعات علمی را با پسوند pdf دانلود می‌کنند. مشتریان در این مقطع پیمایش نیز داشته‌اند. دو پایگاه lexisnexis.comو ebsco.com نیز وضعیتی مشابه به umi.com دارند و از پایگاه‌های پر مصرف در مقطع لیسانس می باشند که در طول روزهای هفته توسط کلیهٔ رشته‌ها در این وب‌سایت مورد استفاده قرار می‌گیرند. پایگاه‌هایی مانند abs.gov.au و ovid.com وجود دارند که تنها در بعضی رشته‌ها مورد استفاده قرار می‌گیرند و در زیر شبکه وابستگی مربوط به ovid.com و نحوهٔ ناوبری کاربرانش آورده شده است:



شکل ٦. شبکهٔ وابستگی پایگاه ovid.com و رفتار مصرف مشتریان مراجعه‌کننده در مقطع لیسانس از رشته‌های مختلف تحصیلی به این پایگاه در روزهای هفته و وضعیت ناوبری آنها در سایت

در مجموع، تمامی رشته‌ها در مقطع لیسانس از پایگاه‌های umi.com, lexisnexi.com , ebsco.com مکرراً در طول روزهای هفته استفاده و اطلاعات علمی نیز دانلود می‌کنند. در این مقطع، مشتریان تمرکز روی روز خاصی در هفته ندارند و در طول روزهای هفته مراجعه داشته‌اند.

مدل داده‌کاوی شماره ۳

در این مدل، رفتارهای اطلاع‌یابی مشتریان مقیم در کشورهای مختلف در مقطع فوق لیسانس پیش‌بینی شده است. نتایج حاصل از اجرای الگوریتم، کشف قوانین وابستگی در این مدل است.

جدول۵. بخشی از مجموعه‌های تکرارپذیر کشف شده پس از اجرای الگوریتم Association Rules با minimum support=7

Row	Support	Size	ItemSet
1	211	1	umi.com = Existing
2	208	2	factiva.com = Existing, umi.com = Existing
3	198	2	ebsco.com = Existing, umi.com = Existing
4	195	3	ebsco.com = Existing, factiva.com = Existing, umi.com = Existing
5	195	2	lexisnexis.com = Existing, umi.com = Existing

ردیف شماره ۵ نشان می‌دهد در این مقطع، بازدید از پایگاه‌های موجود در lexisnexis.comو umi.com در تراکنشهای کاربران به تکرار با هم رخ داده است.

مجموعه قوانین کشف شده

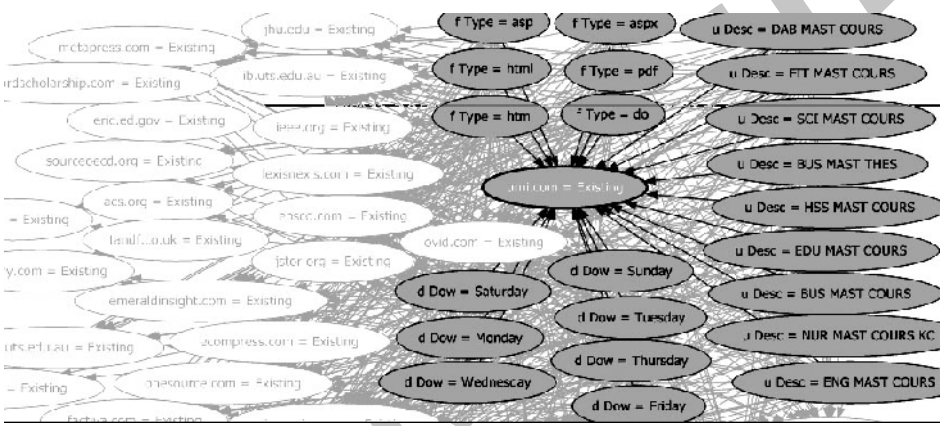
از مجموعه‌های تکرارپذیر، رابطه‌های وابستگی جستجو و قوانین وابستگی کشف شد. بخشی از این قوانین، در جدول ۶ نشان داده شده است.

جدول٦. قوانین کشف شده پس از تعیین مجموعه‌های تکرارپذیر ارائه شده در مرحلهٔ قبل با minimum probability=0.4

Row	Probability	Rule
1	0.909	d Dow = Saturday, f Type = html -> umi.com = Existing
2	0.909	u Desc = BUS MAST THES -> umi.com = Existing
3	0.889	f Type = pdf, u Desc = BUS MAST COURS ->umi.com =Existing
4	0.889	d Dow = Sunday, u Desc = BUS MAST COURS -> umi.com = Existing
5	0.889	f Type = pdf -> umi.com = Existing
6	0.889	f Type = do -> umi.com = Existing
7	0.874	f Type = html, u Desc = BUS MAST COURS -> umi.com = Existing
8	0.856	u Desc = NUR MAST COURS KC -> umi.com = Existing
9	0.856	u Desc = BUS MAST COURS -> umi.com = Existing

برای مثال، قانون۳ نشان می‌دهد مشتریانی که در رشتهٔ تجارت (BUS Master Cours) در مقطع فوق لیسانس بوده و فایلهای نوع pdf دانلود می‌کنند، با احتمال ۸۸٪ از پایگاه umi.com استفاده می‌کنند. همچنین، قانون شماره ۴ نشان می‌دهد مشتریانی که در رشتهٔ تجارت (BUS Master Cours) در مقطع فوق لیسانس هستند و در روزهای یکشنبه وارد وب‌سایت می‌شوند، با احتمال ۸۸٪ این پایگاه را ملاقات می‌کنند.

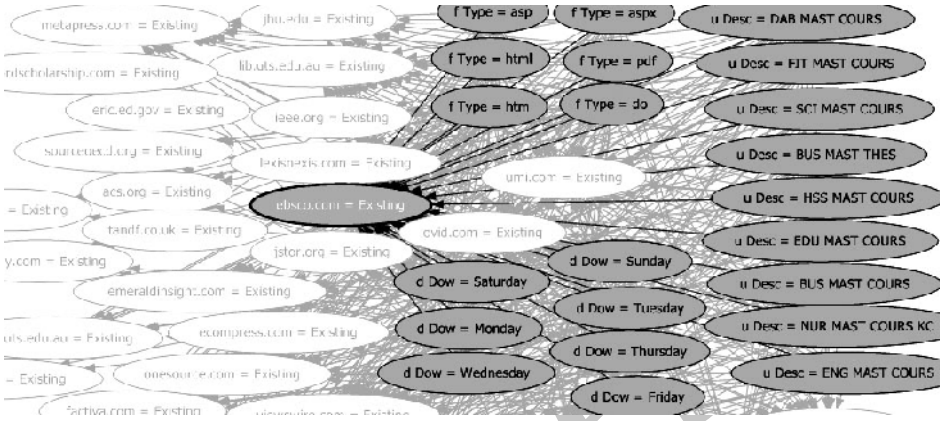
پس از کشف قوانین وابستگی برای درک بهتر، ارتباطهای کشف شده در قالب شبکهٔ وابستگی، در شکل ۷ نمایش داده شده است:



شکل۷. شبکهٔ وابستگی پایگاه umi.com و رفتار مصرف مشتریان مراجعه‌کننده در مقطع فوق لیسانس از رشته‌های مختلف تحصیلی به این پایگاه در روزهای هفته و وضعیت ناوبری آنها در سایت

در شبکهٔ فوق، همان‌طور که نمایش داده شده است، تمامی رشته‌ها در مقطع فوق لیسانس از پایگاه‌های موجود در Umi.com در طول روزهای هفته استفاده و اطلاعات علمی را با پسوند pdf دانلود می‌کنند. همچنین، در این پایگاه پیمایش نیز انجام داده‌اند. از نتایج دیگر از کشف وابستگی‌ها، پایگاه‌هایی است که در این مقطع به تکرار توسط کاربران ملاقات می‌شوند. این پایگاه‌ها ebsco.com، lexisnexis.com, saiglobal.com ,viewSwier.com ,newsbank.com springerlink.com

,galegroup.com ,netlibrary.com, jstor.org, factiva.com
intersicence.wiley.com, می‌باشند و وضعیتی شبیه به umi.com دارند.



شکل۸. شبکهٔ وابستگی پایگاه ebsco.com و رفتار مصرف مشتریان مراجعه‌کننده در مقطع فوق لیسانس از رشته‌های مختلف تحصیلی به این پایگاه در روزهای هفته و وضعیت ناوبری آنها در سایت

پایگاه ieee.org توسط سه رشته در این مقطع در روزهای دوشنبه، سه‌شنبه، پنج‌شنبه و جمعه مکرراً استفاده شده است. در این پایگاه، عمدتاً عمل ناوبری انجام شده است.

در کل، تمامی رشته‌ها در مقطع فوق لیسانس از پایگاه‌های شامل ebsco.com ، lexisnexis.com, springerlink.com ,newsbank.com ,viewSwier.com ,saiglobal.com ,galegroup.com ,netlibrary.com, jstor.org, factiva.com intersicence.wiley.com, مکرراً در طول روزهای هفته استفاده و اطلاعات علمی نیز از آنها دانلود می‌کنند. در این مقطع، مشتریان تمرکز روی روز خاصی در هفته ندارند و در طول روزهای هفته مراجعه داشته‌اند. همچنین، تعداد زیادی از پایگاه‌های علمی به صورت مشترک بین رشته‌های مختلف در کل روزهای هفته توسط مشتریان استفاده می‌شود.

مدل داده‌کاوی شماره ٤

در این مدل، رفتارهای اطلاع‌یابی مشتریان مقیم در کشورهای مختلف در مقطع دکتری پیش‌بینی شده است. در جدول ۷، بخشی از این عناصر آورده شده است.

جدول۷. بخشی از مجموعه‌های تکرارپذیر پس از اجرای الگوریتم Association Rules با minimum support=1

Row	Support	Size	ItemSet
1	183	1	umi.com = Existing
2	156	2	springerlink.com = Existing, umi.com = Existing
3	112	2	interscience.wiley.com = Existing, umi.com = Existing
4	107	2	lexisnexis.com = Existing, umi.com = Existing
5	104	2	galegroup.com = Existing, umi.com = Existing

برای مثال، در ردیف ۵ عناصر تکرار شده نشان می‌دهد که پایگاه‌های galegroup.com و umi.com توسط کاربران این مقطع در یک تراکنش بارها رخداد داشته‌اند.

قوانین کشف شده

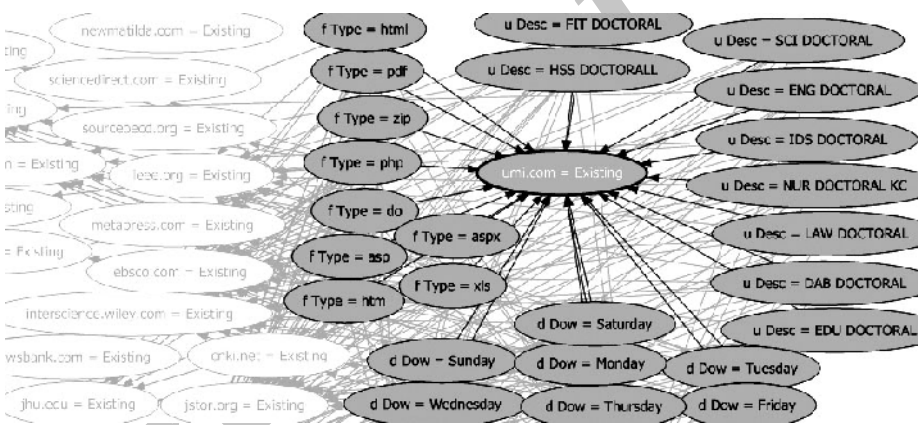
پس از کشف مجموعه‌های تکرار شونده با استفاده از قانون تکرارپذیری در تراکنشهای اجرا شده توسط کاربران، قوانین وابستگی کشف شد. بخشی از این قوانین در جدول ۸ آورده شده است.

جدول۸. قوانین کشف شده پس از تعیین مجموعه‌های تکرارپذیر ارائه شده در مرحله قبل با Minimum Probability=0.4

Row	Probability	Rule
1	0.957	u Desc = IDS DOCTORAL -> umi.com = Existing
2	0.95	u Desc = ENG DOCTORAL -> umi.com = Existing
3	0.947	f Type = html -> umi.com = Existing
4	0.946	u Desc = SCI DOCTORAL -> umi.com = Existing
5	0.933	d Dow = Tuesday -> umi.com = Existing
6	0.933	d Dow = Monday -> umi.com = Existing
7	0.923	u Desc = SCI DOCTORAL, f Type = pdf -> umi.com = Existing
8	0.9	d Dow = Monday, f Type = pdf -> umi.com = Existing
9	0.889	d Dow = Monday, u Desc = HSS DOCTORALL -> umi.com = Existing

برای مثال، قانون شماره ۷ نشان می‌دهد مشتریان در رشتهٔ SCI در مقطع دکتری که فایل از نوع pdf دانلود می‌کنند، با احتمال ۹۲٪ از پایگاه umi.com استفاده می‌کنند. همچنین، قانون ۸ نشان می‌دهد مشتریانی که در مقطع دکتری فایلهای نوع pdf دانلود می‌کنند و در روزهای دوشنبه مراجعه داشته‌اند، با احتمال ۹۰٪ از پایگاه umi.com استفاده می‌کنند. قانون شماره ۹ نشان می‌دهد مشتریانی که در رشته علوم انسانی در مقطع دکتری هستند و در روزهای دوشنبه وارد وب‌سایت می‌شوند، با احتمال ۸۸٪ از این پایگاه بازدید می‌کنند.

پس از کشف قوانین وابستگی برای درک بهتر، ارتباطهای کشف شده در قالب شبکه وابستگی نشان داده شده است. در شکل ۹ پایگاه umi.com و رفتار کاربران استفاده‌کننده در طول روزهای هفته نشان داده شده است:



شکل ۹. شبکهٔ وابستگی پایگاه umi.com و رفتار مصرف مشتریان مراجعه‌کننده در مقطع دکتری از رشته‌های مختلف تحصیلی به این پایگاه در روزهای هفته و وضعیت ناوبری آنها در سایت

در شبکهٔ فوق، همان‌طور که نمایش داده شده است، تمامی رشته‌ها در مقطع دکتری از پایگاه‌های موجود در Umi.com در طول روزهای هفته استفاده و اطلاعات علمی با پسوند pdf ، zip، xls، دانلود می‌کنند. همچنین، در این پایگاه پیمایش نیز انجام

داده‌اند. همچنین، شبکهٔ وابستگی مربوط به دیگر پایگاه ebsco.com و رفتار حرکت ملاقات‌کننده نشان داده شده است.

در کل، تمامی رشته‌ها در مقطع دکتری از پایگاه‌های:

ebsco.com, lexisnexis.com, springerlink.com galegroup.com, jstor.org, factiva.com, intersicence.wiley.com

مکرراً در طول روزهای هفته استفاده و از این پایگاه‌ها اطلاعات علمی نیز دانلود می‌کنند. در این مقطع، مشتریان تمرکز روی روز خاصی در هفته ندارند و در طول روزهای هفته مراجعه داشته‌اند. تعداد پایگاه‌هایی که در این مقطع توسط مشتریان مورد استفاده قرار می‌گیرد، نسبت به دو مقطع لیسانس و فوق لیسانس کمتر است. در این مقطع، نسبت به دو مقطع دیگر، مشتریان انواع متنوع‌تری از اطلاعات علمی را استفاده کرده‌اند. چنانکه در شکل نشان داده شده است، مشتریان، وب‌سایتهای متنوعی را ناوبری کرده‌اند اما عمدتاً هیچ دانلود اطلاعات علمی در طول ناوبری نداشتند. این می تواند دلیلی بر ضعیف بودن پایگاه‌ها در این مقطع باشد که نیاز این دسته از مشتریان را پاسخگو نبوده است.

مدل کاوش شماره۵

این مدل، داده‌کاوی با توجه به سه مقطع تحصیلی ذکر شده، پیش‌بینی می‌کند که مشتریان در مراجعات خود عموماً از چه پایگاه‌هایی با هم استفاده می‌کنند.

مجموعه اقلام تکرارپذیر کشف شده

در این مدل minimum_support=۱۸ مقدار پیشنهادی الگوریتم توسط نرم‌افزار می‌باشد و تعیین‌کننده حداقل مقدار برای قابل قبول بودن اقلام وابسته است. برای مثال، ردیف شماره ۱ بیان می‌کند ۴۰ مرتبه پایگاه‌های galegroup.com و umi.com در یک تراکنش ملاقات کاربران، رخداد همزمان داشته‌اند. در زیر، چند نمونه از خروجی حاصل در این مرحله نشان داده شده است:

جدول۹. بخشی از مجموعه‌های تکرارپذیر کشف شده پس از اجرای الگوریتم Association Rules

Row	Support	Size	Itemset
1	40	2	galegroup.com = Existing, umi.com = Existing
2	37	2	springerlink.com = Existing, umi.com = Existing
3	36	3	springerlink.com = Existing, galegroup.com = Existing, umi.com = Existing
4	36	2	interscience.wiley.com = Existing, umi.com = Existing
5	35	3	interscience.wiley.com = Existing, galegroup.com = Existing, umi.com = Existing
6	35	2	Ebsco.com = Existing, umi.com = Existing

قوانین کشف شده

در این مرحله، به کشف روابط وابستگی بین اقلام هر مجموعه پرداخته شد. قوانین وابستگی در واقع رابطهٔ موجود بین اقلام را با توجه به قوانینی که قبلاً اشاره شد، کشف و برای هر قانون مقدار عددی Probability که تعیین‌کننده احتمال رخداد قانون است، ارائه شده است. در زیر، بخشی از قوانین حاصل از مجموعه‌های تکرارپذیر آمده است.

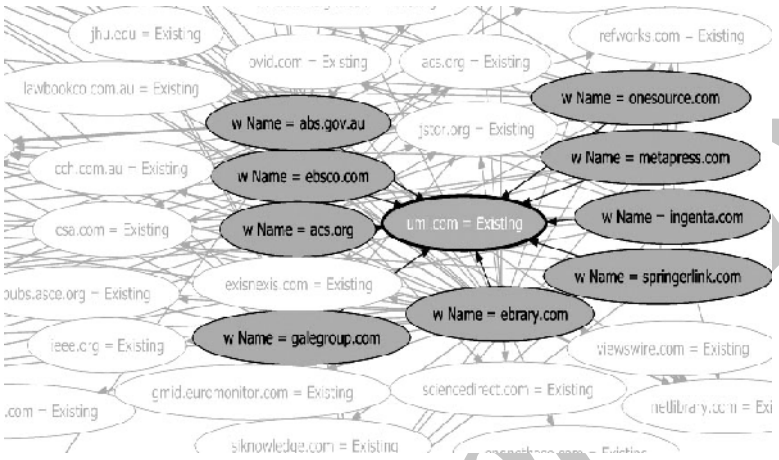
جدول۱۰. قوانین کشف شده پس از کشف مجموعه‌های تکرارپذیر با Minimum probability=0.4

Row	Probability	Rule
1	0.874	w Name = acs.org -> umi.com = Existing
2	0.865	w Name = metapress.com -> umi.com = Existing
3	0.865	w Name = springerlink.com -> umi.com = Existing
4	0.85	w Name = ingenta.com -> umi.com = Existing
5	0.789	w Name = ebsco.com -> umi.com = Existing
6	0.756	w Name = abs.gov.au -> umi.com = Existing

برای مثال، قانون شماره ۲ مطرح می‌کند مشتریانی که از پایگاه‌های موجود در metapress.com استفاده می‌کنند، با احتمال ۸۶٪ به پایگاه‌های موجود در umi.com نیز مراجعه داشته‌اند. همچنین، در قانون شماره ۵، مشتریانی که از پایگاه‌های موجود در ebsco.com استفاده می‌کنند، با احتمال ۷۹٪ به پایگاه‌های موجود در umi.com مراجعه داشته‌اند.

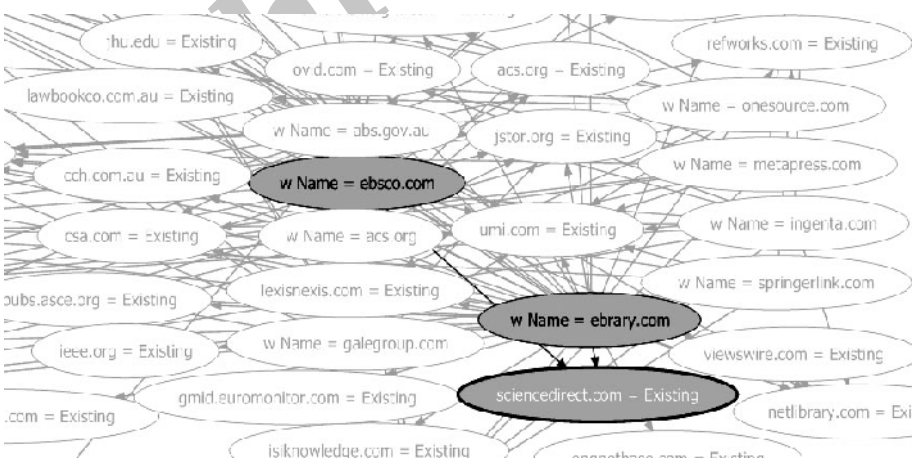
پس از کشف قوانین وابستگی برای درک بهتر، ارتباطهای کشف شده در قالب شبکه وابستگی نمایش داده شده است. در شبکه نشان داده شده مشتریانی که از

پایگـــــاه‌هـــــای abs.gov.au، ebrary.com، Ebsco.com، metapress.com، springerlink.com، unsource.com و galegroup.com استفاده کرده‌اند، به پایگاه umi.com نیز مراجعه داشته‌اند.



شکل ۱۰. شبکهٔ وابستگی مربوط به پایگاه‌هایی که در کنار پایگاه umi.com ملاقات شده‌اند

همچنین، در شبکهٔ وابستگی زیر نشان داده شده است کاربرانی که از پایگاه‌های موجود در ebsco.com و ebrary.com دیدن کرده‌اند، به پایگاه‌های موجود در sciencedirect.com نیز مراجعه کرده‌اند:



شکل ۱۱. بخشی از شبکهٔ وابستگی مربوط به پایگاه‌هایی که در کنار پایگاه sciencedirect.com ملاقات شده‌اند

مدل کاوش شماره٦

در این مدل، با توجه به سه مقطع تحصیلی ذکر شده، پیش‌بینی می‌کند که مشتریان مقیم کشورهای مختلف چه کالاهای علمی را با هم استفاده می کنند.

مجموعه اقلام تکرارپذیر کشف شده مدل کاوش ٦ پس از اجرای الگوریتم

در این مدل، با توجه به minimum support=1 ، مقدار پیشنهادی الگوریتم، تعدادی از مجموعه‌های ۱ و ۲ و ۳ عنصری کشف شدند که بخشی از آنها در جدول ۳ نمایش داده شده است:

جدول۱۱. بخشی از مجموعه‌های تکرارپذیر مدل۲ پس از اجرای الگوریتم Association Rules

Row	Support	Size	Itemset
1	1	2	y662p1l0r8x65235.pdf = Existing, x61m545652q08048.pdf = Existing
2	1	2	y044m8w3571u4j15.pdf = Existing, x61m545652q08048.pdf = Existing
3	1	2	xnn5yvarbuxrffng.pdf = Existing, x61m545652q08048.pdf = Existing
4	1	2	x83n556l41736q78.pdf = Existing, x61m545652q08048.pdf = Existing
5	1	2	x61m545652q08048.pdf = Existing, x312wbfbxe169wad.pdf = Existing
6	1	2	x61m545652q08048.pdf = Existing, x2363l28387g8131.pdf = Existing

قوانین کشف شده مدل کاوش ٦

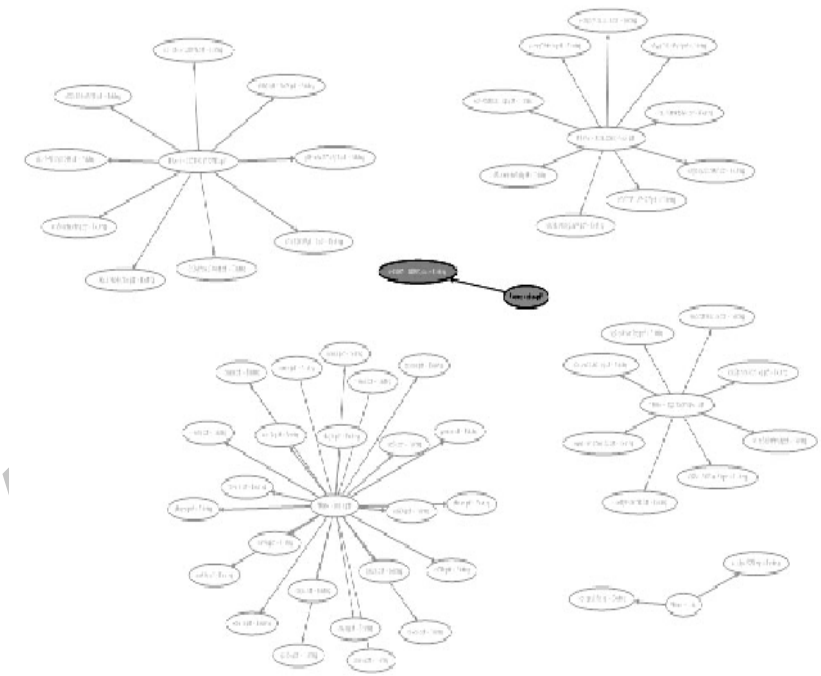
این مجموعه‌ها بر اساس قوانین وابستگی و با استفاده از قانون تکرارپذیری در تراکنشهای اجرا شده توسط مشتریان، کشف و سپس به دنبال کشف روابط وابستگی بین این مجموعه‌ها و اقلام هر مجموعه جستجو کرده و قوانین وابستگی توسط این مدل ارائه شد.

جدول۱۲. قوانین کشف شده پس از تعیین مجموعه‌های تکرارپذیر با minimum probability=0.4

Row	Probability	Rule
1	0.852	f Name = ct-us.pdf -> s-63697-11602827.doc = Existing
2	0.832	f Name = adajia.pdf -> zfa6xa.pdf = Existing
3	0.80	f Name = 0673546165327426.pdf -> x61m545652q08048.pdf = Existing

Row	Probability	Rule
4	0.793	f Name = adajia.pdf -> zdaw1a.pdf = Existing
5	0.788	f Name = adajia.pdf -> zdalra.pdf = Existing
6	0.788	f Name = adajia.pdf -> zcawoa.pdf = Existing

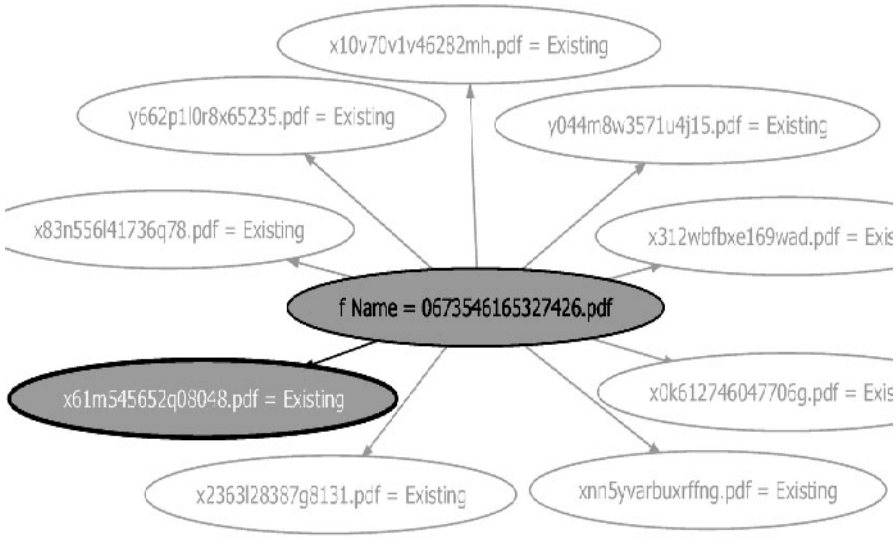
برای مثال، قانون ۳ بیان می‌کند ۸۰٪ مشتریانی که کالای علمی 673546165327426.pdf را دانلود[1] کرده‌اند، کالای علمی x61m545652q08048.pdf را نیز دانلود کرده‌اند. قانون ۴ بیان می‌کند مشتریانی که کالای علمی adajia.pdf را دانلود کرده‌اند، کالای علمی zdaw1a.pdf را نیز دانلود کرده‌اند. پس از کشف قوانین وابستگی برای درک بهتر، ارتباطهای کشف شده در قالب شبکهٔ وابستگی در شکل ۱۲ نمایش داده شده است:



شکل ۱۲. نمای کامل از شبکهٔ وابستگی مدل ۲

1. Download.

در شکل زیر، یکی از ارتباطهای کشف شده به صورت واضح نشان داده شده است. چنانکه در شکل مشاهده می‌شود، وابستگی بین دو فایل اطلاعاتی pdf ۰۶۷۳۵۴۶۱۶۵۳۲۷۴۶. و x61m545652q08048.pdf وجود دارد.



شکل۱۳. شبکه وابستگی مربوط به بخشی از قوانین مدل۲

بحث

با بررسی دقیق «رفتار کاربران در استفاده از پایگاه‌ها و صفحات پر استفاده توسط آنها» مسیرهای پر استفاده در پایگاه‌های مختلف کشف شد که می‌تواند یک منبع غنی به منظور بهبود طراحی سایت بوده و همچنین در بعضی موارد در تصمیم گیریهای اساسی از آنها استفاده نمود. از مسیرهای پراستفاده، صفحه راهنمای سایت csa.com می‌باشد. از دلایل این امر، می‌تواند ناآشنا بودن کاربران با این سایت و پایگاه‌های موجود در آن باشد که در زمان ورود، به دلیل ناآگاهی از هدفها و یا موفق نبودن در یافتن اطلاعات مورد نیازشان در سایت دچار سردرگمی شده و درصدد رفع نیاز خود به راهنمای سایت مراجعه می‌کنند و این می‌تواند از ضعفهای طراحی سایت باشد. از معیارهای مهم در طراحی یک سایت، بالا بودن قابلیت استفادهٔ آن است که مانع از اتلاف وقت کاربران می‌شود. کاربران

تمایل دارند با ورود به سایت بتوانند به سرعت نیاز اطلاعاتی خود را برآورده کنند (Jakob Nielsen, 1990). برای رفع این مشکل، ارائه نقشه سایت، قرار دادن اطلاعاتی در مورد هدفهای سایت در صفحه اول وب‌سایت و ارائه راهنمای غنی از محتوا و چگونگی دسترسی به آنها در هدایت این دسته از کاربران می‌تواند مفید باشد.

از دیگر مسیرهای پر استفاده، پایگاه‌های موجود در مسیر rpsv/cw/cgi-bin/fastforward در sourceoecd.com می‌باشد. دلیل آن می‌تواند وجود اطلاعات مفید در این مسیر باشد که کاربران زیادی را جذب نموده است. با توجه به این قوانین، می‌توان با بررسی بیشتر اطلاعات موجود در مسیرهای پر استفاده، با دسته‌بندی کارآمدتر اقلام علمی و در نظر گرفتن سیاستهای مفیدتر در چینش اقلام و قرار دادن آنها در مسیرهای کوتاه‌تر، ترافیک شبکه را کنترل نمود و به این ترتیب کاربران در یافتن نیازهای اطلاعاتی خود به رضایت بیشتری دست می‌یابند. این امر در کارایی طراحی وب‌سایت و همچنین در امر تصمیم‌گیری به منظور تهیهٔ نیازهای بیشتر کاربران، مفید است.

مدل رفتار اطلاع‌یابی کاربران در مقطع لیسانس، حاکی از استفادهٔ مستمر پایگاه‌های موجود در Umi.com ، ebsco.com و lexisnexis.com در طول روزهای هفته است که عمدتاً اطلاعات علمی با پسوند pdf دانلود کرده‌اند. این پایگاه‌ها نیازهای کاربران زیادی را برطرف می‌کند و این امر می‌تواند به تامین‌کنندگان اصلی اطلاعات علمی این وب‌سایت در اتخاذ تصمیمهای مهم تجاری یاری رساند. همچنین، پایگاه‌هایی مانند abs.gov.au و ovid.com نیز وجود دارند که تنها در بعضی رشته‌ها مورد استفاده قرار می‌گیرند. در مقطع فوق لیسانس، پایگاه‌های موجود در Umi.com ، ebsco.com، lexisnexis.com, springerlink.com ,newsbank.com ,viewSwier.com ,saiglobal.com ,galegroup.com ,netlibrary.com, jstor.org, factiva.com intersicence.wiley.com, از جمله موارد پر استفاده توسط کاربران این مقطع است که تعداد بیشتری از پایگاه‌ها را پوشش داده و این امر به دلیل نیاز کاربران این مقطع به پژوهش بیشتر می‌باشد. در مقطع دکتری، در تمامی رشته‌ها از پایگاه‌های موجود در Umi.com، ebsco.com، springerlink.com ،lexisnexis.com,

galegroup.com , jstor.org, factiva.com ,intersicence.wiley.com در طول روزهای هفته استفاده شده و اطلاعات علمی با پسوند pdf ، zip، xls، دانلود می‌شود. تعداد پایگاه‌های استفاده شده در این مقطع توسط کاربران با توجه به دو مقطع لیسانس و فوق لیسانس کمتر بوده و این امر می‌تواند به دلیل نیاز گستردهٔ کاربران این مقطع به کاوش و بررسی‌های بیشتر و نبود اطلاعات لازم و مورد نیاز این مقطع در پایگاه‌های موجود باشد.

برخی پایگاه‌ها در کنار یکدیگر مورد استفاده قرار می‌گیرند؛ بدین معنا که کاربران در طول ملاقات خود در وب‌سایت، به چندین پایگاه در طول اتصالشان رجوع داشته‌اند. برای نمونه، کاربرانی که از پایگاه‌های metapress.com، Ebsco.com، ebrary.com، abs.gov.au، springerlink.com، unsource.com و galegroup.com استفاده کرده‌اند، به پایگاه umi.com نیز مراجعه داشته‌اند. همچنین، کاربرانی که به پایگاه‌های موجود در ebsco.com و ebrary.com مراجعه داشته‌اند، به پایگاه‌های موجود در sciencedirect.com نیز رجوع کرده‌اند. بنابراین، این پایگاه‌ها دارای اطلاعاتی هستند که می‌تواند در کنار یکدیگر میزان بیشتری از نیازهای کاربران را برطرف کند و هر کدام به تنهایی نمی‌توانند پاسخگوی تمام نیازهای اطلاعاتی آنها باشند. از طرفی، بررسی این امر می‌تواند در ارائهٔ پیشنهادهای بهینه به کاربران و قرار دادن پایگاه‌هایی که عموماً در کنار یکدیگر به تکرار استفاده می‌شوند در یک مکان، به بازیابی اطلاعات سرعت بخشد.

در این وب‌سایت، بسیاری از کالاهای اطلاعات علمی به تکرار با هم استفاده شده‌اند. برای مثال، کاربرانی که کالای علمی 673546165327426.pdf را دانلود کرده‌اند، کالای علمی x61m545652q08048.pdf را نیز دانلود نموده‌اند. همچنین، کالای علمی adajia.pdf در کنار کالای علمی zdaw1a.pdf به تکرار با هم دانلود شده‌اند. این امر نشان می‌دهد این کالاهای علمی دارای وابستگی اطلاعاتی هستند و این وابستگی در سیاست چینش آنها در کنار هم بسیار مهم است. کشف وابستگی‌های کالاها می‌تواند در سرعت بخشیدن به بازیابی اطلاعات و کاهش ترافیک سایت، تأثیر زیادی داشته باشد.

نتیجه‌گیری

وب جهان‌گستر، یک منبع داده‌کاوی غنی است و به یک مدیریت توانا نیاز دارد تا با استفاده از تکنیکهای داده‌کاوی، دانش و اطلاعات موجود در داده‌های پیشین را کشف کند و بر اساس آن پیش‌بینی‌هایی به منظور تصمیم‌گیریهای مهم در مورد پایگاه‌های اطلاعاتی کارآمد و همچنین سیاستهای طراحی وب‌سایت انجام دهد. در نمونهٔ مورد مطالعه، پایگاه‌های اطلاعاتی پراستفاده توسط کاربران در مقاطع مختلف شناسایی شد. از جمله پایگاه‌های پرمصرف در هر سه مقطع لیسانس و فوق لیسانس و دکتری، پایگاه‌های موجود در umi.com و factiva.com و newbanks.com و lexisnexis.com و ebsco.com هستند. استفادهٔ مکرر از پایگاه‌های موجود در umi.com نشان‌دهندهٔ وجود اطلاعات علمی مناسب برای کاربران در سطوح مختلف تحصیلی است که توجه عمدهٔ کاربران را به خود جلب کرده است.

پایگاه‌های اطلاعات علمی ارائه شده در مقطع فوق لیسانس، نسبت به دو مقطع دیگر سطح مراجعه بالاتری دارند. در مقاطع تحصیلی لیسانس و فوق لیسانس، بیشترین استفادهٔ اطلاعات علمی از نوع pdf بوده و در مقطع دکتری از انواع اطلاعات علمی مانند pdf ، xip ، xls و swf استفاده شده که تنوع بیشتری دارند. زمانهای استفاده از پایگاه‌های اطلاعات علمی توسط هر سه مقطع، اکثر روزهای هفته است.

از دیگر نتایج قابل توجه در این پژوهش، کشف مسیرهای پراستفاده توسط کاربران و وجود مشکلات بازیابی اطلاعات در بعضی پایگاه‌ها مانند csa.com و همچنین شناسایی گلوگاه‌هاست. بعضی از این مسیرهای پر ترافیک به مسیرهای مربوط به اطلاعات علمی پرمراجعه مربوط است که عمدتاً در مسیرهای طولانی قرار گرفته‌اند و این می‌تواند دلیلی بر نبود دسته‌بندی مناسب اطلاعات در پایگاه‌ها باشد.

با بررسی قوانین وابستگی روی پایگاه‌های استفاده شده توسط کاربران مختلف، وابستگی‌های موجود بین این پایگاه‌ها کشف شد. مجموعه‌ای از پایگاه‌ها مکرراً با هم توسط کاربران مختلف استفاده شد و کاربران در مراجعات خود به تکرار این کالاهای علمی را در کنار هم درخواست نموده‌اند. این امر، وابستگی موجود بین پایگاه‌ها را نشان

می‌دهد. رعایت چینش پایگاه‌ها با توجه به وجود وابستگی اطلاعاتی آنها، می‌تواند تأثیر عمیقی را به همراه داشته باشد.

منابع

- پاتکار، ویوک.ان. (۱۳۸۰). «کاربردهای داده‌کاوی در کتابخانه‌ها و مؤسسات دانشگاهی». ترجمهٔ مریم صراف‌زاده و افسانه حاضری. شماره سوم دوره پنجم. مجله الکترونیکی پژوهشگاه اطلاعات و مدارک علمی ایران [mmsarraf@yahoo.com]

- Cooley, Robert; Mobasher, Bamshad; Srivastava, Jaideep (1999). "Data Preparation for Mining World Wide Web Browsing Patterns". Department of computer Science and Engineering University of Minnesota. Knowledge and Information Systems, maya.cs.depaul.edu. Available on [www.google. com]

- Frawley, William J.; G. Piatetsky-Shapiro and C. Matheus (1992). "Knowledge discovery in database, ed" G. Piatetsky-Shapiro and w. Frawley, Menlo Park, CA:AAAIPress.

- Fayyad, U. , Piatetsky-Shapiro, G. , Smyth, P (1996). "From Data Mining to Knowledge Discovery in Database". American Association for Intelligence, California: AAAI Press. [aaai.org]. Fall.

- Quinlan, Ross (1992). "C4.5: Programs for Machine Learning". Morgan Kaufmann Publishers. San Mateo (1-25).

- Russell, Michael Randy (1998). "World Wide Web Site Visitor Studies Techniques Using Server Log File Data". A dissertation Submitted to Michigan State University in partial fulfillment of the requirement for the degree of Doctor of Philosophy. UMI Number: 9922370. Available on [www.Proquest.com]

- Subramaniam, Sharmila (2006). "Optimizing Business Processes through Log Analysis". University of California Riverside. Dissertation of Philosophy in Computer Science. June.

- Yang, Zhijian (2005). "Web Log Analysis: Experimental Studies". Florida Atlantic University. A thesis for Degree of Master of Science. UMI Number: 1425339. Available on: [www.proquest.com].

- Zaiane, Osmar (2001). "Web Usage Mining for a Better Web-Based Learning Environment". Conference on Advanced Technology, University of Alberta, Canada-cs.ualberta.ca.
[email: zaianecs.ualberta.ca].

- Zhang, Sen (2005). "Pattern Discovery In Structural Databases With Applications to Bioinformatics". A Dissertation submitted to the faculty of New Jersey Institute of Technology in Partial fulfillment of the Requirements for the Degree of Doctor of Philosophy in computer science. UMI: 3186460. Available on [www.proquest.com].

- Breeding, Marshall (2005). "Analyzing Web Server Logs to Improve a Site's Usage". Computers in Libraries, October.

- Huntington, Paul; David Nicholas and Hamid R. Jamali (2007). "The information seeking behaviour of the users of digital scholarly journals". Journal of Information Science OnlineFirst, Published on April 10 as doi: 10,1177/0165551506077407.

- Nicholas, David. Paul Huntington and Anthony Watkinson (2005). "Scholarly journal usage: the results of deep log analysis". Journal of Documentation, Vol. 61, No. 2, PP. 248-280. Emerald Group Publishing Limited 0022-0418. DOI 10, 1108/00220410510585214.

- Nicholas, David. Paul Huntington, Hamid R. Jamali and Carol Tenopir (2006). "Finding Information in (Very Large) Digital Libraries: A Deep Log Approach to Determining Differences in Use According to Method of Access". The Journal of Academic Librarianship, Volume 32, Number 2, PP.: 119–126, Available online February.