

استخراج دانش از پایگاه داده نرم افزارهای مدیریت اطلاعات و مستندات با تکنیک‌های داده کاوی

مهدی فشارکی

دانشیار،

دانشگاه صنعتی مالک اشتر

حسین شیرازی

دانشیار،

دانشگاه صنعتی مالک اشتر

اکرم بخشی*

کارشناس،

دانشگاه صنعتی مالک اشتر تهران

پذیرش: ۱۳۸۸/۱۲/۲۵

دریافت: ۱۳۸۸/۱۰/۲۸

فصلنامه علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران
شاپا (چاپی) ۵۲۰۶-۱۷۳۵
شاپا (الکترونیکی) ۵۵۸۳-۲۰۰۸
نمایه در SCOPUS، LISA و ISC
<http://jlist.irandoc.ac.ir>
دوره ۲۶ | شماره ۲ | صص ۲۵۹-۲۸۳
زمستان ۱۳۸۹

چکیده: حجم بزرگ داده‌ها به تنهایی به مدیران سازمان‌ها در تصمیم‌سازی و تصمیم‌گیری هیچ کمکی نمی‌کند، بلکه باعث سردرگمی مدیران سازمانها نیز می‌شود. بنابراین مدیریت داده‌های خام و تبدیل داده‌های خارجی و داخلی سازمان به اطلاعات و دانش با استفاده از تکنیک‌های گوناگون، نقش اساسی و محوری دارد. از تکنیک‌های معروف در این زمینه داده کاوی است، که می‌تواند بر روی بانک اطلاعاتی انجام شود و دانش مورد نیاز را به دست آورد. «دی اسپیس» سیستمی نرم‌افزاری، آرشیو دیجیتالی یا مخزن اطلاعات سازمانی «وب»، فضایی برای ذخیره‌سازی، بازیافت و مدیریت اطلاعات سازمان‌ها و مؤسسات ایجاد می‌کند. هدف از ارائه این مقاله استخراج دانش از سوابق و تاریخچه جستجوهای کاربران در پایگاه داده «دی اسپیس» با استفاده از الگوریتم‌های داده کاوی است، تا با به دست آوردن ارتباطات و وابستگی‌های میان عبارات‌های جستجو شده بتوان امکان جستجوی هوشمند را در «دی اسپیس» اضافه کرد. این کار علاوه بر افزایش سرعت و کارایی جستجوی کاربر، توانایی «دی اسپیس» را هم افزایش می‌دهد. با این ایده هم اینکه از جستجوهای بسیار کاربر برای مفهومی جلوگیری می‌شود، که در صرفه‌جویی در زمان بسیار مؤثر است و هم برای ایده دادن به کاربر در راستای نشان دادن عبارتهایی که با هم ارتباط منطقی دارند و در رساندن کاربر به هدف از جستجو، کمک می‌کند. سرانجام در بخش ارزیابی پژوهش نشان داده شده است که تعداد رکورد‌های بازیابی شده در «دی اسپیس» مجهز به داده کاوی بیش تر و هدفمندتر از «دی اسپیس» بدون این توانایی است و چون کاربر را در راستای جستجوی یاری داده و ایده هوشمندانه به او می‌دهد، رضایت‌مندی کاربر را نیز، به همراه دارد.

کلیدواژه‌ها: داده کاوی؛ «دی اسپیس» مدیریت دانش؛ بانک‌های اطلاعاتی؛ مدیریت داده؛ مدیریت اطلاعات و مستندات.

* پدیدآور رابط: akram390@yahoo.com

۱. مقدمه

در سال‌های اخیر مدیریت دانش به موضوعی مهم و حیاتی در صنعت‌ها و شرکت‌ها تبدیل شده است. در واقع آنچه تضمین‌کننده برتری‌های بلندمدت برای سازمان‌ها و جوامع است، میزان بهره‌گیری آن‌ها از منابع دانش است و اگر سازمانی نتواند دانش مورد نیاز را در جایگاه مناسب آن تشخیص دهد، در عرصه‌های رقابتی با مشکل مواجه خواهد شد. با این دیدگاه، مدیریت دانش ابزاری است که امکان بهره‌گیری مؤثرتر از دانش را در راستای اهداف و برنامه‌های سازمان‌ها فراهم آورده و رقابت‌پذیری سازمان را ممکن خواهد ساخت.

در راستای رقابت‌پذیری سازمان‌ها، فناوری اطلاعات در تبدیل داده‌های خارجی و داخلی سازمان به اطلاعات و دانش با استفاده از تکنیک‌های گوناگون نقشی اساسی و محوری دارد. اگرچه حجم زیادی از داده‌ها در سازمان‌ها وجود دارد، اما این داده‌ها به تنهایی در نظر تصمیم‌سازان و تصمیم‌گیران دارای ارزش زیادی نیستند اما برای ایجاد و اجرای آنالیزهای کمی و کیفی و ایجاد اطلاعات و دانش مورد نیاز سازمان برای تصمیم‌گیری قاطع، به این داده‌ها نیاز خواهد بود. زمانی که حجم اطلاعات از حدی فراتر می‌رود، عملیات مدیریت و دست‌یابی به اطلاعات و کیفیت دسترسی به این اطلاعات از اهمیت زیادی برخوردار است.

«دی اسپیس» سیستم آرشیو دیجیتال یا مخزن اطلاعات سازمانی وب و دارای فضایی برای ذخیره‌سازی، بازیافت و مدیریت اطلاعات سازمان‌ها و مؤسسات است (Smith 2003). از سوی دیگر، الگوریتم‌ها و تکنیک‌های داده‌کاوی در حجم زیاد داده‌ها، می‌تواند اطلاعات و دانش مورد نیاز سازمان را استخراج کند. این دانش می‌تواند سازمان‌ها را در راستای اهداف و برنامه‌هایشان یاری داده و رقابت‌پذیری سازمانی را امکان‌پذیر سازد.

در بخش نخست مقاله ابتدا نرم‌افزار مدیریت اطلاعات و مستندات متن باز «دی اسپیس»^۱ را توضیح و در بخش دوم به تشریح داده‌کاوی پرداخته می‌شود سپس در بخش سوم به انجام مراحل داده‌کاوی بر روی پایگاه داده «دی اسپیس» و تشریح طراحی نتایج داده‌کاوی بر روی نرم‌افزار مدیریت اطلاعات و مستندات متن باز «دی اسپیس» اشاره و نتایج به‌دست آمده بر روی یک مجموعه داده نمایان می‌شود.

۲. پیشینه پژوهش

محورهای علمی این مقاله شامل موضوعات مدیریت دانش، مدیریت داده و داده‌کاوی است. هدف اصلی این مقاله ارائه کاری جدید در لبه فناوری است و ادعایی در لبه‌های دانش ندارد.

اجرای این کار فنی نیازمند شناخت حوزه‌های مرتبط علمی در راستای موضوع‌های مدیریت دانش، مدیریت داده و داده‌کاوی به شرح زیر است:

۱. بیان مفاهیم، الگوها و کاربردهای داده‌کاوی (اکبرپور شیرازی ۱۳۸۷).
۲. داده‌کاوی و کاربرد آن در تصمیم‌گیری‌ها (تقوی فرد ۱۳۸۶).
۳. مدیریت دانش چه انجام می‌دهد؟ (Firestone 2005).
۴. مدیریت دانش در ارتباط با مشتری (غلامیان ۱۳۸۵).
۵. ارتباط مدیریت و انبار داده در پشتیبانی مدیریت دانش (Erdmann 2001).
۶. مدیریت دانش در محیط انبار داده‌های ناهمگون (Kerschberg 2000).
۷. سیستم‌های مدیریت بر مبنای دانش و انبار داده (Firestone 1999).
۸. مدیریت دانش انبار داده و داده‌کاوی (Hahsler 2001).

بنابراین هدف مقاله این است که بر روی مرور فناوری‌های پیشرفته مدیریت اطلاعات و مستندات و داده‌کاوی تمرکز شود. نرم‌افزارهای چندی برای مدیریت داده ارائه شده که معروف‌ترین آن‌ها «دی اسپیس» است. در ادامه به توضیح نرم‌افزارهای پردازد:

الف. «کوها»^۱: نخستین سیستم کتابخانه‌ای کد منبع باز در دنیا است. سیستم کوها که در نیوزیلند ایجاد شده است، بسته کامل فهرست‌نویسی، سیستم گردش «اوپیک»^۲ و مدیریت اعضا است. هرچند که این سیستم برای کتابخانه‌های بزرگ به اندازه کافی قوی نیست (از زبان برنامه‌نویسی «پرل»^۳ و پایگاه اطلاعاتی «ام.وای.اس.کیو.ال»^۴ استفاده می‌کند) اما تنها می‌تواند برای کتابخانه‌های کوچک تا متوسط استفاده شود (ارسطوپور ۱۳۸۶).

ب. «گرین استون»^۵: سیستم گرین استون امکانات جستجو و کاوش متون و تصاویر دیجیتالی شده را ارائه می‌دهد (Witten 2005).

پ. «دی اسپیس»^۶: این سیستم به این منظور طراحی شده است که به گروه بزرگی از افراد اجازه می‌دهد فایل‌های دیجیتالی را بارگذاری کنند، دی اسپیس اجازه می‌دهد به آن فایل‌ها دسترسی داشته باشید و آنها را مدیریت کنید (dspace 2008). در ادامه به بخشی از توانایی‌های «دی اسپیس» اشاره می‌شود:

۱. دسته‌بندی موضوعی و محتوایی مناسب اطلاعات
۲. توانایی شناسایی انواع قالب پرونده‌های دیجیتالی اعم از متن، عکس، نرم‌افزار، صدا و فیلم
۳. ذخیره مجموعه مناسبی از فراداده^۷ در مورد اطلاعاتی از جمله موضوع، عنوان، ویرایش، سال نشر، نام پدیدآورنده، انتشارات، شابک، تصویر، حجم، دسته‌بندی موضوعی، مشارکت‌کننده و ...

1. koha
4. MySQL

2. (OPAC)
5. Greenstone

3. Perl
6. dspace

7. metadata

۴. امکان تعیین سطوح دسترسی به کاربران و گروه‌های کاربران
۵. امکان بررسی تغییرات در اطلاعات و تضمین کیفیت نگهداری اطلاعات
۶. نمایش استاندارد اطلاعات در تمامی سیستم‌های عامل و مرورگرهای وب

ت.مای لایبراری^۱: با استفاده از مای لایبراری کاربران می‌توانند پنجره‌هایی برای مجموعه‌ها و منابع کتابخانه ایجاد کنند. با وارد شدن به سیستم و ویرایش کردن منابع انتخاب شده به صورت پیش فرض، کاربران کتابخانه می‌توانند جعبه‌های جستجوی مربوط به پایگاه‌های اطلاعاتی مورد علاقه‌شان را به علاوه لینک‌های مربوط به صفحات وب کلیدی در گروه‌های موضوعی گوناگون، به آسانی در دسترس داشته باشند (شریف ۱۳۸۶).

با بررسی و مطالعه دقیق «دی اسپیس» و با توجه به برتری توانایی‌ها و کارایی‌های آن نسبت به نرم‌افزارهای ذکر شده در بالا، «دی اسپیس» به عنوان سیستمی که کار مدیریت داده را به همراه پایگاه دانش انجام می‌دهد، برای انجام عمل داده کاوی انتخاب شده است.

۳. نرم‌افزار متن باز مدیریت اطلاعات و مستندات «دی اسپیس».

سیستم‌های مخزن سازمانی همانند «دی اسپیس» محیط‌های کنترل شده‌ای هستند که با رعایت چارچوب‌های قانونی و تجاری مناسب و فارغ از محل نگهداری فیزیکی و حوزه تصدی‌گیشان، دسترسی یک‌پارچه‌ها به متن کامل اطلاعات ذخیره شده به شکل دیجیتالی از طریق شبکه فراهم می‌آورند. در حقیقت مخزن سازمانی، فضایی برای ذخیره‌سازی، بازیافت و مدیریت اطلاعات سازمان‌ها و مؤسسات است. «دی اسپیس» سیستم آرشیو دیجیتالی یا مخزن اطلاعات سازمانی زیر نظر وب است که توسط دانشگاه «ام.آی.تی.»^۲ و شرکت «اچ.پی.»^۳ تولید شده است و گسترش یافته است (Smith 2003). این سیستم از پرکاربردترین و قدرتمندترین نرم‌افزارهای مدیریت اطلاعات سازمانی است که در بیش از سیصد مؤسسه بزرگ دنیا در ۵۵ کشور مورد استفاده شرکت‌ها، سازمان‌ها، کتابخانه‌ها، مراکز پژوهشی و دانشگاه‌های معتبر جهان مانند دانشگاه ملی استرالیا، کلمبیا، «ام.آی.تی.»^۴، اوهایو، کمبریج، تورنتو و واشنگتن قرار گرفته است. «دی اسپیس» در واقع گنجینه و مخزن الکترونیکی و مرکزیت یافته حجم زیادی از آرایه‌های معنوی است که از طریق نهادهای پژوهشی معتبر ایجاد شده‌اند. «ام.آی.تی.» سالانه ده‌هزار مجموعه تولید می‌کند که به صورت الکترونیکی درآمده‌اند. این مجموعه‌ها شامل مجموعه مقاله‌های کنفرانس‌ها و گزارش‌های فنی دانشگاه‌ها است. بسیاری از این اطلاعات بسیار طولانی و غیرقابل دسترس هستند. متن باز بودن این محصول و بسترهای تولید آن، باعث افزایش سرعت و سهولت افزودن امکانات به سیستم شده است که توسعه آن در نسخه‌های آینده، این نرم‌افزار را تسهیل می‌کند.

1. mylibrary

2. MIT

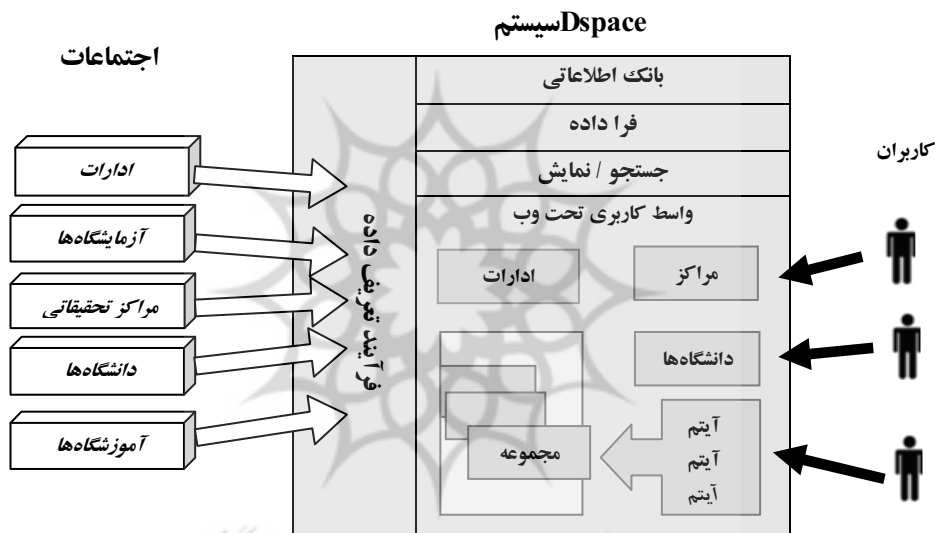
3. HP

4. MIT

سهولت دسترس جهانی به منابع محتوایی سازمان، جمع آوری محتوا در محلی متمرکز، نگهداری و ذخیره سازی دارایی های دیجیتال مانند پرونده های متنی و تصویر و صدا و فیلم، امکان انتقال اطلاعات و هماهنگی میان سیستم های نگهداری اطلاعات، از مهم ترین اهداف استفاده از این مخزن سازمانی است (Smith 2003).

۱-۳. مدل اطلاعاتی دی اسپیس»

در شکل ۱ اجزای مدل اطلاعاتی «دی اسپیس» نمایش داده شده است، که در ادامه به تشریح این اجزاء پرداخته می شود (dspace 2008):



شکل ۱. مدل اطلاعاتی «دی اسپیس» (Smith 2003)

۱-۳-۱. فراداده

برای تشریح فراداده ها از استاندارد «دوبلین کور» که شامل اطلاعاتی از قبیل موضوع، عنوان، ویرایش، سال نشر، نام پدید آورنده، انتشارات، زبان، شابک، تصویر، حجم، دسته بندی موضوعی، مشارکت کننده و ... است، استفاده می شود. در این سیستم سه فیلد عنوان، زبان و تاریخ اجباری و بقیه فیلدها اختیاری است. این اطلاعات برای نمایش و جستجو باید ایندکس گذاری شوند.

۱-۳-۲. واسط کاربری

واسط کاربری این سیستم زیر نظر وب است و برای دسترسی کاربر ارشد واسطی با تمامی دسترسی وجود دارد ولی برای کاربران عادی واسطی برای جستجو و نمایش وجود دارد.

۳-۱-۳. گردش کار

- برای تعریف فراداده در «دی اسپیس» باید ابتدا قالب و چارچوب کار که به صورت سلسله مراتب زیر است، ساخته شود.
- ✓ هر سند دیجیتالی به شکل جریانی از بیت‌ها درآمده (رشته های بیتی^۱) و در قالب فایلی ذخیره سازی می شود.
 - ✓ مجموعه ای از فایل های مرتبط با هم به همراه فراداده اختصاص داده شده به آن‌ها به عنوان یک آیت^۲ تعریف می شوند مثل مقاله ای به همراه کدها و فایل های ضمیمه آن
 - ✓ آیت های داده مربوط به هم در گروه‌ها^۳ دسته بندی می شوند مثل شاخه های پژوهش‌ها، سری گزارش های فنی
 - ✓ گروه های داده در قالب مجموعه‌ها^۴ دسته بندی می شوند مثل ادارات، مراکز پژوهشی و آزمایشگاه های پژوهشی

Communities

- ✓ Collections
 - Items (metadata plus bitstreams)
 - Digital files(bitstreams)

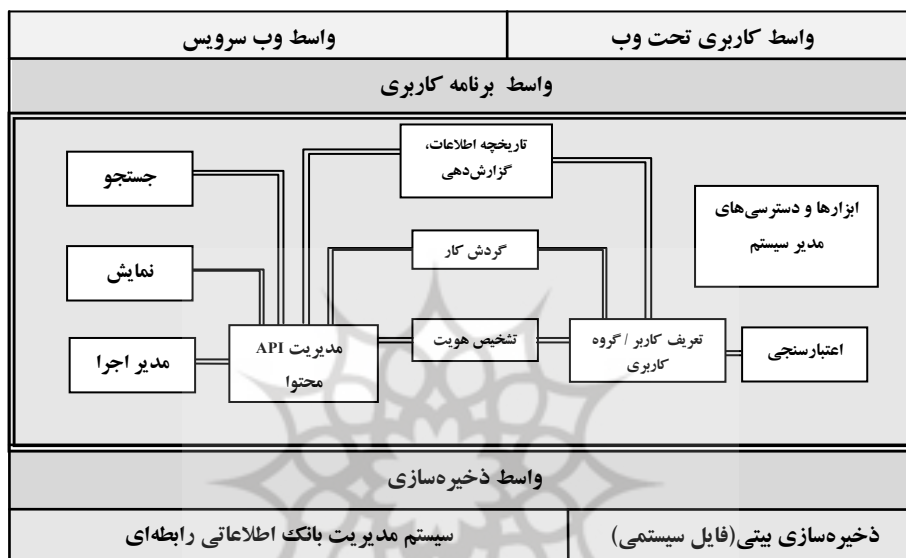
چون هدف از این پژوهش داده کاوی بر روی پایگاه داده «دی اسپیس» است، بنابراین برای انتخاب داده و تعریف داده های مورد نیاز در فرایند داده کاوی، نیاز به شناخت کامل مدل اطلاعاتی «دی اسپیس» و پایگاه داده آن و ارتباطات میان جداول بانک اطلاعاتی است.

۳-۲. مشخصات فنی «دی اسپیس»

«دی اسپیس» سیستم متن بازی است که سازمان‌ها با کمترین منبع و هزینه می توانند به آسانی آن را اجرا کنند. این سیستم طوری طراحی شده است که بر روی هر نوع سیستم عاملی اجرا می شود. کد آن به زبان برنامه نویسی «جاوا» نوشته شده است. سیستم مدیریت پایگاه داده رابطه ای آن «پستگرس. کیوال»^۵ است و از دیگر تکنولوژی های به کار رفته در آن وب سرورها و موتور «جاوا سرولت»^۶ «تامکت و ایچ»^۷ است.

1. bitstream	2. item	3. collection	4. community
5. PostgreSQL	6. Java Servlet	7. Apache , Tomcat	

«دی اسپیس» دارای معماری ای سه لایه است (شکل ۲) که شامل لایه های ذخیره سازی^۱، منطق کسب و کار^۲ و برنامه^۳ است و هر کدام از لایه ها با «ای.پی.آی.»های^۴ خاص خود، اجازه بهینه سازی و توسعه را دارد (dspace 2008 Smith2003).



شکل ۲ معماری Dspace (Smith2003)

- لایه ذخیره سازی از فایل سیستم ها استفاده می کند. که این فایل سیستم ها توسط پایگاه داده «پستگرس.کیو.ال.» مدیریت می شوند. در این پایگاه داده اطلاعات اجتماعات، مجموعه ها، آیتم ها، فراداده ها و ارتباطات آن ها ذخیره می شود.
- لایه منطق کسب و کار لایه ای است که وظایف و فرآیندهای «دی اسپیس» در آن انجام می شود که شامل گردش کار، تاریخچه اطلاعات، جستجو، نمایش، تعیین دسترسی ها، تعریف کاربر و گروه ... است. هر «ماجول» با «ای.پی.آی.» مربوط به خود کار می کند و وظایف آن ماجول را در ارتباط با ماجول های دیگر انجام می دهد.
- لایه برنامه، لایه ارتباط با کاربر است و مهم ترین بخش آن واسط کاربری زیر نظر وب است. هدف از این پژوهش این است که پس از انجام داده کاوی بر روی پایگاه داده «دی اسپیس»، برای به کار بستن نتایج داده کاوی در «دی اسپیس»، در لایه منطق کسب و کار در «ماژول» جستجو،

1. storage
2. business logic
3. application
4. Application Programming Interface

توانایی جدید جستجوی هوشمند طرّحی و پیاده‌سازی گردد، به طوری که با وصل شدن به بانک اطلاعاتی علاوه بر جستجوی کلمه درخواستی از طریق کاربر، کلمه مرتبط با آن را نیز اگر در فرایند داده کاوی کشف شده باشد، جستجو کند.

۴. داده کاوی

تعاریف گوناگونی از داده کاوی بیان شده‌است که نمونه‌ای از آنها به شرح زیر ارائه می‌شوند (Han 2006):

- داده کاوی (یا روش‌های اکتشاف دانش) پردازشی است که داده را از جنبه‌های متفاوت و با تلخیص‌های متفاوت آنالیز می‌کند تا از حجم زیاد داده‌ها، اطلاعات کاربردی استخراج نماید. در این تعریف منظور از اطلاعات کاربردی اطلاعاتی است که بتواند بهره‌وری را افزایش یا هزینه را کاهش دهد و هر دو منظور را همزمان برآورده سازد. نرم‌افزار داده کاوی نیز یکی از ابزارهایی است که می‌تواند این هدف را محقق سازد. این ابزار باید بتواند داده‌ها را در کلاس‌های گوناگون دسته‌بندی و آنها را از طریق آنالیز مؤلفه‌های متفاوت آنها بررسی و تحلیل نماید. در این صورت می‌تواند نتایجی را استخراج نماید که همان اطلاعات کاربردی در مدیریت منابع یا دیگر زمینه‌های کارکرد مورد نظر سازمان بهره‌بردار باشد.

- داده کاوی در مواردی از جمله خرده‌فروشی (تعیین الگوهای خرید مشتریان، تجزیه و تحلیل سبد خرید بازار، پیش‌گویی میزان خرید مشتریان از طریق پست (فروش الکترونیکی)، بانکداری (پیش‌بینی الگوهای کلاه‌برداری از طریق کارت‌های اعتباری، تشخیص مشتریان ثابت، تعیین میزان استفاده از کارت‌های اعتباری بر اساس گروه‌های اجتماعی)، بیمه (تجزیه و تحلیل دعاوی، پیش‌گویی میزان خرید بیمه‌نامه‌های جدید توسط مشتریان)، پزشکی (تعیین نوع رفتار با بیماران و پیش‌گویی میزان موفقیت عمل‌های جراحی، تعیین میزان موفقیت روش‌های درمانی در برخورد با بیماری‌های سخت)، تجارت (تبلیغات، مدیریت ارتباط با مشتری) (شاه‌سمندی ۱۳۸۴) و غیره کاربرد دارد.

در ادامه به تشریح فنون داده کاوی پرداخته شده و علل انتخاب حوزه تحلیل وابستگی در این پژوهش بیان می‌شود.

فن‌های داده کاوی به دسته‌های زیر تقسیم می‌شوند (Han 2006)

- توصیف و تمایز گذاری در داده | از طریق دسته‌بندی‌ها و مفاهیم حاکم بر داده‌ها
 - توصیف داده‌ها از طریق گردآوری و خلاصه کردن داده‌های گروهی زیر نظر موضوع از عنوان‌های عمومی (که گروه، موضوع هدف نامیده می‌شود)

- تمایز گذاری داده‌ها از طریق مقایسه گروه هدف با گروه‌های دیگر (که البته توانایی مقایسه آنها با یکدیگر وجود داشته باشد (که به طور عموم تضاد گروهی-کلاسی نامیده می شود).
- و از طریق توصیف و تمایز گذاری داده‌ها که اجتماع دو حالت بالا رخ خواهد داد.
- تحلیل وابستگی
 - تحلیل وابستگی، کشف قوانین وابستگی میان اطلاعات است به نحوی که نمایان گر مقادیری از صفات اطلاعاتی است که بیش تر اوقات همراه با یکدیگر و در مجموعه‌ای داده در داده‌های بانک اطلاعاتی ظاهر می شوند.
 - در اساس ارتباط میان مجموعه اشیا (چیزها) وابستگی‌های جالب توجهی هستند که منجر به امکان آشکار سازی الگوهای مفید و قوانین وابستگی برای پشتیبانی تصمیم، پیش بینی‌های مالی، سیاست‌های بازاریابی، وقایع پزشکی و کاربردهای دیگر می شود. (S.Gyrorodi 2001)
 - تحلیل وابستگی‌ها حالت غیر نظارتی داده کاوی است که به جستجو برای یافتن ارتباط در مجموعه داده‌ها می پردازد. یکی از کاربردی ترین حالات تحلیل وابستگی‌ها، تجزیه و تحلیل سبد بازار است که در آن هدف، یافتن کالاهایی است که به طور همزمان خریداری می شوند. این کار کمک می کند که خرده‌روشان بهتر بتوانند کالاهای خود را سازمان دهی کرده و چیدمان بهتری از محصولات خود داشته باشند (Ye 2003).
- دسته بندی و پیش گویی
 - دسته بندی پردازشی است برای یافتن مجموعه‌ای از مدل‌ها (یا توابع) که مفاهیم یا کلاس‌های داده را برای تشخیص داده‌هایی که در هیچ کلاسی قرار نمی گیرند یا به صورتی نامشخص توصیف شده‌اند را توصیف نموده یا تشخیص دهند. درحقیقت دسته بندی داده‌ها، نوعی توانمندی را برای تشخیص یا پیش گویی داده‌های نامشخص ایجاد می کند و در عمل نیز مدل‌های برگرفته شده بر پایه داده‌های جامعه آماری نمونه تعریف و توصیف می شوند و پس از آن بر روی داده‌های دیگر تعمیم داده شده و باعث تشخیص کلاس داده یا پیش گویی داده‌هایی می شوند که به صورت خام وارد جامعه آماری مورد کاوش گردیده‌اند.
 - در سیستم داده کاوی یا سیستم‌های مشابه از فرم‌های گوناگونی برای نمایش و ارائه مدل‌های برگرفته شده استفاده می شود که از آن جمله می توان به موارد زیر اشاره کرد:
 - قوانین دسته بندی (قوانین if-then)؛ درخت تصمیم، فرمول‌های ریاضی و شبکه‌های عصبی

- تحلیل خوشه‌بندی
 - برعکس کلاس‌بندی و پیش‌گویی که داده‌ها را براساس نشان کلاس آنها بررسی و تحلیل می‌کرد، در روش خوشه‌بندی، تحلیل‌های انجام شده در این روش، بدون توجه به نشان کلاس داده‌ها انجام می‌شوند (اکبر پورشیرازی ۱۳۸۷).
- تحلیل تکامل و انحراف (تکمیل و خطا)
 - تحلیل تکامل داده در حقیقت مدل‌سازی و توصیف سیر تکاملی و تغییرات گرایش‌های موجود در داده‌ها در یک بازه زمانی است. این نوع مدل‌سازی ممکن است شامل هر کدام از انواع توصیف، تمایز، خوشه‌بندی، کلاس‌بندی و ارتباط میان داده‌هایی شود که به زمان بستگی دارند که این تحلیل‌ها به‌طور طبیعی از مقایسه الگوهای مربوط به داده‌های سری - زمانی، متناوب و یا تحلیل داده‌های مشابه یا موارد این‌چنینی، مجزا هستند.
- در این پژوهش به دلایل زیر حوزه تحلیل وابستگی انتخاب شده است:
 - حوزه کار در داده‌کاوی متناسب با هدف بیان شده در پژوهش است. در این نوشتار هدف، پیدا کردن قوانین و ارتباطات میان داده‌ها است.
 - عدم وابستگی به زمان در هدف بیان شده
 - عدم نیاز به دسته‌بندی داده‌ها در هدف بیان شده
 - عدم نیاز به توزیع آماری در این حوزه
 - عدم نیاز به داده‌های آموزشی در این حوزه
 - درک بهتر برای کاربران
- با بررسی الگوریتم‌های این حوزه، الگوریتم معروف «آپریوری»^۱ که کاربرد زیادی در این حوزه دارد و در این حوزه با این الگوریتم کار می‌شود، انتخاب گردیده است.

۵. فرایند داده‌کاوی بر روی پایگاه «دی اسپیس»

از آن جایی که «دی اسپیس» جستجوی عادی و پیشرفته را بر اساس کلمات ورودی انجام می‌دهد این نیاز وجود دارد که اگر «دی اسپیس» بتواند به کاربران خود به‌طور هوشمند، کلمات مرتبط با درخواست را معرفی کند، در رساندن کاربران به اهدافشان از جستجو، کمک شایانی کرده‌است. به‌علاوه در کم شدن زمان جستجو نیز بسیار مؤثر بوده‌است و این باعث افزایش سرعت و کارایی جستجوی کاربر و در نهایت رضایت‌مندی او می‌شود. بنابراین برای کشف این کلمات و

عبارات مرتبط باید عمل داده کاوی بر روی پایگاه «دی اسپیس» انجام شود. در ادامه ابتدا فرایند استاندارد داده کاوی به طور کلی بیان می شود (Han 2006) سپس این فرایند گام به گام بر روی داده های این پژوهش به صورت عملی تشریح می شود:

- تعیین بازه و دامنه سیستم مورد کاربرد و هدف نهایی ایجاد دانش؛
 - جمع آوری داده های موجود از تمامی حوزه های مورد نیاز؛
 - پاک سازی و پیش پردازش داده ها که ممکن است تا ۶۰ درصد کل کار را از نظر حجم فعالیت پوشش دهد؛
 - کاهش و تغییر شکل دادن داده ها (که ممکن است از طریق طراحی و ساخت انباره داده ها انجام شود)؛
 - انتخاب داده ها بر اساس نوع مفاهیم مورد نظر؛
 - انتخاب پردازش های مورد نظر در داده کاوی (مانند گردآوری، دسته بندی، رگرسیون، تحلیل وابستگی، خوشه بندی و...)
 - انتخاب الگوریتم مورد نظر در داده کاوی؛
 - انجام پردازش های داده کاوی (جستجو به دنبال الگوی مفاهیم مورد نظر)؛
 - ارزیابی الگوها و ارائه دانش حاصله از راه های بصری، دگرگونی داده ها، حذف الگوهای تکراری و...؛
 - و در پایان، دستیابی به دانش الگوهای رفتاری سیستم و استفاده از آن در کارکردهای عملیاتی.
- در این میان برخی از گام ها قابل تکرار شدن بوده و برخی (مانند تشکیل انباره داده ها) قابل گذر است، اما نکته مهم، طی مسیر صحیح و منطبق بر هدف کاوش در داده ها است.
- متناسب با فرایند استاندارد بالا، فرایند داده کاوی بر روی پایگاه نرم افزار متن باز مدیریت اطلاعات و مستندات «دی اسپیس» به صورت دیاگرام در شکل ۳ مطرح می شود.



شکل ۳. فرایند انجام داده کاوی بر پایگاه «دی اسپیس». مراحل داده کاوی بر روی پایگاه داده «دی اسپیس»

طبق دیاگرام بالا در ادامه تشریح می شود:

۱. بیان هدف از داده کاوی:

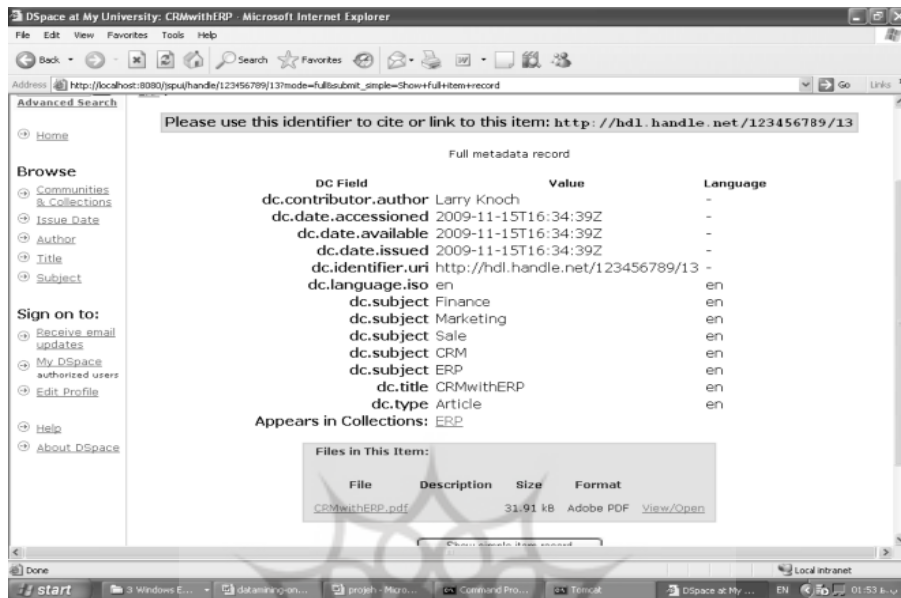
انجام داده کاوی بر روی سوابق و تاریخچه جستجوهای کاربران در نرم افزار متن باز «دی اسپیس» برای به دست آوردن ارتباطات و وابستگی های میان عبارت های جستجو شده است تا با به دست آمدن این ارتباطات، بتوان امکان جستجوی هوشمند را در «دی اسپیس» اضافه کرد که علاوه بر افزایش توانایی «دی اسپیس»، به کاربر در هر چه سریع تر رسیدن به اهدافش کمک کرد.

۲. جمع آوری و انتخاب داده برای داده کاوی متناسب با هدف:

در این بخش دودسته داده وجود دارد. دسته نخست، فراداده ها که بر اساس استاندارد «دوبلین کور» شامل اطلاعاتی مثل موضوع، عنوان، ویرایش، سال نشر، نام پدیدآورنده، انتشارات، شابک، تصویر، حجم، دسته بندی موضوعی، مشارکت کننده و ... است. در شکل ۴ بخشی از این اطلاعات در مورد تعریف فراداده نشان داده می شود.



استخراج دانش از پایگاه داده نرم افزارهای مدیریت اطلاعات | فناوری و دیگران



شکل ۴. نمونه‌ای از تعریف فراداده

اطلاعات فراداده‌ها در قالب فرایند زیر تعریف می‌شوند:

تشریح فیله‌ها ← بارگذاری فایل ← بازبینی داده‌ها ← ارائه مجوز ← تکمیل و اتمام فرایند. در فرایند بالا ابتدا فیله‌های فراداده پر می‌شوند، سپس فایل سند به آن ضمیمه می‌شود و در مرحله بعد اطلاعات وارد شده، بازنگری شده و اگر ایرادی داشت اصلاح می‌شود و در مرحله بعد مجوز ثبت آن تأیید شده در پایان اطلاعات فراداده به‌طور کامل در بانک اطلاعاتی «دی اسپیس» در جدول فراداده‌ها ثبت می‌شود. دسته دوم، کلمات و عبارات‌های جستجو شده توسط کاربران است (جدول ۱). مجموعه داده‌های آزمایشی در این پژوهش، حدود ۲۰۰۰ کلمه جستجو شده از حدود ۳۰ کاربر است. این داده‌ها بخشی از اطلاعاتی است که توسط جستجوهای کاربران شرکتی در پورتال سازمانی آن ذخیره شده و برای نمونه آزمایشی این پژوهش انتخاب شده است. این نهاد، شرکتی کامپیوتری است که در حوزه‌های نرم‌افزار، سخت‌افزار، شبکه و مشاوره پروژه‌های فناوری اطلاعات فعالیت می‌کند و حدود ۴۰۰ کارمند و دو شرکت زیرمجموعه دارد. داده‌های ذکر شده در جدولی در بانک اطلاعاتی «پستگرس.کیو.آل»، با فیله‌های کد کاربر، کلمه جستجو شده و تاریخ جستجو ذخیره می‌شوند. این داده‌ها تمام حوزه‌ها را از جمله حوزه‌های علمی، پژوهشی، فناوری، اطلاعات عمومی و... پوشش می‌دهند.

برای انجام داده کاوی در راستای هدف بیان شده، به دسته دوم داده‌ها نیاز است و بر روی کلمات جستجو شده فرایند داده کاوی انجام می‌شود ولی برای آزمایش و ارزیابی «دی اسپیس» مجهز به داده کاوی، به دسته نخست داده‌ها نیاز است.

جدول ۱ بخشی از «دیتاسیت»^۱ مورد آزمایش

ID	UserCode	WordSearch	DateSearch
2	1	erp	1388/05/01
3	1	customer	1388/05/01
4	1	protocol	1388/05/01
5	1	linux	1388/05/01
6	1	sap	1388/05/01
7	1	irankhodro	1388/05/01
8	1	enterprise	1388/05/01
9	1	scm	1388/05/01
10	2	linux	1388/05/01
11	2	windows	1388/05/01
12	2	os	1388/05/01
13	2	crm	1388/05/01
14	2	customer	1388/05/01
15	2	olap	1388/05/01
16	2	datawarehouse	1388/05/02
17	2	oltp	1388/05/02
18	2	scm	1388/05/02
19	2	architecture	1388/05/02
20	2	soa	1388/05/02
21	2	memari	1388/05/02
22	3	olap	1388/05/02
23	3	oltp	1388/05/02
24	3	datawarehouse	1388/05/02
25	3	web	1388/05/02

1. DataSet

۳. اصلاح و پاک‌سازی مورد نیاز در داده‌ها:

در این مرحله داده‌ها مورد پاک‌سازی و اصلاح قرار گرفته و عمل پیش پردازش انجام می‌شود. در این پژوهش اطلاعات تا حدودی کامل بوده و داده‌های دارای نقص، حذف شده‌است.

۴. بررسی و مطالعه حوزه‌ها و الگوریتم‌های داده‌کاوی و انتخاب آن‌ها متناسب با هدف:

با مطالعه و بررسی‌های فن‌های متنوع داده‌کاوی و با ارائه دلایلی که در پایان بند چهار آمده‌است، فن تحلیل وابستگی به عنوان یکی از فنون داده‌کاوی بر روی پایگاه داده «دی اسپیس» انتخاب شده و از میان الگوریتم‌های این حوزه، با الگوریتم معروف «آپریوری» کار می‌شود.

۵. بررسی و اجرای نرم‌افزارهای داده‌کاوی در حوزه تحلیل وابستگی:

برای به‌دست آوردن ارتباطات و وابستگی‌های داده‌ای، نرم‌افزارهای «وکا» و «مگنیوم آپوس» مورد بررسی و مطالعه قرار گرفته‌است که به طور مختصر در ادامه به تشریح دو نرم‌افزار پرداخته می‌شود:

- نرم افزار «وکا» مجموعه‌ای از الگوریتم‌های روز یادگیری ماشینی و ابزارهای پیش پردازش داده‌ها است. این نرم‌افزار به گونه‌ای طراحی شده‌است که می‌توان به سرعت، روش‌های موجود را به صورت انعطاف‌پذیری روی مجموعه‌های جدید داده‌ها آزمایش کرد. این نرم‌افزار، پشتیبانی‌های ارزشمندی را برای کل فرایند داده‌کاوی‌های تجربی فراهم می‌کند. این پشتیبانی‌ها، آماده‌سازی داده‌های ورودی، ارزیابی آماری چارچوب‌های یادگیری و نمایش گرافیکی داده‌های ورودی و نتایج یادگیری را در بر می‌گیرند. همچنین، هماهنگ با دامنه وسیع الگوریتم‌های یادگیری، این نرم‌افزار شامل ابزارهای گوناگون پیش‌پردازش داده‌ها است. این جعبه‌ابزار متنوع و جامع، از طریق واسطی متداول در دسترس است، به طوری که کاربر می‌تواند روش‌های متفاوت را در آن با یکدیگر مقایسه کند و روش‌هایی مناسب‌تر را تشخیص دهد (عطایی ۱۳۸۶).
- نرم افزار «مگنیوم آپوس»:
- «مگنیوم آپوس» از نرم‌افزارهای داده‌کاوی برای کشف ارتباطات میان داده‌ها است. این نرم‌افزار ارتباطات و وابستگی‌هایی را که بیش‌ترین یکسانی با معیارهای انتخاب شده را دارد، به‌دست می‌آورد که با مشخصات زیر است:

۱. دارای سابقه‌ی بالاتر از ده سال در خدمت رسانی به انجمن‌های داده‌کاوی
۲. طراحی شده توسط کارشناسان داده‌کاوی
۳. تبدیل داده به دانش در این سیستم
۴. امکان تشریح ارتباطات به شکل ساده
۵. امکان حذف خودکار ارتباطات نادرست و اشتباه (برای به‌دست آوردن ارتباطات صحیح و قابل اعتماد)
۶. طراحی «مگنیوم آپوس» برای تحلیل روی بانک‌های اطلاعاتی بزرگ و شایان توجه با میلیون‌ها رکورد
۷. سرعت عمل پردازش و تحلیل

حال برای اجرای نرم‌افزارها، داده‌ها به فرمت‌های ورودی مورد نیازشان تبدیل شده و پارامترهای ورودی هر دو نرم‌افزار تنظیم می‌شود. نرم‌افزار «وکا» فایلی ورودی به فرمت «ای.آر.اف.اف.» و نرم‌افزار «مگنیوم آپوس» فایلی ورودی به فرمت «آی.تی.ال» برای انجام داده‌کاوی دریافت می‌کنند.

در نرم‌افزار «وکا»، قسمت «اسوسی‌ایشن»^۳ را انتخاب کرده و الگوریتم «اپریوری» را از لیست الگوریتم‌ها گزینش می‌کنیم سپس پارامترهای مورد نیاز این الگوریتم که به صورت پیش فرض موجود است را در صورت نیاز تغییر داده با اجرای نرم‌افزار، ارتباطات میان داده‌ها کشف شده و قوانین در قالب زیر خارج می‌شود.

keyword2=knowledgediscovery 6 ==> keyword1=datamining 6 conf:(1)

keyword1=datamining 6 ==> keyword2=knowledgediscovery 6 conf:(1)

keyword40=oracle 5 ==> keyword39=dbms 5 conf:(1)

keyword39=dbms 5 ==> keyword40=oracle 5 conf:(1)

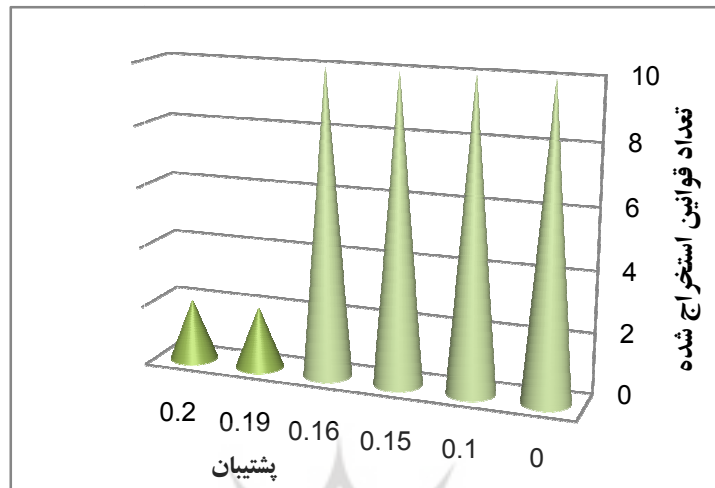
keyword41=sqlserver 5 ==> keyword39=dbms 5 conf:(1)

keyword39=dbms 5 ==> keyword41=sqlserver 5 conf:(1)

keyword41=sqlserver 5 ==> keyword40=oracle 5 conf:(1)

keyword40=oracle 5 ==> keyword41=sqlserver 5 conf:(1)

در این پژوهش به‌ازای مقادیر گوناگون پارامتر پشتیبان^۴، نرم‌افزار «وکا» اجرا شده و تعداد قوانین گوناگونی به‌دست آمده‌است که آن در شکل پنج نشان داده شده‌است ولی در پایان همان مقدار پیش فرض نرم‌افزار (پشتیبان=۰.۱۵) برای اجرا و خروج قوانین در نظر گرفته شده‌است.



شکل ۵ نمودار مقایسه تعداد قوانین خارج شده به ازای پشتیبان‌های گوناگون در «وکا»

نرم افزار «مگنیوم آپوس» که ویژه کشف ارتباطات داده‌ای است، با تعیین پارامترهای آن، قابل اجراست و قوانین در قالب زیر ارائه می‌شود.

```

apriorialgorithm -> apriorisofware
Coverage=0.424 (14); Support=0.424 (14); Strength=1.000; Lift=2.36;
Leverage=0.2443 (8.1)
apriorisofware -> apriorialgorithm
Coverage=0.424 (14); Support=0.424 (14); Strength=1.000; Lift=2.36;
Leverage=0.2443 (8.1)
backup -> recovery
Coverage=0.485 (16); Support=0.485 (16); Strength=1.000; Lift=1.94;
Leverage=0.2351 (7.8)
recovery -> backup
Coverage=0.515 (17); Support=0.485 (16); Strength=0.941; Lift=1.94;
Leverage=0.2351 (7.8)
oracle -> sqlserver
Coverage=0.333 (11); Support=0.333 (11); Strength=1.000; Lift=3.00;
Leverage=0.2222 (7.3)
sqlserver -> oracle
Coverage=0.333 (11); Support=0.333 (11); Strength=1.000; Lift=3.00;
Leverage=0.2222 (7.3)
windows -> linux
Coverage=0.364 (12); Support=0.364 (12); Strength=1.000; Lift=2.54;
Leverage=0.2204 (7.3)
linux -> windows
Coverage=0.394 (13); Support=0.364 (12); Strength=0.923; Lift=2.54;
Leverage=0.2204 (7.3)
soa -> architecture
Coverage=0.485 (16); Support=0.485 (16); Strength=1.000; Lift=1.83;
Leverage=0.2204 (7.3)

```

نتیجه داده کاوی پس از ارزیابی و حذف الگوهای تکراری برای انجام تغییر در کد «دی اسپیس»، در بانک اطلاعاتی «دی اسپیس»، جدول «ریزالتماینینگ»^۱ ذخیره می‌شود.

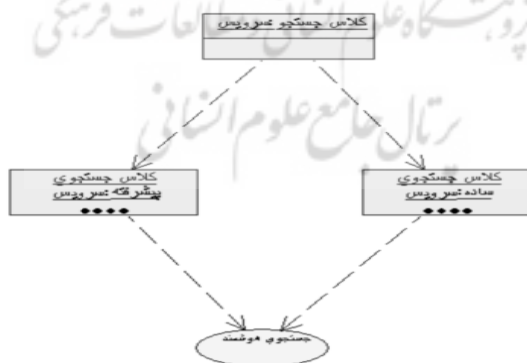
۶. به کار بستن نتایج داده کاوی در نرم افزار متن باز «دی اسپیس».

در این قسمت به طراحی سیستم مدیریت دانش با توجه به دانش‌های برگرفته از طریق داده کاوی در نرم افزار متن باز «دی اسپیس» پرداخته می‌شود. برای انجام طراحی از روش «آر.یو.پی»^۳ و ابزار «یو.ام.ال»^۴ استفاده می‌شود.

برای به کار بستن نتایج داده کاوی در «دی اسپیس» باید در طراحی ماجول جستجو تغییراتی انجام شود. در حالت پیشاز داده کاوی، «دی اسپیس» جستجوی خود را بر اساس کلمات ورودی کاربر انجام و نتیجه را نمایش می‌دهد، ولی با انجام تغییراتی در «دی اسپیس» که در زیر بیان شده است، پس از این که کاربر کلمات مورد جستجویش را وارد می‌کند، در ماجول جستجو (جستجوی ساده یا جستجوی پیشرفته)، پس از وصل به بانک اطلاعاتی به جدول «ریزالتماینینگ»^۵ که نتایج داده کاوی و قوانین کشف شده (کلمات مرتبط به هم) در آن ذخیره شده است، متصل می‌شود و اگر کلمات مرتبط با پارامترهای ورودی را پیدا کند، کلمات معادل را هم جستجو می‌کند و اگر پیدا نکند، فقط همان کلمه ورودی کاربر را جستجو کرده و نمایش می‌دهد.

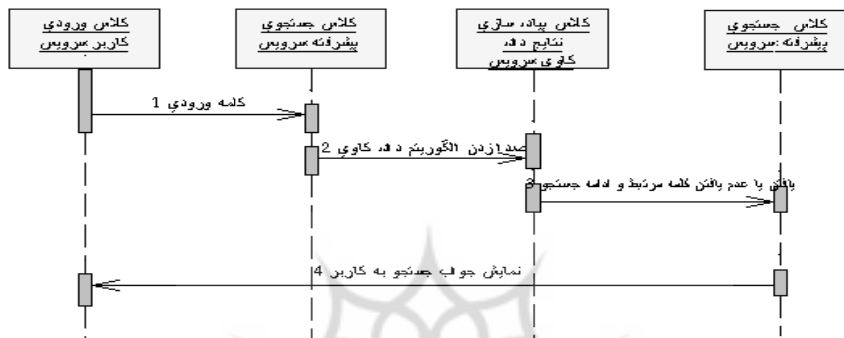
حال برای طراحی توانایی جستجوی هوشمند به تشریح نمودارهای کلاس، توالی و فرایند ریز جستجوی هوشمند می‌پردازیم.

در شکل ۶ نمودار کلاس فرایند جستجوی هوشمند نشان داده شده است. کلاس جستجو به دو کلاس جستجوی ساده و پیشرفته تقسیم شده و هر دو کلاس از توانایی جستجوی هوشمند استفاده می‌کنند.



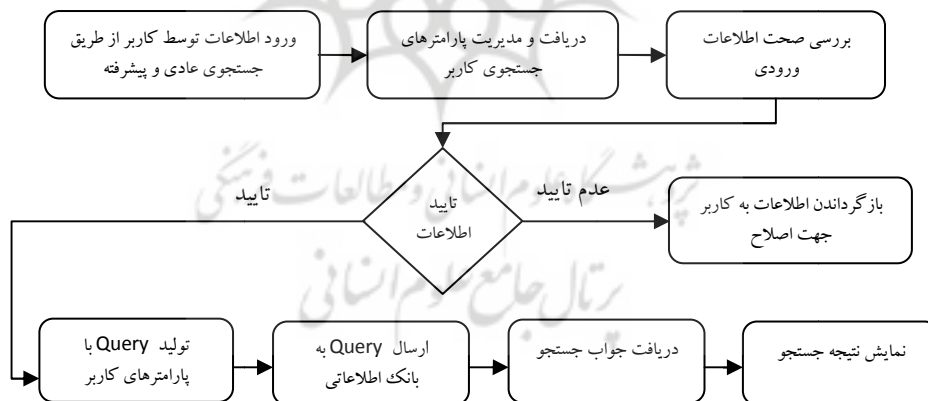
شکل ۶. نمودار کلاس فرایند جستجوی هوشمند

در شکل ۷، نمودار توالی فرایند جستجوی هوشمند ترسیم شده است. در این نمودار مراتب انجام فعالیت‌ها که در بالا بیان گردید، نشان داده می‌شود.



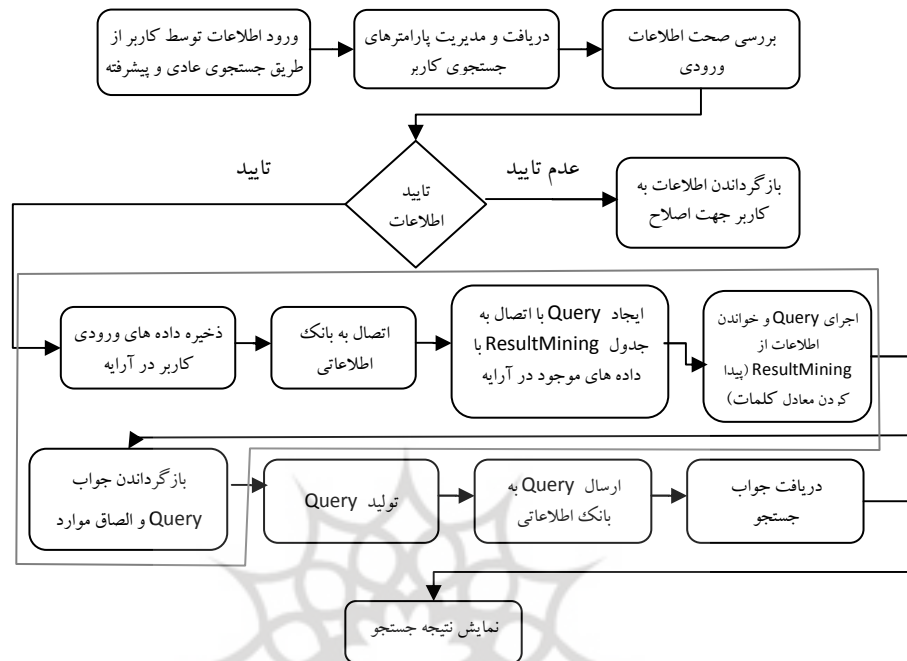
شکل ۷. نمودار توالی فرایند جستجوی هوشمند

در شکل ۸ فرایند ریز جستجو در نرم افزار متن باز «دی اسپیس» در حالت پیش از انجام تغییرات، نشان داده می‌شود.



شکل ۸. فرایند جستجو در «دی اسپیس»

در شکل ۹ فرایند ریز جستجو در نرم افزار متن باز «دی اسپیس» بعد از انجام تغییرات، نشان داده می‌شود (حلقه فرمز تغییرات جدید است).



شکل ۹ فرایند جستجوی هوشمند در «دی اسپیس»

با توجه به تغییرات بیان شده در بالا، در بخش ماجول جستجو در کد برنامه «دی اسپیس»، توابع زیر اضافه می شود^۱:

- تابع «مایننگ»: این تابع کلمات مورد نظر کاربر را دریافت کرده و در آرایه‌ای به نام «فریس»^۲ ذخیره می کند و در پایان عبارت جستجوی تولید شده را به «کواری»^۳ اصلی اضافه می کند.
- تابع «گت مایننگز»: این تابع داده‌ها را از آرایه «فریس» خوانده و با اتصال به جدول «ریزالت مایننگ»^۴ (جدول نگه‌داری نتایج داده کاوی) از بانک اطلاعاتی «پستگر»^۵، جستجوی مربوطه را آغاز می کند و نتیجه را در یک «ریزالت‌ست»^۶ نگهداری می‌کند.
- تابع «سرچ. آی.ان.دی.بی.»^۷: این تابع برای اتصال به بانک اطلاعاتی تعریف شده است.

۱. علاقه‌مندان می توانند از طریق تماس با نویسنده متن توابع را به دست آورند.

2. phrase
6. postgre

3. Query
7. resultset

4. getMinings
8. searchInDB

5. Resultmining

نمایش نتیجه توابع اضافه شده بالا و توانایی جدید جستجوی هوشمند بر روی نمونه آزمایشی در قسمت ارزیابی نشان داده می شود.

۷. ارزیابی

این نوشتار، پژوهشی فناورانه است، که بحث فناوری یعنی (انجام تغییر در ساختار و معماری نرم افزار متن باز «دی اسپیس»، برای ایجاد توانایی جدید جستجوی هوشمند بعد از انجام فرایند داده کاوی است. این پژوهش بر لبه فناوری گام بر می دارد نه بر لبه دانشی. لیکن برای ارزیابی آن، اثبات کارکرد درست «دی اسپیس» با اضافه شدن توانایی جدید (جستجوی هوشمند) به آن کافی است. در ادامه به مقایسه «دی اسپیس» در حالت های پیش و بعد از انجام تغییرات پرداخته می شود. در شکل ۱۰ «دی اسپیس» با روش گذشته، پس از ورود کلمه «نالچ» از طریق کاربر، فقط خود کلمه را جستجو کرده و ۱۸ مورد بازیابی کرده است.



شکل ۱۰ نمونه ای از کارکرد قدیمی «دی اسپیس»

در شکل ۱۱ «دی اسپیس» با روش جدید و تغییر یافته اش و با توانایی جستجوی هوشمند، پس از ورود کلمه «نالچ» توسط کاربر ابتدا در پایگاه داده اش بررسی می کند که آیا کلمه «نالچ» با کلمه دیگر ارتباطی منطقی دارد یا خیر؟ و پس از پیدا کردن کلمه مرتبط «اینفیشن»^۲ با آن، هر دو را جستجو کرده و ۲۰ مورد را بازیابی می کند. تشریح این فرایند از این قرار است:

1. knowledge
2. information

- نتایج داده کاوی نرم افزارهای «و کا» و «مگنیوم اوپوس» در پایگاه داده «دی اسپیس»، در جدول «ریزالت ماینینگ» ذخیره می شود.
- در جستجوی هوشمند ابتدا تابع «ماینینگ» صدا زده می شود: این تابع کلمات جستجو شده توسط کاربر را دریافت کرده و در آرایه‌ای ذخیره می کند. این جا کلمه «نالچ» در آرایه ذخیره می شود.
- سپس تابع «سرج. آی. ان. دی. بی.» صدا زده می شود. این تابع برای اتصال به بانک اطلاعاتی تعریف شده است.
- در مرحله بعد تابع «گت ماینینگر» صدا زده می شود. این تابع داده‌ها را از آرایه خوانده و با اتصال به جدول «ریزالت ماینینگ» (جدول نگهداری نتایج داده کاوی) از بانک اطلاعاتی «پُستگر»، جستجوی مربوطه را آغاز می کند که آیا داده‌های آرایه (مثل «نالچ»)، در جدول «ریزالت ماینینگ» موجود هستند یا خیر؟ در صورت وجود، کلمه معادل که در این مثال «اینفمیشن» است را پیدا کرده و در یک «ریزالتست» نگهداری می کند و به تابع «ماینینگ» برمی گرداند.
- سپس در تابع «ماینینگ» طبق نتایج موجود در «ریزالتست»، عبارت جستجوی جدید ساخته می شود. که در این عبارت هم کلمه «نالچ» است و هم کلمه «اینفمیشن».



شکل ۱۱. نمونه‌ای از کارکرد جدید «دی اسپیس»

در جدول ۲ نتایج چند نمونه از داده‌های ورودی به نرم افزار «دی اسپیس» در حالت‌های پیش و پس از تغییرات جستجوی هوشمند نشان داده می‌شود.

جدول ۲ مقایسه نتایج «دی اسپیس» در پیش و پس از تغییرات

نتیجه جستجو در نرم افزار Dspace بعد از تغییرات		نتیجه جستجو در نرم افزار Dspace قبل از تغییرات	کلمه ورودی توسط کاربر
تعداد	کلمه مرتبط		
۲۰ رکورد	Information	۱۸ رکورد	Knowledge
۷ رکورد	Information	۵ رکورد	Data
۱۴ رکورد	Recovery	۸ رکورد	Backup
۱۶ رکورد	Linux	۱۰ رکورد	Windows
۹ رکورد	Sqlserver	۵ رکورد	Oracle
۱۵ رکورد	Dbms	۵ رکورد	oracle
۱۹ رکورد	Architecture	۱۱ رکورد	Soa
۱۳ رکورد	Baby	۲ رکورد	child
۷ رکورد	Apriorialgorithm	۲ رکورد	apriorisoftware
۱۳ رکورد	Knowledge discovery	۸ رکورد	datamining
۱۴ رکورد	Sqlserver	۱۰ رکورد	dbms

در نمونه‌های بالا علاوه بر افزایش نتیجه جستجو، نتایج غیر مستقیم زیر نیز به دست می‌آید:

- افزایش سرعت جستجوی کاربر
- ارائه هوشمند کلمات مرتبط با کلمه ورودی کاربر
- ارائه ایده کلمات مرتبط به کاربر علوم انسانی
- جلوگیری از جستجوهای متعدد توسط کاربر درباره مفهوم
- خشنودی کاربر
- صرفه جویی در زمان

۸. نتیجه گیری

در این پژوهش ابتدا نگاهی کوتاه و گذرا بر توانایی‌ها، کارکردها و مشخصات فنی نرم افزار متن باز «دی اسپیس» انجام شد. «دی اسپیس» که سیستم مدیریتی اطلاعات و مستندات است،

پیاده‌سازی آن در سازمان‌ها به منظور پیاده‌سازی مدیریت دانش در آن سازمان است. با اجرای آن می‌توان چرخه ساده مدیریت دانش که شامل شناسایی و جمع‌آوری، ذخیره‌سازی و سازمان‌دهی، انتشار و به‌کارگیری اطلاعات است، را در سازمان پیاده کرد. در گام بعد با انجام فرایند داده‌کاوی بر روی پایگاه داده آن و با استفاده از دو ابزار داده‌کاوی «وکا» و «منگیوم.آپوس» برای استخراج دانش، این نرم‌افزار مجهز به توانایی داده‌کاوی گشته تا علاوه بر افزایش کارایی و توانایی‌های آن، افزایش سرعت جستجوی کاربر را در رسیدن به هدف و کاهش زمان صرف شده، داشته‌باشد. این نوشتار پژوهشی فناورانه است و بر لبه فناوری حرکت می‌کند و در گام پایانی، در ساختار طراحی فرایند جستجو در نرم‌افزار متن باز «دی اسپیس» نفوذ کرده‌است و این فرایند، در راستای جستجویی که از نتایج داده‌کاوی استفاده می‌کند، تغییر کرده و با این تغییر امکان جستجوی هوشمند در «دی اسپیس» فراهم شده‌است. همچنین در این پژوهش چهارچوب استاندارد داده‌کاوی بر روی پایگاه داده معرفی شده و تمامی مراحل آن بر روی مجموعه داده آزمایشی انجام شده‌است. برای ارزیابی پژوهش نیز تعداد رکوردهای بازیابی شده در «دی اسپیس» مجهز به توانایی داده‌کاوی، با تعداد رکوردهای بازیابی شده در «دی اسپیس»‌های بدون این توانایی مقایسه شده که در «دی اسپیس» تغییر داده شده تعداد رکوردها بیش‌تر است و کاربر را سریع‌تر به هدف می‌رساند.

نتایج پژوهش که برای استخراج دانش از پایگاه داده «دی اسپیس» با استفاده از فنون داده‌کاوی به‌دست می‌آید، به شرح زیر است:

پیاده‌سازی سیستم مدیریت دانش «دی اسپیس» در سازمان و فرهنگ‌سازی آن، اجرای فرایند داده‌کاوی بر روی نمونه داده آزمایشی، طراحی و پیاده‌سازی توانایی جدید جستجوی هوشمند در «دی اسپیس» که پیامدهای این توانایی جدید شامل: افزایش نتایج جستجوی کاربر، افزایش سرعت جستجوی کاربر، ارائه هوشمند کلمات مرتبط با کلمه ورودی کاربر، جلوگیری از جستجوهای متعدد توسط کاربر درباره مفهوم، خشنودی کاربر و صرفه‌جویی در زمان است.

۹. فهرست منابع

- ارسطوپور شعله. ۱۳۸۶. نرم‌افزار نظامی با کد منبع باز: نگاهی کوتاه به قابلیت‌ها و ماژول‌های نظام یکپارچه کتابخانه‌ای کوها (KOHA). مجله الکترونیکی پژوهشگاه اطلاعات و مدارک علمی ایران. دوره هفتم شماره دوم
- اکبرپور شیرازی، محسن، و حسن تویچی. ۱۳۸۷. داده‌کاوی: مفاهیم، روش‌ها و کاربردها. تهران: دانشگاه خواجه نصیرالدین طوسی
- تقوی فرد محمد تقی، و طاها منصور. ۱۳۸۶. داده‌کاوی و کاربرد آن در تصمیم‌گیری‌ها. مجله دانش مدیریت. دوره ۲۰ شماره ۷۹.
- شاه‌سمندی، پرستو. ۱۳۸۴. داده‌کاوی در مدیریت ارتباط با مشتری. ماهنامه علمی آموزشی تدبیر. سال شانزدهم شماره ۱۵۶.

شریف، عاطفه. ۱۳۸۶. نگاهی به یک نرم افزار کتابخانه‌ای متن باز از دیدگاه یک کتابدار. مجله الکترونیکی پژوهش گاه اطلاعات و مدارک علمی ایران. دوره هفتم شماره اول
عطایی، شروان، و سید محمدزاده. ۱۳۸۶. کاوش پایگاه اطلاعات خطوط راه آهن
غلامیان، محمدرضا، و مریم خواجه‌افضلی. ۱۳۸۵. مدیریت دانش در ارتباط با مشتری. تدبیر شماره ۱۸۷.

Dspace .2008. <http://www.dspace.org> . DSpace 1.5.1 Manual.

Erdmann Michael. 2001. The Data Warehouse as a Means to Support Knowledge Management.

Firestone Joseph. 1999. Knowledge Base Management Systems and The Knowledge Warehouse.

Firestone Joseph et al. 2005. Doing Knowledge Management. Learning Organization Journal, Vol. 12.

Hahsler Michael. 2001. Knowledge Management Data Warehouses and Data Mining.

Han Jiawei and Kamber Micheline. 2006. Data Mining: Concepts and Techniques. publisher:

Elsevier.

Kerschberg Larry. 2000. Knowledge Management in Heterogeneous Data Warehouse Environments.

S.Gyorodi Robert. 2001. A Comparative Study of Iterative Algorithm in Association Rules Mining . Studies in Information and control, Vol.12, No.3.

Smith MacKenzie et al . 2003 . DSpace: An Open Source Dynamic Digital Repository. D-Lib Magazine. Volume 9 Number 1, ISSN 1082-9873

Witten.H and Lan. 2005. StoneD A Bridge between Greenstone and DSpace. D-Lib Magazine. volume11Number9 . ISSN 1082-9873

Ye N. 2003. The Handbook of Data Mining. Lawrence Erlbaum Association (LEA).

Knowledge Acquisition from Database of Information Management and Documentation Softwares by DataMining Techniques

Mehdi Fesharaki

Associate professor of Malek Ashtar Industrial University, Scientific Community of Malek Ashtar Industrial University of Tehran

Hossein shirazi

Associate professor of Malek Ashtar Industrial University, Scientific Community of Malek Ashtar Industrial University of Tehran

Akram Bakhshi*

Malek Ashtar Industrial University of Tehran

Information
Sciences
& Technology

Iranian Research Institute
For Science and Technology
ISSN 1735-5206
eISSN 2008-5583
Indexed in LISA, SCOPUS & ISC
Vol.26 | No.2 | pp: 259-283
Winter 2011

Abstract: Using the high amount of data, not only isn't useful for the managers in deciding and decision making, but also will be so much problematic for the organization's managers. So it's very important and essential to manage the raw data, and changing the internal and external data to information and knowledge by using different techniques. One of the typical techniques is data mining, it can be used in databases to acquire the needed knowledge. Dspace is a software system of digital archive or a kind of repository for organizational information on the web, that will provide the possibility of saving, recovery and managing the organizational information. The purpose of this paper is to acquire information from the previous data and the history of user's researching in the database of Dspace by using data mining algorithms. By procuring the communications and dependencies of researches terms, it would be possible to use data mining for having intelligent research in Dspace software, so the efficiency of user's research will be increased. By using this idea, it would be prevented from multiple researching of every user for each concept; it would be effective for time reducing and helping users in finding the appropriate keywords of researching. So, by using a data mining based Dspace, the retrieved records will be more than the Dspace which hasn't this capability. Dspace, will make users more satisfied by helping the user in researching and preparing the intelligent ideas.

Keywords: data mining, Dspace, knowledge management, database, data management, information and document management.

* Corresponding Author: akram390@yahoo.com