**نام مقاله: كشف مسير حركت كاربران اطلاعات الكترونيكي با استفاده از الگوريتم قوانين وابستگي در داده كاوي: مطالعه موردي وب سايت كتابخانه دانشگاه يوتي اس، استراليا**

**نام نشريه: فصلنامه كتابداري و اطلاع رساني (اين نشريه در www.isc.gov.ir نمايه مي شود)**

**شماره نشريه: 49 \_ شماره اول، جلد 13**

**پديدآور: دكتر زهير حياتي، مرجان صادقي مجرد، نيما جعفري**

**چكيده**

**هدف اصلي اين تحقيق، جستجوي روشهايي براي مطالعه رفتار كاربران در ارتباط با هدفهاي آموزشي آنها در يك وب‌سايت مشخص است. در حال حاضر، داده‌كاوي، مهم‌ترين فناوري براي بهره‌برداري مؤثر، صحيح و سريع از داده‌هاي حجيم است. موضوع داده‌كاوي، شناخت دانش جديد و مفيد، رابطه‌هاي منطقي و الگوهاي موجود در داده‌هاست و پل ارتباطي بين علم آمار، رايانه، هوش مصنوعي، الگوشناسي، فراگيري ماشين و بازنمائي بصري داده‌ها مي‌باشد. پژوهش حاضر با استفاده از تكنيك داده‌كاوي و بهره‌گيري از الگوريتم «قوانين وابستگي» روي داده‌هاي جمع‌آوري شده در قالب فايل ثبت وقايع وب‌سايت كتابخانه دانشگاه UTS استراليا، به كشف الگوي مسير حركت كاربران در سايت پرداخته است. نتايج حاصل از اين پژوهش، بينش وسيعي از رفتار كاربران و عملكرد آنها در وب‌سايت را در اختيار مديران و طراحان آن كتابخانه قرار مي‌دهد.**

**كليدواژه‌ها: داده كاوي، قوانين وابستگي، تجارت الكترونيكي، كاوش كاربردي وب، وب‌سايت، كتابخانه، دانشگاه يو تي اس، استراليا**

**مقدمه**

داده كاوي در سالهاي اخير، به دليل در دسترس بودن حجم انبوهي از داده‌ها، توجه بسيار زيادي را در جوامع علمي و صنعت اطلاعات، به خود جلب كرده است و به عنوان يكي از پيشرفتهاي اخير در راستاي فناوريهاي مديريت داده‌ها به شمار مي‌رود. فناوري بر پايه وب، به دليل فراهم نمودن امكانات مفيد از جمله در دسترس بودن منابع، سادگي گسترش و به روز كردن و نگهداري آنها روي وب، به عنوان يك فناوري مناسب معرفي شده است و در بسياري از محيطهاي آموزشي توسعه‌يافته تحت وب در سرتاسر دنيا در حال استفاده از آن هستند. اگرچه ابزارهاي هوشمندي براي درك رفتارهاي كاربران برخط به منظور افزايش فروش و سود، توسعه يافته است، اما كارهاي اندكي بر روي كشف و دسترسي به الگوهاي كاربران برخط براي درك رفتارهاي آموزشي آنها صورت گرفته است. مربياني كه از ابزارها و محيطهاي الكترونيكي براي آموزش استفاده مي‌كنند، به منظور ارزيابي فعاليتها و تمايز بين رفتارهاي مختلف يادگيرنده‌هاي برخط با مشكلاتي مواجهند (زيان[[1]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn1" \o ")، 2001).

با افزايش محبوبيت شبكه جهاني وب، مقدار حجيمي از داده‌ها توسط وب سرورها در قالب فايلهاي ثبت وقايع وب[[2]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn2" \o ") جمع‌آوري مي‌شوند. اين فايلها كه در آنها تمامي فعاليتهاي و رخ‌داده در سيستم وب‌سرور ثبت مي‌شود، مي‌توانند به عنوان منابع بسيار غني از اطلاعات براي درك و تشخيص رفتار كاربران وب، استفاده شوند. كاوش كاربردي وب[[3]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn3" \o ") كه آن را كاوش فايل ثبت وقايع در وب[[4]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn4" \o ") نيز مي‌نامند، در واقع استفاده از الگوريتمهاي داده كاوي بر روي فايلهاي ثبت وقايع وب به منظور پيدا كردن مسير حركت و نظم موجود در الگوهاي جستجوي كاربران وب است (سن[[5]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn5" \o ")، 2005).

هدف از انجام اين پژوهش، دست يافتن به رفتار كاربران با استفاده از فناوري داده كاوي در وب‌سايت كتابخانه دانشگاه يو تي اسUTS  استراليا و كشف قوانين موجود در داده‌هاي جمع‌آوري شده در طول 7 ماه در قالب فايل ثبت وقايع است. اين قوانين مي‌توانند مديران كتابخانه و بخش فناوري اطلاعات اين دانشگاه را در تصميم‌گيريهاي مهم توسعة مجموعه اطلاعاتي و طراحي كارآمد وب‌سايت به منظور افزايش رضايت كاربران ياري دهد.

**پيشينة پژوهش**

داده‌كاوي، فرايندي است كه در آغاز دهه 90 پا به عرصه ظهور گذاشته است و با نگرشي نو به مسئله استخراج اطلاعات از پايگاه داده‌ها مي‌پردازد. در سالهاي 1989 و 1991، كارگاه‌هاي كشف دانش از پايگاه داده‌ها توسط «پياتتسكي و همكارانش» و در فاصله سالهاي 1991 تا 1994 كارگاه‌هاي فوق، توسط «فاياد و پياتتسكي» برگزار شد. به طور رسمي، اصطلاح داده‌كاوي براي اولين بار توسط «فياض» در اولين كنفرانس بين‌المللي كشف دانش و داده‌كاوي در سال 1995 مطرح شد. از سال 1995 داده‌كاوي به صورت جدي وارد مباحث آمار گرديد (فياد، پياتتسكي و اسميت، 1996). در سال 1996 اولين شمارة مجله «كشف دانش» از پايگاه داده‌ها منتشر شد.

امروزه كنفرانسهاي مختلفي در اين زمينه در سراسر دنيا برگزار مي‌شود. داده‌كاوي با همه‌گير شدن استفاده از پايگاه‌هاي داده‌اي به عنوان يك علم مطرح شده است (كوئين لن[[6]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn6" \o ")،1992). «راسل» (1998) معتقد است افزايش رشد شبكه جهاني وب، يك منبع جديد گسترده و بزرگ از اطلاعات قابل دسترس به وجود آورده است كه بسياري از وب‌سايتها تمايل دارند هدفهاي آموزشي خود را از طريق آن انجام دهند. سرعت توسعه و رشد وب، از ميزان توسعة روشهاي مطالعه كارآمد وب‌سايتها به عنوان ابزاري براي پشتيباني آموزش و يادگيري پيشي گرفته است.

هدف اصلي اين تحقيق، جستجوي روشهايي براي مطالعة رفتار كاربران در ارتباط با هدفهاي آموزشي آنها در يك وب‌سايت مشخص بود. هدف اين مطالعه، كمك و ياري به توسعه‌دهندگان وب‌سايتها به منظور انتخاب تكنيكهاي كارآمد براي ارزيابي سايت است. پژوهشگران ديگري چون «نيكولاس»، «هانتينگتون» و «جمالي» (2006) نيز به مطالعة رفتار اطلاع‌يابي كاربران با به‌كارگيري فنون داده‌كاوي وب پرداخته‌اند. آنها دريافتند كه بسياري از كاربران وب براي زمانهاي طولاني صفحات وب را مطالعه نكرده و قبل از ترك منابع وبي، تنها به بررسي اجمالي اقلام و صفحات وبي محدودي مشغول بوده‌اند. «بريدينگ» (2005) با به‌كارگيري گروه ويژه از كاربران وب‌سايتها و نرم‌افزارهاي تجزيه و تحليل وب‌لاگ‌ها، رفتار اطلاع‌يابي كابران را در سطوح عميق‌تر مطالعه كرده است. او به مطالعة رفتار اطلاع‌يابي كاربران انفرادي اكتفا نكرده، بلكه به مطالعة گروهي از آنها از طريق جلساتي كه برگزار كرده‌اند نيز پرداخته است.

«هانتينگتون، نيكولاس و جمالي» (2007) با مطالعة تراكنشهاي جستجوي وب‌لاگ‌ها بيان مي‌دارند كه ابزارهاي اندازه‌گيري كه از اين منابع كشف مي‌شود، وسايل سودمندي براي بررسي ميزان كارايي و همچنين ميزان رضايت و عدم رضايت از موتورهاي جستجو مي‌باشند. آنها دو معيار اندازه‌گيري زمان سپري شده ميان جلسات جستجو و تعداد جستجوهاي انجام شده در هر جلسه را براي مطالعة رفتار اطلاع‌يابي كاربران موتورهاي جستجو به‌كار گرفتند. مطالعة ديگري در همين سطح توسط «نيكولاس، هانتينگتون و واتكينسون» (2005) در مورد رفتار اطلاع‌يابي كاربران كتابخانه‌هاي مجله‌هاي ديجيتالي انجام گرديد. تمركز آنها بر روي كاربران پايگاه اطلاعاتي Blackwell Synergy بود و معيارهاي تعداد جلسات برگزار شده و اقلام مورد مشاهده و مورد تقاضا را براي بررسي رفتار اطلاع‌يابي اعضاي هيئت علمي مجله‌هاي ديجيتالي پايگاه مذكور به كار گرفتند. اين پژوهشگران بيان مي‌دارند چنانچه اين نوع مطالعات با مطالعات كيفي رفتار اطلاع‌يابي كاربران تكميل گردد، به نتايج بهتر و واقعي‌تري مي‌توان دست يافت.

منبع داده در اين پژوهشها، تمامي صفحات رؤيت شده توسط مشتريان سايت در يك فايلِ ثبت وقايع روي وبِ سرور بوده است. تحليل اين فايلهاي داده، به ارزيابي‌كنندگان سايت كمك مي‌كند تا نقاط اصلي مسير حركت سطوح پرترافيك در سايت را تشخيص دهند. جستجوكنندگان مي‌توانند در مورد هويت ديداركنندگان سايت، صفحات و بخشهايي را كه در يك سايت توسط ديداركنندگان ديده شده است، استخراج كنند.

كاوش كاربردي وب، به عنوان يكي از كاربردهاي تكنيك داده‌كاوي به منظور استفاده از فايلهاي ثبت وقايع براي بهبود طراحي وب‌سايتهاست (كولي[[7]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn7" \o ")، مباشر[[8]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn8" \o ") و سريواستاوا[[9]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn9" \o ")، 1999). فايلهاي ثبت وقايع وب سرورها به صورت بالقوه شامل داده‌هاي تجربي مفيدي براي بهبود كارايي وب‌سايتها هستند و منافعي را براي بعضي از كاربردها، بخصوص موارد تجاري، در بردارند. با تحليل اين فايلها مي‌توان به پيش‌بيني لينكهايي پرداخت كه در افزايش كارايي وب‌سايت تأثير مثبت دارند و براي طراحان وب‌سايت بسيار مفيدند (يانگ[[10]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn10" \o ")، 2005). به عنوان مثال، پيش‌بيني لينك‌ها مي‌تواند براي بارگذاري اسنادي كه ممكن است ديداركننده از آنها ديدن كند، در زماني كه وي در حال خواندن صفحة جاري است، تأثيري بسيار مثبت در كار وي داشته باشد. با استفاده از فايل ثبت وقايع جريان كاري مي‌توان راهكارهايي را براي حلّ مشكلات موجود در بهبود فرايندهاي كسب و كار ارائه نمود (سابرامينام[[11]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn11" \o ")، 2006).

**روش شناسي پژوهش**

در اين پژوهش، از روش تجزيه و تحليل الگوريتمها و اجراي الگوريتم قوانين وابستگي روي‌ داده‌هاي موجود در بانك اطلاعات به منظور كشف وابستگي بين اطلاعات و اقلام موجود در بانك اطلاعات و پيش‌بيني قوانين وابستگي به منظور بهبود طراحي وب‌سايت، استفاده شده است.

**مجموعه داده‌ها**

 از فايلهاي ثبت وقايع كتابخانه دانشگاه UTS استراليا به منظور كشف قوانين وابستگي در اين مجموعه اطلاعات استفاده شده است. اين اطلاعات مجموعه داده‌هاي عمومي جمع‌آوري شده حاصل از تمامي فعاليتها و وقايع مربوط به كاربران دانشجو در مقاطع كارشناسي، كارشناسي ارشد و دكتري است كه از سراسر جهان به پايگاه‌هاي اطلاعاتي موجود در كتابخانه دانشگاه UTSمراجعه كرده و سپس مجوز استفاده از اين پايگاه‌ها و اطلاعات موجود در آنها را دريافت كرده‌اند. عموماً، توليدات علمي در اين مجموعه، اطلاعاتي در قالب فايلهايي با انواع مختلف مانند Pdf ، Doc، Zip، Rtf ، exe ، Txt و PPt به كاربران عرضه شده است.

 سپس با استفاده از فناوري OLAM[[12]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn12" \o ") و به‌كارگيري الگوريتم قوانين وابستگي بر روي اطلاعات جمع‌آوري شده از مراجعات كاربران به اين مركز در مدت 7 ماه  (سپتامبر 2006،‌ژانويه ـ ژوئيه 2007) اقدام به كشف مسير و الگوي حركت كاربران مي‌شود. اطلاعات كاربران اين مركز با توجه به اطلاعات حاصل از فايل ثبت وقايع، شاملدانشكدة محل تحصيل، محل اشتغال و همچنين مكان جغرافيايي استقرار كاربران مي‌باشد.

**شناسايي قالب اطلاعات ذخيره شده**

فايلهاي ثبت وقايع، اغلب براي كاوش كاربردي وب استفاده و در سه فُرمت عمومي[[13]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn13" \o ") ، توسعه‌يافته[[14]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn14" \o ") و اختصاصي دسته‌بندي مي‌شوند. در اين پژوهش، فرمت فايل ثبت وقايع استفاده شده جهت ذخيرة اطلاعات از نوع قالب عمومي بوده و داراي فيلدهاي زير است:

 

**شكل 1. فيلدهاي موجود در فايل ثبت وقايع عمومي**

فايل ثبت وقايع در دسترس داراي حجم حدود 16 گيگا بايت بود كه پس از عمل پاكسازي، يكپارچه‌سازي و انتقال به حجم 500 مگابايت، در قالب بانك اطلاعات در محيط sqlserver 2005 كاهش يافت.در اين مجموعه، 169 وب‌سايت منحصر به فرد موجود است (وب‌سايتها دربرگيرندة بانكهاي اطلاعاتي هستند كه كاربران به آنها مراجعه مي‌كنند). همچنين، حدود 213000 آدرس اينترنتي و 200000 فايل در اين مجموعه دانلود شده است. مجموعه فايلهاي استفاده شده در اين مجموعه پس از انجام پاكسازي اطلاعات، 10 نوع فايل مي‌باشد كه شامل txt ،doc ،zip،rtf،ppt،pdf، xls، html، asp، do، php، htm، aspx هستند. پس از انجام مراحل پيش پردازش 2000000 درخواست حاصل و در بانك اطلاعات ثبت شد.

**ساختار انباره داده****[[15]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn15" \o ")**

 پس از بررسي ساختار فايل ثبت وقايع و اطلاعات موجود در آن، انباره داده ايجاد گرديد. انباره داده به عنوان يك منبع نگهداري مجموعه‌اي از داده‌هاي جمع‌آوري شده از چندين مرجع يا منبع داده است كه معمولاً ناهمگن و هدف آن ايجاد مجموعه‌اي تحت يك طرح و ساختار به صورت يكپارچه است. در پژوهش حاضر، از ساختار برف‌دانه‌اي به منظور ايجاد انباره داده استفاده شده است كه داراي يك مركز با عنوان جدول اصلي و چندين بُعد مي‌باشد و ابعاد از طريق كليد اصلي با مركز ارتباط دارند.

 

**شكل2. ساختار مدل برف‌دانه‌اي در ايجاد انباره داده**

**پيش پردازش****[[16]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn16" \o ")**

در مرحله پيش پردازش، سه عمل اصلي بر روي داده‌هاي موجود در فايل ثبت وقايع انجام مي‌شود كه شامل پاكسازي و يكپارچه‌سازي، تبديل داده‌ها و در نهايت بارگذاري در انباره داده است. در مرحلة پاكسازي، اطلاعاتي كه در انجام و اجراي مراحل داده‌كاوي ضرورتي به وجودشان نبود، حذف شدند تا در محاسبات شركت داده نشوند. اين اطلاعات شامل تراكنشهاي موجود همراه با آدرسها و فايلهاييكه صرفاً جهت ساخت يك صفحه وب استفاده مي‌شوند، بود. فايلهاي تصويري و كدهاي جاوا اسكريپت و فايلهاي مربوط به قالب و شكل ظاهري صفحات وب‌سايت در طي اين مرحله حذف شدند. در اين پژوهش تنها منبع اطلاعات، فايل ثبت وقايع بوده، بنابراين مرحله يكپارچه‌سازي در طول فرايند پيش‌پردازش حذف گرديد. بعضي از فيلدهاي موجود در فايل مانند تاريخ و زمان كه داراي مقادير تركيبي بودند، به منظور كاوش عميق‌تر در داده‌ها تجزيه شدند. پس از انجام مراحل فوق، داده‌هاي حاصل از سه مرحله قبل به درون انباره داده، انتقال يافت.

**شناسايي و معرفي قوانين وابستگي در داده**

منظور از قوانين وابستگي، كشف وابستگي بين اقلامي است كه رخداد آنها در يك زمان است؛ براي مثال، اجناسي كه در يك فروشگاه احتمال خريد آنها با هم در يك تراكنش خريد زياد است. اين اقلام داراي وابستگي هستند كه اين وابستگي‌ها به صورت A->B  نمايش داده مي‌شود. بهA  مقدم و به B مؤخر يا نتيجه گفته مي‌شود. كشف مجموعه عناصر تكرار شونده، به كشف وابستگي بين عناصر در مجموعه داده‌ها با حجم زياد منجر مي‌شود. بسياري از صنايع مشتاقند تا با داشتن حجم عظيمي از داده‌هايي كه به طور پيوسته جمع‌آوري و ذخيره مي‌شوند، چنين الگوهايي را از بانكهاي اطلاعاتي خود استخراج كنند. كشف روابط وابستگي قابل توجه در بين حجم عظيمي از تراكنشهاي كسب و كار ثبت شده، در بسياري از فرايندهاي تصميم‌گيري كسب و كار مانند طراحي كاتالوگ، بازاريابي عرضي و تحليل رفتار خريد مشتريان كمك مي‌كند.

از مقياسهاي مهم در قوانين وابستگي كه به منظور ارزيابي قوانين كشف شده مورد استفاده قرار مي‌گيرد، Support و  confidence هستند كه به ترتيب سودمندي و قطعيت قوانين كشف شده را نتيجه مي‌دهند.

**·Confidence:** زماني كه خريد يك قلم به خريد اقلام ديگري منجر مي‌شود، احتمال رخداد با استفاده از اين معيار اندازه‌گيري مي‌شود.

**·Support:** اگر خريد دو كالا با هم انجام شود، ميزان احتمال رخداد آن با اين معيار، اندازه‌گيري و ميزان درصد خريد آنها با هم، با عدد support نشان داده مي‌شود.

قوانين وابستگي داراي يك آستانه حداقل support  و يك آستانه حداقل confidence هستند كه با توجه به اين مقدار آستانه، معناداري قوانين تشخيص داده مي‌شود. اين آستانه مي‌تواند توسط كارشناسان و يا نرم‌افزار، تنظيم شود. تحليلهاي بيشتر مي‌تواند براي كشف وابستگي‌هاي قابل توجه بين عناصر وابسته به‌كار گرفته شود.

كشف قوانين وابستگي، داراي دو مرحلة تكرارپذير است:

1. كشف تمامي مجموعه يا itemset هاي تكرارپذير

2. توليد قوانين محكم از itemset هاي تكرارشونده

در اين پژوهش، به منظور كشف مجموعه‌هاي تكرارپذير در قوانين وابستگي منطقي، از الگوريتم اي‌پريوري[[17]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn17" \o ") استفاده شده است. «اي پريوري» يك روش تكرارپذير به كار مي‌گيرد كه  k-itemsetها براي يافتن (K+1)-itemset ها مورد استفاده قرار مي‌گيرند و از دو بخش الحاق[[18]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn18" \o ") و هرس***[[19]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn19" \o ")***تشكيل شده‌اند. زماني كه itemset هاي تكرارپذير از بين تراكنشها در بانك اطلاعات به دست آمدند، ايجاد قوانين وابستگي محكم از آنها به راحتي امكان‌پذير است كه با استفاده از معادله زير انجام مي‌شود:

Confidence (AÞB) = P (B|A) = support\_count (AÈB)/ support \_ count (A)

احتمال شرطي فوق بيان مي‌كند كه support\_count(AÈB) تعداد تراكنشهاي شامل AÈB و support\_count(A) تعداد تراكنشهاي شامل itemset ، Aمي‌باشند. بر اساس اين معادله، قوانين وابستگي مي‌تواند به صورت زير توليد شود:

·براي هر itemset تكرارپذير L، همه زير مجموعه‌هاي غير تهي آن ايجاد شود.

·براي هر زير مجموعه غير تهي S از L قانون:

S Þ(l-s): if (support\_count (l)/support\_count(s)) ³min\_conf, where min\_conf is the minimum confidence threshold.

يعني اگر تعداد تكرارهاي Itemset انتخابي l بر تعداد تكرارهاي زيرمجموعه انتخاب شده از آن، از مقدار min\_conf بزرگتر باشد آنگاه sÞ(l-s) يك قانون وابستگي است. چون قوانين ازitemset  هاي تكرارپذير توليد مي‌شوند، هر كدام به صورت خودكار، minimum support مورد نظر را دارند.

**تجزيه و تحليل داده‌ها**

در اين بخش يافته‌ها بر اساس الگوريتم «قوانين وابستگي» و اجراي اين الگوريتم بر روي داده‌هاي موجود در انباره داده حاصل از عمل پيش‌پردازش فايلهاي ثبت وقايع، تجزيه و تحليل شده است. در ابتدا، مدلهاي داده كاوي روي داده‌هاي موجود در جدولهاي مختلف موجود در انباره داده، طراحي شده و با اجراي الگوريتم مورد نظر، قوانين وابستگيو رابطه‌هاي موجود بين اقلام اطلاعاتي، كشف و بر اساس اين قوانين، پيش‌بيني‌‌هايي صورت گرفته كه هر كدام به صورت مدل جداگانه در اين بخش آورده شده است. در زمان طراحي مدل، تعريف متغيرهاي ورودي و متغيرهايي كه پيش بيني روي آنها انجام مي‌شود، ضروري است. تعيين اين متغيرها و انتخاب آنها به عنوان ورودي و متغير پيش‌بيني شونده، اهميت بسياري دارد و مي‌تواند در روند ايجاد مدل و نتايج خروجي و همچنين معناداري قوانين كشف شده، تأثير بسياري بگذارد. بنابراين، اين مرحله به دانش و مطالعه روي فيلدهاي تعريف شده در انباره داده و تسلط كافي بر درك داده‌ها نياز دارد.

الگوريتم «قوانين وابستگي» در داده‌هاي موجود در انباره داده به دنبال مجموعه‌هاي تكرارپذير معنادار كه معناداري آنها بر اساس معيار minimum\_support ارزيابي مي‌شود، جستجو كرده و به فهرست Itemset هاي معنادار تكرارپذير دست مي‌يابد. سپس در اين مجموعه‌ها به دنبال كشف روابط وابستگي نهفته بين اقلام هر مجموعه و مجموعه‌ها با يكديگر، قوانيني را با ضرايب معناداري مختلف كه بر اساس معيار minimum\_probabilty ارزيابي مي‌شود، استخراج مي‌كند. قوانين كشف شده داراي مقادير مختلف Confidence (كه در نرم‌افزار Sql Server2005  با عنوان Probability نام برده شده است) بوده و بيانگر احتمال رخداد آن قانون است. در تمام مدلهاي ارائه شده، از مقدار پيشنهادي نرم‌افزار براي minimum\_probabilty و minimum\_support استفاده شده است. براي استفاده ازالگوريتم، نرم‌افزار Sql Server  نسخه 2005 و Sql Server Analysis Services نرم‌افزارMicrosoft Visual Studio.net  نسخه 2005 استفاده شده است. مدلهاي طراحي شده روي سه مقطع تحصيلي كارشناسي، كارشناسي ارشد و دكتري اجرا شده است.در مدلهاي استفاده شده، كلية اطلاعات علمي و اطلاعاتي با پسوندهاي pdf ، doc، txt، zip، xls، ppt و rtf و پسوندهاي htm, html, asp, aspx ,php, do به عنوان صفحات ملاقات شده توسط مشتريان در نظر گرفته شده است.

**مدل كاوش شماره1**

اين مدل رفتار مشتريان در استفاده از پايگاه‌ها و صفحات پر استفاده توسط آنها پيش‌بيني شده است. در جدول 1، بخشي از عناصر تكرارپذير كه رخداد آنها با هم بوده، آورده شده است:

**جدول1. بخشي از مجموعه‌هاي تكرارپذير پس از اجراي الگوريتم Association Rules**

**با minimum support=1**

|  |  |  |  |
| --- | --- | --- | --- |
| **Row** | **Support** | **Size** | **ItemSet** |
| 1 | 7 | 2 | help/whgdata/ = Existing, w Name = csa.com |
| 2 | 7 | 1 | rpsv/cw/vhosts/oecdthemes/99980037/v1998n1/ = Existing |
| 3 | 7 | 3 | ids70/ = Existing, w Name = csa.com, help/ = Existing |
| 4 | 7 | 2 | ids70/ = Existing, help/ = Existing |
| 5 | 7 | 3 | csaillumina/ = Existing, w Name = csa.com, help/ = Existing |

رديف 2 نشان مي‌دهد كه مسير rpsv/cw/vhosts/oecdthemes/99980037/v1998n1 چندين مرتبه و به تكرار ملاقات شده است. همچنين، رديف 3 نشان مي‌دهد مسير help و وب‌سايت csa.com به تكرار با هم ديده شده‌اند.

**قوانين كشف شدةمدل كاوش 1**

از مجموعه‌هاي تكرارپذير حاصل شده در مرحلة قبل، روابط وابستگي بين اقلام يك مجموعه و مجموعه‌ها با يكديگر كشف و قوانين وابستگي ارائه شده، در جدول 2 ارائه شده است:

**جدول2. قوانين كشف شده پس از تعيين مجموعه هاي تكرارپذير 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/ = |

براي مثال، قانون شماره 7 بيان مي‌كند 80% مشترياني كه از پايگاه‌هاي موجود روي csa.com استفاده مي‌كنند، وارد صفحه راهنما مي‌شوند. همچنين، قانون شماره 9 بيان مي‌كند 90% مشترياني كه از پايگاه‌هاي sourcecode.com استفاده مي‌كنند، وارد صفحه rpsv/cw/cgi-bin/fastforward مي‌شوند.

همچنانكه در شبكة وابستگي مربوط به اين مدل نشان داده شده، مسير rpsv/cw توسط مشتريان بسيار استفاده شده است.

 

**شكل3. شبكة وابستگي پايگاه csa.com و رفتار حركت مشتريان مراجعه‌كننده به اين وب‌سايت**

در اين شبكه، به صورت واضح نشان داده شده است كه مشتريان در زمان ورود به سايت csa.com،  به دفعات وارد صفحة راهنما مي‌شوند. دليل اين امر مي‌تواند ناآشنا بودن مشتريان با اين سايت و پايگاه‌هاي موجود در آن باشد كه در زمان ورود به دليل ناآگاهي از هدفهاي سايت دچار مشكل شده‌اند و اين از دلايل طراحي ضعيف سايت است. همچنين، اين احتمال وجود دارد كه مشتريان در يافتن اطلاعات مورد نياز خود با مشكلاتي روبه‌رو مي‌شوند؛ لذا در صدد رفع نياز خود، به راهنماي سايت مراجعه مي‌كنند. در قانون ديگري كه توسط اين الگوريتم كشف شده، مسير پر استفاده در پايگاه‌هاي موجود در sourceoecd.com  مي‌باشد. در زير، شبكة وابستگي اين قانون نشان داده شده است.

 

**شكل4. شبكة وابستگي پايگاه sourceoecd.com و رفتار مصرف مشتريان مراجعه كننده به اين وب‌سايت**

شبكة فوق، ميزان استفادة مشتريان مراجعه‌كننده به وب‌سايت sourceoecd.com از مسير rpsv/cw و صفحات پرمصرف را نشان مي‌دهد. مشتريان در زمان ورود به سايتsourcecode.comبه منظور استفاده از پايگاه‌هاي موجود روي آن، به تكرار وارد صفحه rpsv/cw/cgi-bin/fastforward مي‌شوندو اين نشان مي‌دهد در اين مسير اطلاعات مفيدي وجود دارد. همچنين، مسير rpsv/cw از ديگر مسيرهاي پر استفاده توسط مشتريان است. با توجه به اين قوانين، مي‌توان با بررسي بيشتر اطلاعات موجود در مسيرهاي پرمصرف، دسته‌بندي بهتري را براي چينش اطلاعات در نظر گرفت و آنها را در مسيرهاي كوتاه‌تر قرار داد. بدين ترتيب، ترافيك شبكه كنترل شده و مشتريان در يافتن نيازهاي خود به رضايت بيشتري دست خواهند يافت.

**مدل داده‌كاوي شماره2**

در اين مدل، رفتار اطلاع‌يابي مشتريان مقيم در كشورهاي مختلف در مقطع ليسانس، پيش‌بيني شده است. در جدول 3 عناصر تكرارپذيري كه رخداد وقوع آنها با هم بوده، آورده شده است.

**جدول3. بخشي از مجموعه‌هاي تكرارپذير كشف شده پس از اجراي الگوريتم**

**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 اين جدول، بخشي از عناصر تكرار شونده آورده شده است. براي مثال، رديف 4 نشان مي‌دهد رخداد ملاقات پايگاه‌هاي موجود در ebsco.com و umi.com به تكرار با هم بوده و تعداد مرتبه اين رخداد 209 است و اين با توجه به حداقل مقدار Support ، مقدارقابل توجهي است.

**قوانين كشف شدهمدل كاوش 2**

پس از ايجاد مجموعه‌ها، رابطه‌هاي وابستگي موجود بين اين مجموعه‌ها و اقلام هر مجموعه كشف شد:

**جدول4. قوانين كشف شده پس از تعيين مجموعه‌هاي تكرارپذير ارائه شده در مرحلة 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 |

براي مثال، قانون 7 نشان مي‌دهد رشتة حقوق در مقطع ليسانس با احتمال 87% از پايگاه umi.com استفاده مي‌كند. همچنين قانون شماره 9 نشان مي‌دهد مشتريان از اين پايگاه با احتمال 87% فايلهايي از نوع pdf دانلود مي‌كنند.

پس از كشف قوانين وابستگي براي درك بهتر، ارتباطهاي كشف شده در قالب شبكه وابستگي در شكل 5 نمايش داده شده است.

 

 **شكل5. شبكة وابستگي پايگاه umi.com و رفتار مصرف مشتريان مراجعه‌كننده از رشته‌هاي مختلف تحصيلي به اين پايگاه در روزهاي هفته و وضعيت ناوبري آنها در سايت**

در شبكة فوق، همان‌طور كه نمايش داده شده است، تمامي رشته‌ها در مقطع ليسانس از پايگاه‌هاي موجود در Umi.com در طول روزهاي هفته استفاده كرده و اطلاعات علمي را با پسوند pdf دانلود مي‌كنند. مشتريان در اين مقطع پيمايش نيز داشته‌اند. دو پايگاه lexisnexis.comوebsco.com نيز وضعيتي مشابه به umi.com دارند و  از پايگاه‌هاي پر مصرف در مقطع ليسانس مي باشند كه در طول روزهاي هفته توسط كلية رشته‌ها در اين وب‌سايت مورد استفاده قرار مي‌گيرند.پايگاه‌هايي مانند abs.gov.au و ovid.com وجود دارند كه تنها در بعضي رشته‌ها مورد استفاده قرار مي‌گيرند و در زير شبكه وابستگي مربوط به ovid.com و نحوة ناوبري كاربرانش آورده شده است:

  

**شكل 6. شبكة وابستگي پايگاه ovid.com و رفتار مصرف مشتريان مراجعه‌كننده در مقطع ليسانس از رشته‌هاي مختلف تحصيلي به اين پايگاه در روزهاي هفته و وضعيت ناوبري آنها در سايت**

در مجموع، تمامي رشته‌ها در مقطع ليسانس از پايگاه‌هاي  umi.com, ebsco.com , lexisnexi.com مكرراً در طول روزهاي هفته استفاده و اطلاعات علمي نيز دانلود مي‌كنند.در اين مقطع، مشتريان تمركز روي روز خاصي در هفته ندارند و در طول روزهاي هفته مراجعه داشته‌اند.

**مدل داده‌كاوي شماره 3**

در اين مدل، رفتارهاي اطلاع‌يابي مشتريان مقيم در كشورهاي مختلف در مقطع فوق ليسانس پيش‌بيني شده است. نتايج حاصل از اجراي الگوريتم، كشف قوانين وابستگي در اين مدل است.

**جدول5. بخشي از مجموعه‌هاي تكرارپذير كشف شده پس از اجراي الگوريتم**

**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 |

رديف شماره 5 نشان مي‌دهد در اين مقطع، بازديد از پايگاه‌هاي موجود در lexisnexis.com و umi.com در تراكنشهاي كاربران به تكرار با هم رخ داده است.

**مجموعه قوانين كشف شده**

از مجموعه‌هاي تكرارپذير، رابطه‌هاي وابستگي جستجو و قوانين وابستگي كشف شد. بخشي از اين قوانين، در جدول 6 نشان داده شده است.

**جدول6. قوانين كشف شده پس از تعيين مجموعه‌هاي تكرارپذير ارائه شده**

**در مرحلة 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 |

براي مثال، قانون3 نشان مي‌دهد مشترياني كه در رشتة تجارت (BUS Master Cours) در مقطع فوق ليسانس بوده و فايلهاي نوع pdf دانلود مي‌كنند، با احتمال 88% از پايگاه umi.com استفاده مي‌كنند. همچنين، قانون شماره 4 نشان مي‌دهد مشترياني كه در رشتة تجارت (BUS Master Cours) در مقطع فوق ليسانس هستند و در روزهاي يكشنبه وارد وب‌سايت مي‌شوند، با احتمال 88% اين پايگاه را ملاقات مي‌كنند.

پس از كشف قوانين وابستگي براي درك بهتر، ارتباطهاي كشف شده در قالب شبكة وابستگي، در شكل 7 نمايش داده شده است:

 

**شكل7. شبكة وابستگي پايگاه umi.com و رفتار مصرف مشتريان مراجعه‌كننده در مقطع فوق ليسانس از رشته‌هاي مختلف تحصيلي به اين پايگاه در روزهاي هفته و وضعيت ناوبري آنها در سايت**

در شبكة فوق، همان‌طور كه نمايش داده شده است، تمامي رشته‌ها در مقطع فوق ليسانس از پايگاه‌هاي موجود در Umi.com در طول روزهاي هفته استفاده و اطلاعات علمي را با پسوند pdf دانلود مي‌كنند. همچنين، در اين پايگاه پيمايش نيز انجام داده‌اند. از نتايج ديگر از كشف وابستگي‌ها، پايگاه‌هايي است كه در اين مقطع به تكرار توسط كاربران ملاقات مي‌شوند. اين پايگاه‌ها  ebsco.com، lexisnexis.com, springerlink.com ,newsbank.com ,viewSwier.com ,saiglobal.com ,galegroup.com ,netlibrary.com, jstor.org, factiva.com ,intersicence.wiley.com  مي‌باشند و وضعيتي شبيه به umi.com دارند.

 

 **شكل8. شبكة وابستگي پايگاه ebsco.com و رفتار مصرف مشتريان مراجعه‌كننده در مقطع فوق ليسانس از رشته‌هاي مختلف تحصيلي به اين پايگاه در روزهاي هفته و وضعيت ناوبري آنها در سايت**

پايگاه ieee.org توسط سه رشته در اين مقطع در روزهاي دوشنبه، سه‌شنبه، پنج‌شنبه و جمعه مكرراً استفاده شده است. در اين پايگاه، عمدتاً عمل ناوبري انجام شده است.

در كل، تمامي رشته‌ها در مقطع فوق ليسانس از پايگاه‌هاي شامل ebsco.com  ، lexisnexis.com, springerlink.com ,newsbank.com ,viewSwier.com ,saiglobal.com ,galegroup.com ,netlibrary.com, jstor.org, factiva.com ,intersicence.wiley.com مكرراً در طول روزهاي هفته استفاده و اطلاعات علمي نيز از آنها دانلود مي‌كنند.در اين مقطع، مشتريان تمركز روي روز خاصي در هفته ندارند و در طول روزهاي هفته مراجعه داشته‌اند. همچنين، تعداد زيادي از پايگاه‌هاي علمي به صورت مشترك بين رشته‌هاي مختلف در كل روزهاي هفته توسط مشتريان استفاده مي‌شود.

**مدل داده‌كاوي شماره 4**

در اين مدل، رفتارهاي اطلاع‌يابي مشتريان مقيم در كشورهاي مختلف در مقطع دكتري پيش‌بيني شده است. در جدول 7، بخشي از اين عناصر آورده شده است.

**جدول7. بخشي از مجموعه‌هاي تكرارپذير پس از اجراي الگوريتم**

**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 |

براي مثال، در رديف 5 عناصر تكرار شده نشان مي‌دهد كه پايگاه‌هاي galegroup.com و umi.com توسط كاربران اين مقطع در يك تراكنش بارها رخداد داشته‌اند.

**قوانين كشف شده**

پس از كشف مجموعه‌هاي تكرار شونده با استفاده از قانون تكرارپذيري در تراكنشهاي اجرا شده توسط كاربران، قوانين وابستگي كشف شد. بخشي از اين قوانين در جدول 8 آورده شده است.

**جدول8. قوانين كشف شده پس از تعيين مجموعه‌هاي تكرارپذير 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 |

براي مثال، قانون شماره 7 نشان مي‌دهد مشتريان در رشتة SCIدر مقطع دكتري كه فايل از نوع pdf دانلودمي‌كنند، با احتمال 92% از پايگاه umi.com استفاده مي‌كنند. همچنين، قانون 8 نشان مي‌دهد مشترياني كه در مقطع دكتري فايلهاي نوع pdf دانلودمي‌كنند و در روزهاي دوشنبه مراجعه داشته‌اند، با احتمال90% از پايگاه umi.com استفاده مي‌كنند. قانون شماره 9 نشان مي‌دهد مشترياني كه در رشته علوم انساني در مقطع دكتري هستند و در روزهاي دوشنبه وارد وب‌سايت مي‌شوند، با احتمال 88% از اين پايگاه بازديد مي‌كنند.

پس از كشف قوانين وابستگي براي درك بهتر، ارتباطهاي كشف شده در قالب شبكه وابستگي نشان داده شده است.  در شكل 9 پايگاه umi.com و رفتار كاربران استفاده‌كننده در طول روزهاي هفته نشان داده شده است:

   

**شكل 9. شبكة وابستگي پايگاه umi.com و رفتار مصرف مشتريان مراجعه‌كننده در مقطع دكتري از رشته‌هاي مختلف تحصيلي به اين پايگاه در روزهاي هفته و وضعيت ناوبري آنها در سايت**

در شبكة فوق، همان‌طور كه نمايش داده شده است، تمامي رشته‌ها در مقطع دكتري از پايگاه‌هاي موجود در Umi.com در طول روزهاي هفته استفاده و اطلاعات علمي با پسوند pdf ، zip، xls، دانلودمي‌كنند. همچنين، در اين پايگاه پيمايش نيز انجام داده‌اند. همچنين، شبكة وابستگي مربوط به ديگر پايگاهebsco.com و رفتار حركت ملاقات‌كننده نشان داده شده است.

در كل، تمامي رشته‌ها در مقطع دكتري از پايگاه‌هاي:

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

مكرراً در طول روزهاي هفته استفاده و از اين پايگاه‌ها اطلاعات علمي نيز دانلودمي‌كنند.در اين مقطع، مشتريان تمركز روي روز خاصي در هفته ندارند و در طول روزهاي هفته مراجعه داشته‌اند. تعداد پايگاه‌هايي كه در اين مقطع توسط مشتريان مورد استفاده قرار مي‌گيرد، نسبت به دو مقطع ليسانس و فوق ليسانس كمتر است.در اين مقطع، نسبت به دو مقطع ديگر، مشتريان انواع متنوع‌تري از اطلاعات علمي را استفاده كرده‌اند.چنانكه در شكل نشان داده شده است، مشتريان، وب‌سايتهاي متنوعي را ناوبري كرده‌اند اما عمدتاً هيچ دانلود اطلاعات علمي در طول ناوبري نداشتند. اين مي تواند دليلي بر ضعيف بودن پايگاه‌ها در اين مقطع باشد كه نياز اين دسته از مشتريان را پاسخگو نبوده است.

**مدل كاوش شماره**5

اين مدل، داده‌كاوي با توجه به سه مقطع تحصيلي ذكر شده، پيش‌بيني مي‌كند كه مشتريان در مراجعات خود عموماً از چه پايگاه‌هايي با هم استفاده مي‌كنند.

**مجموعه اقلام تكرارپذير كشف شده**

در اين مدل 18 minimum\_support=مقدار پيشنهادي الگوريتم توسط نرم‌افزار مي‌باشد و تعيين‌كننده حداقل مقدار براي قابل قبول بودن اقلام وابسته است. براي مثال، رديف شماره 1 بيان مي‌كند 40 مرتبه پايگاه‌هاي galegroup.com و umi.com در يك تراكنش ملاقات كاربران، رخداد همزمان داشته‌اند. در زير، چند نمونه از خروجي حاصل در اين مرحله نشان داده شده است:

**جدول9. بخشي از مجموعه‌هاي تكرارپذير كشف شده پس از اجراي الگوريتم 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 |

**قوانين كشف شده**

در اين مرحله، به كشف روابط وابستگي بين اقلام هر مجموعه پرداخته شد.قوانين وابستگي در واقع رابطة موجود بين اقلام را با توجه به قوانيني كه قبلاً اشاره شد، كشف و براي هر قانون مقدار عددي Probability كه تعيين‌كننده احتمال رخداد قانون است، ارائه شده است. در زير، بخشي از قوانين حاصل از مجموعه‌هاي تكرارپذير آمده است.

**جدول10. قوانين كشف شده پس از كشفمجموعه‌هاي تكرارپذير 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 |

براي مثال، قانون شماره 2 مطرح مي‌كند مشترياني كه از پايگاه‌هاي موجود در metapress.com استفاده مي‌كنند، با احتمال 86% به پايگاه‌هاي موجود در umi.com نيز مراجعه داشته‌اند. همچنين، در قانون شماره 5، مشترياني كه از پايگاه‌هاي موجود در ebsco.com استفاده مي‌كنند، با احتمال 79% به پايگاه‌هاي موجود در umi.com مراجعه داشته‌اند.

پس از كشف قوانين وابستگي براي درك بهتر، ارتباطهاي كشف شده در قالب شبكه وابستگي نمايش داده شده است. در شبكه نشان داده شده مشترياني كه از    پايگاه‌هاي metapress.com، Ebsco.com، ebrary.com، abs.gov.au، springerlink.com،  unsource.com و galegroup.com استفاده كرده‌اند، به پايگاه umi.com نيز مراجعه داشته‌اند.

 

**شكل 10.  شبكة وابستگي مربوط به پايگاه‌هايي كه در كنار پايگاه umi.com ملاقات شده‌اند**

همچنين، در شبكة وابستگي زير نشان داده شده استكاربراني كه از پايگاه‌هاي موجود در ebsco.com  و ebrary.com  ديدن كرده‌اند، به پايگاه‌هاي موجود در sciencedirect.com نيز مراجعه كرده‌اند:

 

**شكل11. بخشي از شبكة وابستگي مربوط به پايگاه‌هايي**

**كه در كنار پايگاه sciencedirect.com ملاقات شده‌اند**

**مدل كاوش شماره6**

در اين مدل، با توجه به سه مقطع تحصيلي ذكر شده، پيش‌بيني مي‌كند كه مشتريان مقيم كشورهاي مختلف چه كالاهاي علمي را با هم استفاده مي كنند.

**مجموعه اقلام تكرارپذير كشف شده مدل كاوش 6 پس از اجراي الگوريتم**

در اين مدل، با توجه به  minimum support=1، مقدار پيشنهادي الگوريتم، تعدادي از مجموعه‌هاي 1 و 2 و 3 عنصري كشف شدند كه بخشي از آنها در جدول 3 نمايش داده شده است:

**جدول11. بخشي از مجموعه‌هاي تكرارپذير مدل2 پس از اجراي الگوريتم 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 |

**قوانين كشف شده مدل كاوش 6**

اين مجموعه‌ها بر اساس قوانين وابستگي و با استفاده از قانون تكرارپذيري در تراكنشهاي اجرا شده توسط مشتريان، كشف و سپس به دنبال كشف روابط وابستگي بين اين مجموعه‌ها و اقلام هر مجموعه جستجو كرده و قوانين وابستگي توسط اين مدل ارائه شد.

**جدول12. قوانين كشف شده پس از تعيين مجموعه‌هاي تكرارپذير 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 |
| 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 |

براي مثال، قانون 3 بيان مي‌كند 80% مشترياني كه كالاي علمي 673546165327426.pdf را دانلود[[20]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn20" \o ") كرده‌اند، كالاي علمي x61m545652q08048.pdf را نيز دانلود كرده‌اند. قانون 4 بيان مي‌كند مشترياني كه كالاي علمي adajia.pdf  را دانلود كرده‌اند، كالاي علمي zdaw1a.pdf  را نيز دانلود كرده‌اند.پس از كشف قوانين وابستگي براي درك بهتر، ارتباطهاي كشف شده در قالب شبكة وابستگي در شكل 12 نمايش داده شده است:

 

**شكل 12. نماي كامل از شبكة وابستگي مدل 2**

در شكل زير، يكي از ارتباطهاي كشف شده به صورت واضح نشان داده شده است. چنانكه در شكل مشاهده مي‌شود، وابستگي بين دو فايل اطلاعاتي pdf .067354616532746 و x61m545652q08048.pdf وجود دارد.

 

**شكل13. شبكه وابستگي مربوط به بخشي از قوانين مدل2**

**بحث**

با بررسي دقيق «رفتار كاربران در استفاده از پايگاه‌ها و صفحات پر استفاده توسط آنها» مسيرهاي پر استفاده در پايگاه‌هاي مختلف كشف شد كه مي‌تواند يك منبع غني به منظور بهبود طراحي سايت بوده و همچنين در بعضي موارد در تصميم‌گيريهاي اساسي از آنها استفاده نمود. از مسيرهاي پراستفاده، صفحه راهنماي سايت csa.com مي‌باشد. از دلايل اين امر، مي‌تواند ناآشنا بودن كاربران با اين سايت و پايگاه‌هاي موجود در آن باشد كه در زمان ورود، به دليل ناآگاهي از هدفها و يا موفق نبودن در يافتن اطلاعات مورد نيازشان در سايت دچار سردرگمي شده و درصدد رفع نياز خود به راهنماي سايت مراجعه مي‌كنند و اين مي‌تواند از ضعفهاي طراحي سايت باشد. از معيارهاي مهم در طراحي يك سايت، بالا بودن قابليت استفادة آن است كه مانع از اتلاف وقت كاربران مي‌شود. كاربران تمايل دارند با ورود به سايت بتوانند به سرعت نياز اطلاعاتي خود را برآورده كنند (Jakob Nielsen, 1990) . براي رفع اين مشكل، ارائه نقشه سايت، قرار دادن اطلاعاتي در مورد هدفهاي سايت در صفحه اول وب‌سايت و ارائه راهنماي غني از محتوا و چگونگي دسترسي به آنها در هدايت اين دسته از كاربران مي‌تواند مفيد باشد.

از ديگر مسيرهاي پر استفاده، پايگاه‌هاي موجود در مسير rpsv/cw/cgi-bin/fastforward در sourceoecd.com  مي‌باشد. دليل آن مي‌تواند وجود اطلاعات مفيد در اين مسيرباشد كه كاربران زيادي را جذب نموده است. با توجه به اين قوانين، مي‌توان با بررسي بيشتر اطلاعات موجود در مسيرهاي پر استفاده، با دسته‌بندي كارآمدتر اقلام علمي و در نظر گرفتن سياستهاي مفيدتر در چينش اقلام و قرار دادن آنها در مسيرهاي كوتاه‌تر، ترافيك شبكه را كنترل نمود و به اين ترتيب كاربران در يافتن نيازهاي اطلاعاتي خود به رضايت بيشتري دست مي‌يابند. اين امر در كارايي طراحي وب‌سايت و همچنين در امر تصميم‌گيري به منظور تهية نيازهاي بيشتر كاربران، مفيد است.

مدل رفتار اطلاع‌يابي كاربران در مقطع ليسانس،حاكي از استفادة مستمر پايگاه‌هاي موجود در Umi.com، ebsco.com و lexisnexis.comدر طول روزهاي هفته است كه عمدتاً اطلاعات علمي با پسوند pdf دانلود كرده‌اند.اين پايگاه‌ها نيازهاي كاربران زيادي را برطرف مي‌كند و اين امر مي‌تواند به تامين‌كنندگان اصلي اطلاعات علمي اين وب‌سايت در اتخاذ تصميمهاي مهم تجاري ياري رساند. همچنين، پايگاه‌هايي مانند abs.gov.au و ovid.comنيز وجود دارند كه تنها در بعضي رشته‌ها مورد استفاده قرار مي‌گيرند. در مقطع فوق ليسانس، پايگاه‌هاي 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 كاربران اين مقطع است كه تعداد بيشتري از پايگاه‌ها را پوشش داده و اين امر به دليل نياز كاربران اين مقطع به پژوهش بيشتر مي‌باشد. در مقطع دكتري، در تمامي رشته‌ها از پايگاه‌هاي موجود در Umi.com، ebsco.com،  lexisnexis.com, springerlink.com galegroup.com , jstor.org, factiva.com ,intersicence.wiley.com  در طول روزهاي هفته استفاده شده و اطلاعات علمي با پسوند pdf ، zip، xls، دانلود مي‌شود. تعداد پايگاه‌هاي استفاده شده در اين مقطع توسط كاربران با توجه به دو مقطع ليسانس و فوق ليسانس كمتر بوده و اين امر مي‌تواند به دليل نياز گستردة كاربران اين مقطع به كاوش و بررسي‌هاي بيشتر و نبود اطلاعات لازم و مورد نياز اين مقطع در پايگاه‌هاي موجود باشد.

برخي پايگاه‌ها در كنار يكديگر مورد استفاده قرار مي‌گيرند؛ بدين معنا كه كاربران در طول ملاقات خود در وب‌سايت، به چندين پايگاه در طول اتصالشان رجوع داشته‌اند. براي نمونه، كاربراني كه از پايگاه‌هاي metapress.com، Ebsco.com، ebrary.com، abs.gov.au، springerlink.com،  unsource.com و galegroup.com استفاده كرده‌اند، به پايگاه umi.com نيز مراجعه داشته‌اند. همچنين، كاربراني كه به پايگاه‌هاي موجود در ebsco.com  و ebrary.com  مراجعه داشته‌اند، به پايگاه‌هاي موجود در sciencedirect.com نيز رجوع كرده‌اند. بنابراين، اين پايگاه‌ها داراي اطلاعاتي هستند كه مي‌تواند در كنار يكديگر ميزان بيشتري از نيازهاي كاربران را برطرف كند و هر كدام به تنهايي نمي‌توانند پاسخگوي تمام نيازهاي اطلاعاتي آنها باشند. از طرفي، بررسي اين امر مي‌تواند در ارائة پيشنهادهاي بهينه به كاربران و قرار دادن پايگاه‌هايي كه عموماً در كنار يكديگر به تكرار استفاده مي‌شوند در يك مكان، به بازيابي اطلاعات سرعت بخشد.

در اين وب‌سايت، بسياري از كالاهاي اطلاعات علمي به تكرار با هم استفاده شده‌اند. براي مثال، كاربراني كه كالاي علمي 673546165327426.pdf را دانلود كرده‌اند، كالاي علمي x61m545652q08048.pdf را نيز دانلود نموده‌اند. همچنين، كالاي علمي adajia.pdf   در كنار كالاي علمي zdaw1a.pdf  به تكرار با هم دانلود شده‌اند. اين امر نشان مي‌دهد اين كالاهاي علمي داراي وابستگي اطلاعاتي هستند و اين وابستگي در سياست چينش آنها در كنار هم بسيار مهم است. كشف وابستگي‌هاي كالاها مي‌تواند در سرعت بخشيدن به بازيابي اطلاعات و كاهش ترافيك سايت، تأثير زيادي داشته باشد.

**نتيجه‌گيري**

وب جهان‌گستر، يك منبع داده‌كاوي غني است و به يك مديريت توانا نياز دارد تا با استفاده از تكنيكهاي داده‌كاوي، دانش و اطلاعات موجود در داده‌هاي پيشين را كشف كند و بر اساس آن پيش‌بيني‌هايي به منظور تصميم‌گيريهاي مهم در مورد پايگاه‌هاي اطلاعاتي كارآمد و همچنين سياستهاي طراحي وب‌سايت انجام دهد. در نمونة مورد مطالعه، پايگاه‌هاي اطلاعاتي پراستفاده توسط كاربران در مقاطع مختلف شناسايي شد. از جمله پايگاه‌هاي پرمصرف در هر سه مقطع ليسانس و فوق ليسانس و دكتري، پايگاه‌هاي موجود در umi.com و factiva.com و  newbanks.com و lexisnexis.com و ebsco.comهستند. استفادةمكرر از پايگاه‌هاي موجود در umi.com نشان‌دهندة وجود اطلاعات علمي مناسب براي كاربران در سطوح مختلف تحصيلي است كه توجه عمدة كاربران را به خود جلب كرده است.

پايگاه‌هاي اطلاعات علمي ارائه شده در مقطع فوق ليسانس، نسبت به دو مقطع ديگر سطح مراجعه بالاتري دارند. در مقاطع تحصيلي ليسانس و فوق ليسانس، بيشترين استفادة اطلاعات علمي از نوع pdf بوده و در مقطع دكتري از انواع اطلاعات علمي مانند pdf ،  xip، xls و swf استفاده شده كه تنوع بيشتري دارند. زمانهاي استفاده از پايگاه‌هاي اطلاعات علمي توسط هر سه مقطع، اكثر روزهاي هفته است.

از ديگر نتايج قابل توجه در اين پژوهش، كشف مسيرهاي پراستفاده توسط كاربران و وجود مشكلات بازيابي اطلاعات در بعضي پايگاه‌ها مانند csa.com و همچنين شناسايي گلوگاه‌هاست. بعضي از اين مسيرهاي پر ترافيك به مسيرهاي مربوط به اطلاعات علمي پرمراجعه مربوط است كه عمدتاً در مسيرهاي طولاني قرار گرفته‌اند و اين مي‌تواند دليلي بر نبود دسته‌بندي مناسب اطلاعات در پايگاه‌ها باشد.

با بررسي قوانين وابستگي روي پايگاه‌هاي استفاده شده توسط كاربران مختلف، وابستگي‌هاي موجود بين اين پايگاه‌ها كشف شد. مجموعه‌اي از پايگاه‌ها مكرراً با هم توسط كاربران مختلف استفاده شد و كاربران در مراجعات خود به تكرار اين كالاهاي علمي را در كنار هم درخواست نموده‌اند. اين امر، وابستگي موجود بين پايگاه‌ها را نشان مي‌دهد. رعايت چينش پايگاه‌ها با توجه به وجود وابستگي اطلاعاتي آنها، مي‌تواند تأثير عميقي را به همراه داشته باشد.

**منابع**

- پاتكار، ويوك.ان. (1380). «كاربردهاي داده‌كاوي در كتابخانه‌ها و مؤسسات دانشگاهي». ترجمة مريم صراف‌زاده و افسانه حاضري. شماره سوم دوره پنجم. مجله الكترونيكي پژوهشگاه اطلاعات و مدارك علمي ايران [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). "WebUsage 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). **"**AnalyzingWeb 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.

1. Osmar Zaiane.

2. Web Access logs.

3. Web Usage Mining.

4. Web Log Mining.

5. Sen Zhang.

1. Ross Quinlan.

1. Robert Cooley.

2. Bamshad Mobasher.

3. Jaideep Srivastava.

4. Zhijian Yang.

5. Sharmila Subramaniam.

1. Online Analytical Mining.

2. Common Log Format (CLF).

3. Extended Common Log Format (ECLF).

1. Data Warehousing.

1. Preprocessing.

1. Aprior Algorithm.

2. Join.

3. Prune.

1. Download.