استفاده از روش‌های داده‌کاوی به منظور تسهیل جستجو در موتورهای جستجوگر متنی

سولماز لطفی آذری داریان [۱]، رضا جاویدان [۲]

[۱] دانشجوی کارشناسی ارشد مهندسی فناوری اطلاعات- دانشگاه شیراز، solmaz.lotfi@gmail.com

[۲] استادیار دانشکده مهندسی کامپیوتر و فناوری اطلاعات ، دانشگاه صنعتی شیراز، reza.javidan@gmail.com

چکیده - شبکه جهانی اطلاعات روز به روز در حال گسترش است. کاربران از موتورهای جستجوی وب برای یافتن اطلاعات مورد نیاز خود بهره می‌گیرند. یکی از مهم‌ترین اهداف شبکه جهانی اطلاعات، طراحی موتورهای جستجویی است که اطلاعاتی را به کاربر نمایش دهد که با پرس‌وجوی ثبت شده او از لحاظ مفهومی مرتبط باشد. حجم زیاد اطلاعات در این شبکه جهانی و همچنین عدم توانایی کاربر در بیان دقیق نیاز اطلاعاتی او، مانع بزرگی برای دستیابی به این هدف است. کاوش فایل‌های رویدادنگاری موتورجستجو، روش‌هایی هستند که هدف آن‌ها استخراج دانش با ارزش از فایل‌های رویدادنگاری پرس‌وجو می‌باشد. در این پژوهش کلیک‌های ثبت شده کاربر در فایل‌های رویدادنگاری موتورجستجو مورد کاوش قرار گرفته است تا الگوی رفتاری کاربران شناسایی شده و بتوان الگوریتمی جهت بهبود دقت نتایج حاصل از پرس‌وجوی کاربر پیشنهاد نمود. در این تحقیق از نرم‌افزار داده‌کاوی رپیدماینر و روش خوشه‌بندی K-Means بهره گرفته شده است. تحلیل صورت گرفته روی حدود سه میلیون رکورد از داده‌های واقعی یک موتور جستجوی تجاری به کار گرفته شده است. با به کارگیری تکنیک‌های خوشه‌بندی قادر به ایجاد خوشه‌های حاوی پرس‌وجوهای مشابه خواهیم بود. با استفاده از این خوشه‌ها، روش‌هایی برای پیشنهاد نتایج بهتر به پرس‌وجوی کاربر جهت بهبود لیست نتایج، ارائه شده است.

کلید واژه - روش خوشه‌بندی K-Means، شبکه جهانی اطلاعات، کاوش فایل‌های رویدادنگاری، موتور جستجو

۱- مقدمه

جستجو یکی از رایج‌ترین عملیات در اینترنت است. میزان اطلاعات در وب با سرعت بسیار زیادی در حال افزایش است. از طریق استفاده از موتورهای جستجو، کاربران قادر به جستجوی اطلاعات مورد نیاز خود می‌باشند. رشد بسیار سریع اسناد در اینترنت، نمایش اسناد مرتبط با پرس‌وجوی انجام شده را دشوارتر ساخته است. موتورهای جستجوی قدیمی در پاسخگویی به نیازهای کاربران دارای برخی مشکلات هستند. زیرا این موتورهای جستجوی قدیمی نمی‌توانند نیاز روز به روز در حال افزایش افراد برای بازیابی اطلاعات به صورت شخصی‌سازی‌شده را برآورده نمایند[۱و۲].

تکنیک‌های کاوش در وب، می‌توانند برای شخصی‌سازی استفاده کاربران از وب به کار روند. برای مثال می‌توان رفتار کاربر را از طریق مقایسه الگوی پیمایش فعلی وی با الگوهای پیمایش استخراج شده از فایل‌های ثبت‌شده، به صورت بلادرنگ پیش‌بینی کرد. هم‌اکنون در بسیاری از جنبه‌های بررسی رفتار کاربران وب، تحلیل‌گران با استفاده از تکنیک‌های داده‌کاوی سعی بر آن دارند تا الگوی رفتاری کاربران را شناسایی کرده و به پیش‌بینی رفتار آنان دست یابند. از طرفی دیگر علاوه بر افزایش حجم داده‌ها و لزوم استفاده از روش‌هایی به منظور کشف دانش پنهان از این داده‌ها، مانند روش‌های داده‌کاوی، قدرت پردازش در رایانه‌ها نیز افزایش یافته است که این امر نیز لزوم استفاده از روش‌های پیشرفته و توسعه‌یافته داده‌کاوی را بر روی داده‌های جستجوی کاربران بیش از پیش ضروری می‌سازد[۳].

در این پژوهش تلاش شده است تا با ترکیب دو روش محتواکاوی و کاربردکاوی وب، از فایل‌های رویدادنگاری ثبت شده‌ی مربوط به جستجوی کاربران بهره برده و روی پرس‌وجوهای ثبت شده، عملیات متن‌کاوی انجام داد تا بتوان الگوهای رفتاری کاربران در جستجوی متنی را شناسایی نمود. ساختار ادامه مقاله به این صورت است : در بخش ۲ پژوهش‌های مرتبط در این زمینه بررسی شده است. در بخش ۳ روش پیشنهادی توضیح داده شده است. در بخش ۴ نتایج حاصل از روش بیان شده و نهایتا در بخش ۵ نتیجه‌گیری ارائه شده است.

۲- مرور کارهای مرتبط

داده‌کاوی به فرآیند تجزیه و تحلیل پایگاه داده‌های بزرگ به منظور یافتن الگوهایی مفید، گفته می‌شود. روش‌های داده‌کاوی شامل دو دسته عمده روش‌های پیش‌بینی کننده و روش‌های تشریحی می‌باشند. هدف این روش‌ها جداسازی و قرار دادن اشیا در تعدادی از کلاس‌هاست که در روش‌های پیش‌بینی کننده مانند روش رده‌بندی، این کلاس‌ها از قبل وجود دارند و اشیا بر اساس ویژگی‌هایشان در این کلاس‌ها قرار داده می‌شوند. یعنی این روش‌ها جهت پیشگویی متغیر هدف به کار گرفته می‌شوند. اما در روش‌های تشریحی نظیر خوشه‌بندی، این کلاس‌ها از قبل وجود ندارند و طی فرآیند و با توجه به ویژگی‌های اشیا ایجاد می‌شوند.

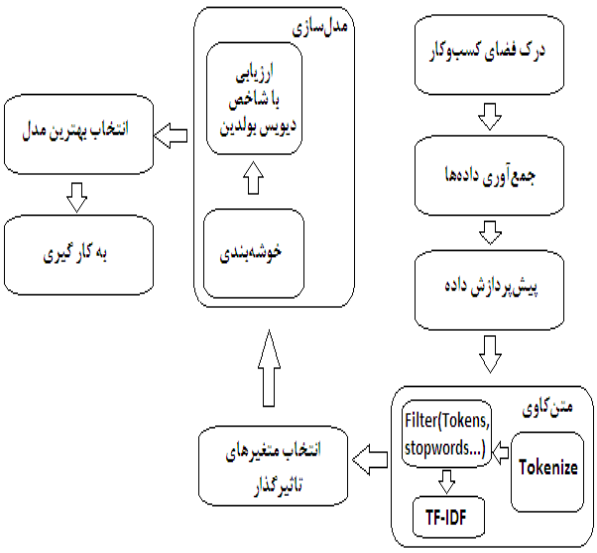
کاوش فایل‌های رویدادنگاری پرس‌وجو به کلیه تکنیک‌هایی که هدفشان کشف الگوهایی از فایل‌های رویدادنگاری پرس‌وجوی موتور جستجو است، گفته می‌شود. فایل‌های رویدادنگاری، ردیابی اطلاعات حاصل از ارتباط بین کاربران و موتور جستجو می‌باشد. در این فایل‌ها، پرس‌وجوهای ارسال شده به موتور جستجو، زمان ثبت پرس‌وجو، آدرس صفحات نتیجه حاصل از پرس‌وجو که توسط موتورجستجو برگردانده می‌شود و همچنین رتبه صفحات مشاهده شده کاربر ثبت می‌شود. این امکان وجود دارد که از اطلاعات حاصل از فایل‌های رویدادنگاری، نشست‌های مربوط به جستجو و مجموعه فعالیت‌های کاربر در بازه زمانی محدود حاصل گردد. پرس‌وجوها همیشه به تنهایی برای درک منظور کاربر از آن کافی نیستند. تحلیل‌گران از معیارهای دیگری نظیر تعداد کلیک‌های کاربر و زمان سپری شده کاربر در هر سایت (نتیجه پرس‌وجو) برای تحلیل رفتار کاربران استفاده می‌کنند. تکنیک‌های مدرن داده‌کاوی که سهم مهمی از حوزه علوم اطلاعاتی را به خود اختصاص داده‌اند می‌توانند به منظور ساخت مدل‌هایی برای پیش‌بینی رفتار جستجوی کاربران در موتورهای جستجوی وب به کار برده شوند.

پژوهش‌های متعددی در سال‌های گذشته در زمینه‌ی کاوش موتورهای جستجو توسط نویسندگان مختلف صورت گرفته است. در سال ۲۰۰۲، Yan Li و همکارانش تحقیقات بسیاری روی موتورهای جستجوی هوشمند بر پایه کاوش انجام داده‌اند. در این مقاله ابتدا فرضیه و مفهوم موتورهای جستجو بیان شده و سپس روش‌های دسته بندی و برنامه‌های کاربردی وب کاوی نمایش داده شده است. در نتیجه‌گیری مقاله نیز برنامه‌های کاربردی متد وب کاوی در موتورهای جستجوی هوشمند با جزئیات مورد بررسی و بحث قرار گرفته است[۴]. Foncesa و همکارانش در سال ۲۰۰۷ یک روش بسط پرس‌وجو بر پایه مفهوم ارائه نمودند که موجب ابهام‌زدایی پرس‌وجوهای ثبت شده در موتور جستجو می‌گردد. مفاهیم از طریق تحلیل و تعیین چرخه‌هایی در نوع خاصی از گراف رابطه استخراج می‌شوند. این گراف یک گراف هدایت شده است که از روابط پرس‌وجوهای کاوش شده با استفاده از قوانین وابستگی شناخته شده است. سپس مفاهیم مرتبط با پرس‌وجوی جاری که کاربر آن را به عنوان مرتبط‌ترین مفهوم به آن پرس‌وجو انتخاب نموده است، به وی نشان داده می‌شود. این مفهوم جهت بسط پرس‌وجوی اصلی استفاده شده و به جای آن روی پرس‌وجوی بسط داده شده اقدام انجام می‌شود.[۵]. Chirita و همکارانش در سال ۲۰۰۷ روشی جهت بهبود پرس‌وجوهای وب از طریق بسط آن‌ها با کلماتی که از مخزن داده‌های شخصی هر کاربر جمع‌آوری شده بود پیشنهاد نمودند. لذا به طور ضمنی خروجی حاصل از جستجو را شخصی‌سازی نمودند. نویسندگان پنج روش عمده جهت تولید کلمات کلیدی اضافی پرس‌وجو ارائه نمودند. تحلیل وسیع تجربی نشان می‌دهد که برخی از این روش‌ها خصوصا روی پرس‌وجوهای مبهم به خوبی اجرا می‌شوند که افزایش بسیاری در کیفیت رتبه‌دهی نتایج را حاصل می‌شوند[۶]. Cui و همکارانش در سال ۲۰۰۹ روشی که ارتباط بین کلمات در اسناد کلیک شده را به کار می‌گیرد ارائه نمودند. استفاده از داده‌های حاصل از کلیک به علت این فرضیه است که اسناد کلیک شده به پرس‌وجو مرتبط‌تر می‌باشند. ارزیابی روش با ۳۰ پرس‌وجویی که به طور تصادفی از ترکیب فایل‌های رویدادنگاری Encarta، مجموعه داده پرس‌وجوی TREC و یک مجموعه کوچک از پرس‌وجوهای ارزیابی شده به طور دستی استخراج شده است صورت پذیرفته است[۷]. Braglia و همکارانش در سال ۲۰۰۹ مدل جدیدی به نام مساله میان‌بر جستجو (The Search Shortcut Problem) ارائه نمودند که شامل توصیه پرس‌وجوهای موفقی است که برای نیاز اطلاعاتی مشابه موجب رضایت کاربران دیگر در گذشته شده است. نویسندگان به طور دقیقی روش‌های پالایش را برای این

مساله اجرا نمودند. روش پیشنهاد پرس‌وجوی ارائه شده روی دو فایل رویدادنگاری بزرگ مورد ارزیابی قرار گرفته است. تکنیک‌های مختلفی برای تحلیل و استخراج اطلاعات از فایل‌های رویدادنگاری به علاوه معیارها و تکنیک‌های جدید برای اندازه‌گیری موثر بودن پیشنهاد پرس‌وجو ارائه و مورد سنجش قرار گرفته است[۸]. پیریا و پویتی در مقاله‌ای در سال ۲۰۱۲ تحت عنوان طراحی و توسعه‌ی موتورهای جستجوی شخصی به بیان این مطالب می‌پردازند که موتور جستجوی شخصی یک سیستم ترکیبی است که برای جستجوی شخصی و استدلال پیشنهاد به کاربر بر اساس پروفایل او می‌باشد. سیستم ترکیبی شامل هر دو پایگاه دانش جهانی و اطلاعات محلی کاربر می‌باشد. پروفایل برای هر کاربر برای جمع‌آوری علایق و ارتباطات او ایجاد می‌شود. این سیستم همچنین بر اساس آنچه کاربر در پروفایل خود کاندید کرده است رفتار او را می‌آموزد. تا نتایج جستجو را برای او بر اساس ویژگی موضوع شخصی‌سازی کند[۹]. آقایان کاووسی و مشیری در اولین کنفرانس بین المللی فناوری اطلاعات و دانش در سال ۱۳۸۲، مقاله‌ای با عنوان استخراج مدل رفتاری موتورهای جستجوی مورد استفاده در یک فراجویشگر هوشمند ارائه دادند که درآن مقاله یک فراجویشگر ارائه شده است، که علایق کاربر را دریافت می‌کند و با توجه به آن پرس‌وجوهای متناسب با چند موتور جستجو تنظیم می‌گردد و به موتورهای جستجو ارسال می‌شود. سپس نتایج بازگشتی از موتورهای جستجو پالایش می‌شوند و پس از تعیین اولویت در اختیار کاربر قرار می‌گیرند. در این مقاله تلاش شده است تا بخشی از معماری یک عامل هوشمند سفارشی مورد بررسی قرار گیرد که قادر است به مرور زمان و با توجه به بازخوردی که از کاربر دریافت می‌کند مدل رفتاری موتورهای جستجو را در خوشه‌های موضوعی مختلف استخراج نماید.[۱۰]. آقایان حسنپور و دهقانی و کنشلو در چهارمین کنفرانس داده‌کاوی در سال ۱۳۸۹، مقاله‌ای تحت عنوان مرتب‌سازی نتایج موتورهای جستجو بر اساس تاریخچه رفتاری کاربر ارائه دادند که در آن راهکار جدیدی برای موتورهای جستجو ارائه شده است. دراین راهکار به منظور شناسایی کاربران صفحاتی که پس از جستجو مورد توجه کاربر قرار می‌گیرند پردازش معنایی شده و به عنوان تاریخچه ملاقات‌های قبلی کاربر مورد توجه قرار می‌گیرد، که در شکل‌گیری الگوی رفتاری وی موثر می‌باشد[۱۱]. آقایان امیر پناه، امید پناه و امین پناه در دومین همایش فناوری اطلاعات، حال، آینده در سال ۱۳۹۰، مقاله‌ای با عنوان بررسی وب‌کاوی و پیشنهاد روشی جدید برای افزایش سرعت جستجو در وب ارائه دادند. در این مقاله، روش جدیدی برای جستجوی اطلاعات موجود در وب به وسیله موتورهای جستجو پیشنهاد شده است. این روش با ترکیب دو روش موجود و به کاربرده شده در دو موتور جستجوی مشهور سبب کاهش زمان جستجو و بالا بردن کارائی جستجو می‌شود[۱۲].

۳- روش پیشنهادی

معماری روش پیشنهادی و روند ارائه روش در شکل ۱ نشان داده شده است و قسمت‌های بعد بر اساس معماری روش توضیح داده شده است.



شکل ۱ : معماری روش پیشنهادی

در این پژوهش از روش فرآیندی کریسپ که یک روش اجرای پروژه‌های داده‌کاوی است، استفاده شده است. مرحله اول فرآیند کریسپ، درک فضای کسب‌وکار و تعیین اهداف مورد انتظار از پروژه است. هدف از این پژوهش شناسایی الگوهای جستجوی کاربران و تشخیص علایق آن‌هاست. گام بعدی استخراج مجموعه داده هدف است. مرحله سوم از این روش فرآیندی نیز شامل فعالیت‌های مرتبط با پیش پردازش داده می‌باشد. پس از آن روش‌های داده‌کاوی جهت کشف الگو به کار گرفته شده و سپس این روش‌ها تحلیل و ارزیابی می‌گردند. در

نهایت نیز برنامه‌ای جهت به‌کارگیری مدل در دنیای واقعی تدوین می‌شود.

۳-۱- جمع‌آوری و پردازش داده

در این پژوهش داده مورد استفاده، فایل رویدادنگاری پرس‌وجوهای ثبت شده توسط ۶۵۰۰۰۰ کاربر در بازه زمانی سه ماه در موتور جستجوی AOL می‌باشد. داده‌ها کاملا واقعی بوده و توسط کاربران واقعی تولید شده‌اند. این مجموعه داده شامل ستون‌های شماره کاربر، عبارت پرس‌وجو، آدرس سایتی که به عنوان نتیجه توسط موتور جستجو برگردانده شده، رتبه مربوط به هر سایت و زمان ثبت پرس‌وجو می‌باشد.

داده‌های موجود عموما دارای اشتباهات و خطاهایی می‌باشند. پس از جمع‌آوری داده، باید اشتباهات موجود در داده‌ها از جمله وجود داده‌های دورافتاده (Noise) و یا داده‌های از دست رفته (Missing Values) را در مجموعه داده بررسی نموده و این اشتباهات را رفع کرد. داده‌های به کار رفته در این پژوهش نیز دارای مقادیر از دست رفته برای برخی از متغیرها بودند. برای مقادیر از دست رفته داده‌های مورد استفاده در این پژوهش به دلیل حجم بسیار زیاد داده‌ها از روش حذف استفاده شده است. داده‌هایی که تکراری بودند و عینا تکرار شده بودند از مجموعه داده مورد نظر حذف شده است.

یکی دیگر از راه‌های پیش‌پردازش داده ایجاد ویژگی می‌باشد. در برخی موارد می‌توان با استفاده از متغیرهای موجود یک متغیر جدید تولید نمود. به عنوان مثال می‌توان مقدار متغیر بدهی افراد تقسیم بر درآمد آن‌ها را به عنوان یک متغیر جدید با عنوان نسبت بدهی به درآمد معرفی کرد. در این پژوهش به کمک ابزارهای سئو (Search Engine Optimization)، سه متغیر "عنوان"، "توضیحات" و "کلمات کلیدی" بر اساس متغیر آدرس صفحه نتیجه پرس‌وجو، به مجموعه داده افزوده شده است. در واقع این سه متغیر از متاتگ‌های سایت‌های مربوطه استخراج شده است. برای آدرس صفحاتی که محتوای آن‌ها اکنون در اینترنت وجود ندارد این سه ستون خالی می‌باشد. افزودن این سه متغیر به مجموعه داده مورد نظر مزایایی دارد : با این کار صفحات قدیمی که محتوای آن‌ها وجود ندارند مشخص می‌شوند و می‌توان در مرحله بعد آن‌ها را فیلتر نموده و در مدل‌سازی لحاظ نکرد. دوم اینکه استفاده از ستون‌های اضافه شده به مجموعه داده موجب نوآوری در تحقیق شده و همچنین سبب افزایش دقت مدل‌سازی نهایی می‌شود.

۳-۲- فرآیند متن‌کاوی

در این پژوهش همچنین بعد از پاک‌سازی داده‌های متنی روی آن‌ها فرآیند متن‌کاوی صورت گرفته است. به این صورت که تمامی اسناد متنی موجود در رکوردهای داده به صورت تکه تکه شده و کلمات به دست آمده از متن‌ها به صورت متغیرهای مجزا درآمدند. برای استخراج همه کلمات یک متن، یک فرآیند Tokenization لازم است که در آن یک متن با حذف کردن همه علائم نقطه گذاری و جایگزین کردن تمامی کاراکترهای غیر متنی با یک کاراکتر فضای خالی، تبدیل به جریانی از کلمات می‌شود و سپس از این نمایش برای پردازش‌های بعدی استفاده می‌شود. برای فیلتر کردن مجموعه کلمات حاصل از متن‌کاوی، می‌توان از گره Filter Tokens استفاده کرد. در این پژوهش نیز کلماتی که کوچکتر از چهار حرف و بیشتر از بیست و پنج حرف هستند در نظر گرفته نشده‌اند، به این خاطر که کلمات زائد و بی‌تاثیر به عنوان متغیر در مدل‌سازی به کار گرفته نشوند.

یکی دیگر از کارهایی که می‌توان در متن‌کاوی به کار برد، تبدیل تمامی حروف موجود در کلمات به حروف کوچک و یا بزرگ است. این کار به این علت انجام می‌شود که اگر کلمه‌ای وجود داشته باشد که چندین بار به صورت یکسان اما با تفاوت در حروف کوچک و بزرگ تکرار شده است، در عملیات متن‌کاوی یک بار در نظر گرفته شود. برای این کار باید از گره Transform Cases استفاده نمود که در این پژوهش تمامی حروف به حروف کوچک تبدیل شده‌اند.

همچنین می‌توان کلمات توقف (Stop Words) را از مجموعه کلمات حاصل از فرآیند متن‌کاوی حذف نمود. کلمات توقف به کلماتی گفته می‌شوند که به تنهایی اطلاعات مفید و متمایز کننده‌ای را در اختیار مدل فضای بردار قرار نمی‌دهند، مانند حروف ربط، اشاره و غیره. برای این کار می‌توان از گره Filter Stopwords(English) استفاده نمود.

ریشه‌یابی (Stemming) نیز از جمله عملیات رایج در فرآیند متن‌کاوی می‌باشد. ریشه‌یابی کلمات را به عبارات پایه آن‌ها تبدیل می‌نماید. برای مثال کلمات "روش‌هایم"، "روشی" و "روشمند" هر سه از ریشه "روش" هستند. روش‌های ریشه‌یابی

اغلب مبتنی بر زبان هستند. الگوریتم‌های Porter و Snowball از الگوریتم‌های ریشه‌یابی محبوب برای زبان انگلیسی می‌باشند که در این پژوهش از الگوریتم Snowball استفاده شده است.

در این پژوهش مقادیر هر متغیر نسبت به اسناد متنی با استفاده از الگوی وزن‌دهی بسامد جمله در مقابل بسامد سند یا TF-IDF (Term Frequency Inverse Document Frequency) محاسبه شده‌اند. این روش باعث می‌شود تا عباراتی که در یک مجموعه کوچک از اسناد وجود دارند وزن بیشتری پیدا کنند و آن‌هایی که در اغلب اسناد موجودند وزن کمتری داشته باشند. این مجموعه داده بعد از اعمال عملیات ذکر شده روی آن دارای ۲۰۰ متغیر (کلمات استخراج شده) می‌باشد.

همچنین در متن کاوی پژوهش حاضر، عباراتی که در کمتر از سه در صد و بیشتر از پنجاه درصد از اسناد تکرار شده‌اند هرس شده‌اند تا در کلمات خروجی حاصل از متن‌کاوی این کلمات لحاظ نشده و موجب دقت در پردازش‌های بعدی شود.

۳-۳- مدل‌سازی

مرحله بعد از فرآیند کریسپ، انتخاب روش مناسب برای مدل‌سازی است. در این تحقیق از روش‌های خوشه‌بندی برای مدل‌سازی استفاده شده است. در واقع با استفاده از روش K-Means و DBSCAN روی پرس‌وجوهای کاربران بر اساس عنوان، توضیحات و کلمات کلیدی خوشه‌بندی صورت گرفته است.

روش K-Means دارای پارامترهای K یا تعداد خوشه، Measure Type یا نوع مقیاس و Divergence یا واگرایی است. روش بر پایه چگالی (DBSCAN) نیز دارای پارامترهای Eps یا شعاع و MinPoints یا حداقل تعداد نقاط می‌باشد.

پیش از ساخت مدل نهایی نیاز است تا مدلی بر روی داده‌ها پیاده سازی شده و نتایج آن به دقت بررسی گردد، زیرا پیش از مدل‌سازی مشخص نیست که مدل نتیجه خوبی را ارائه خواهد داد یا خیر. پس از ساخت مدل اولیه و حصول نتایج مثبت باید قدم بعدی را برداشت. در حقیقت ابتدا توسط حجم کمی از داده‌ها مدل‌سازی انجام می‌شود و مدل‌هایی که بیشترین دقت را نتیجه بدهند به منظور بررسی بیشتر وارد مرحله بعدی شده‌اند. در این پژوهش در مرحله درست کردن مدل اولیه روش‌های DBSCAN و K-Means با تکرارهای مختلف پیاده سازی شدند.

روش K-Means از روش‌های خوشه‌بندی داده‌ها در داده‌کاوی می‌باشد. این روش علیرغم سادگی آن یک روش پایه برای بسیاری از روش‌های خوشه‌بندی دیگر (مانند خوشه‌بندی فازی) محسوب می‌شود. این روش روشی انحصاری و مسطح محسوب می‌شود. گام‌های الگوریتم K-Means به شرح زیر است:

۱-تعیین K یا تعداد خوشه‌ها

۲-تعیین مراکز خوشه تصادفی به تعداد K به صورت تصادفی

۳-محاسبه فاصله تمامی نمونه‌ها با مراکز تعیین شده اولیه

۴-تخصیص نمونه‌ها به مراکز نزدیک‌تر

۵-محاسبه مرکز هندسی خوشه‌های تشکیل‌شده

۶-تکرار مراحل ۳ تا ۵ تا رسیدن به همگرایی (زمانی که مراکز خوشه‌ها ثابت بماند)

روش بعدی به کار گرفته شده در این پژوهش، روش DBSCAN می‌باشد. در این روش ابتدا باید دو شاخص را تعیین نمود : شاخص Eps یا شعاع و شاخص MinPoints یا حداقل تعداد نقاط. این الگوریتم به این صورت عمل می‌کند که به ازای تک تک نمونه‌ها دایره‌ای به شعاع تعیین شده ترسیم می‌کند و تعداد نمونه‌های موجود در آن دایره (خوشه) را بررسی می‌کند. اگر تعداد نمونه‌ها بیشتر از حداقل تعداد نمونه تعیین شده در ابتدا بود، آن نمونه یک Core Point یا نمونه مرکزی محسوب می‌شود، اگر تعداد نمونه‌ها اندکی کمتر از حداقل تعداد نمونه تعیین شده بود آن نمونه یک Border Point یا نمونه لب مرزی خواهد بود و نهایتا اگر تعداد نمونه‌ها در آن خوشه بسیار کمتر از حداقل تعداد نمونه‌ها بود آن نمونه به عنوان Noise Point یا نمونه دور افتاده شناسایی می‌شود.

پس از ساخت مدل اولیه، نوبت به ساخت مدل نهایی می‌رسد. مدل نهایی با استفاده از روش(های) انتخاب شده از مرحله ساخت مدل اولیه، ساخته و اجرا می‌شود. در این مرحله همچنین باید پارامترهای مدل را به منظور رسیدن به مدل بهینه‌تر تغییر داده و نتایج را بررسی نمود. روش انتخاب شده

پارامترهای زیادی دارد و با تغییر آنها دقت مدل نیز تغییر خواهد کرد. تمامی مقادیر ممکن این پارامترها (تا حد امکان) باید تنظیم شده و نتیجه بررسی گردد. به عنوان مثال در روش K-Means می‌توان تعداد خوشه‌ها، نوع مقیاس و واگرایی را تغییر داد.

۳-۴- ارزیابی خوشه‌بندی

برای روش‌های تشریحی داده‌کاوی نظیر خوشه‌بندی، نتایج حاصل از اعمال الگوریتم‌های خوشه‌بندی روی یک مجموعه داده با توجه به انتخاب‌های پارامترهای الگوریتم‌ها می‌تواند بسیار متفاوت از یکدیگر باشد. هدف از اعتبارسنجی خوشه‌ها یافتن خوشه‌هایی است که بهترین تناسب با داده‌های مورد را داشته باشد. دو معیار پایه اندازه‌گیری پیشنهاد شده برای ارزیابی و انتخاب خوشه‌های بهینه عبارتند از[۱۳] :

۱- معیار چسبندگی (Cohesion) یا همان فواصل درون خوشه‌ای : داده‌های متعلق به یک خوشه بایستی تا حد ممکن به یکدیگر نزدیک باشند. معیار رایج برای تعیین میزان چسبندگی داده‌ها، واریانس داده‌هاست.

۲- معیار جدایش (Separation) یا فواصل بین خوشه‌ای : خوشه‌ها خود بایستی به اندازه کافی از یکدیگر جدا باشند. سه راه برای سنجش میزان جدایی خوشه‌ها مورد استفاده قرار می‌گیرد که عبارتند از : فاصله بین نزدیک‌ترین داده‌ها از دو خوشه، فاصله بین دورترین داده‌ها از دو خوشه، و فاصله بین مراکز خوشه‌ها.

همچنین روش‌های ارزیابی، خوشه‌های حاصل از خوشه‌بندی را به صورت سه دسته تقسیم می‌کنند که عبارتند از :

۱-معیارهای خروجی (External Criteria)

۲-معیارهای درونی (Internal Criteria)

۳-معیارهای نسبی (Relative Criteria)

هم معیارهای خروجی و هم معیارهای درونی بر مبنای روش‌های آماری عمل می‌کنند و پیچیدگی محاسباتی بالایی دارند. معیارهای خروجی عمل ارزیابی خوشه‌ها را با استفاده از بینش خاص کاربران انجام می‌دهند. معیارهای درونی، عمل ارزیابی خوشه‌ها را با استفاده از مقادیری که از خوشه‌ها و نمای آن‌ها محاسبه می‌شود، انجام می‌دهند.

پایه معیارهای نسبی، مقایسه بین شماهای خوشه‌بندی مختلف(الگوریتم به علاوه پارامترهای آن) است. یک و یا چندین روش مختلف خوشه‌بندی چندین بار با پارامترهای مختلف روی یک مجموعه داده اجرا می‌شوند و بهترین شمای خوشه‌بندی از بین تمام شماها انتخاب می‌شود. در این روش مبنای مقایسه، شاخص‌های اعتبارسنجی (Validity Index) هستند. شاخص‌های ارزیابی بسیار متنوعی پیشنهاد شده‌اند که در این پژوهش از شاخص ارزیابی دیویس بولدین استفاده شده است. این شاخص[۱۴] از معیار شباهت بین دو خوشه(R_{ij}) استفاده می‌کند که بر اساس پراکندگی یک خوشه(s_i) و عدم شباهت بین دو خوشه(d_{ij}) تعریف می‌شود. شباهت بین دو خوشه را می‌توان به صورت‌های مختلفی تعریف کرد ولی بایستی شرایط زیر را دارا باشد:

- $R_{ij} \geq 0$
- $R_{ij} = R_{ji}$
- اگر s_i و s_j هر دو برابر صفر باشند آنگاه R_{ij} نیز برابر صفر باشد.
- اگر $s_j > s_k$ و $d_{ij} = d_{ik}$ آنگاه $R_{ij} > R_{ik}$
- اگر $s_j = s_k$ و $d_{ij} < d_{ik}$ آنگاه $R_{ij} > R_{ik}$

معمولا شباهت بین دو خوشه به صورت زیر تعریف می‌شود :

$$R_{ij}=\frac{si+sj}{dij} \qquad (۱)$$

که در آن d_{ij} و s_i با روابط زیر محاسبه می‌شوند :

$$d_{ij}=d(v_i, v_j) \qquad (۲)$$

$$si=\frac{1}{|ci|}\sum_{x\in ci} d(x, vi) \qquad (۳)$$

با توجه به مطالب بیان شده و تعریف شباهت بین دو خوشه شاخص دیویس بولدین به صورت زیر تعریف می‌شود :

$$DB=\frac{1}{nc}\sum_{i=1}^{nc} Ri \qquad (۴)$$

که R_i در آن به صورت زیر محاسبه می‌شود :

$$R_i= \max(R_{ij}),i=1\ldots n_c \qquad (۵)$$

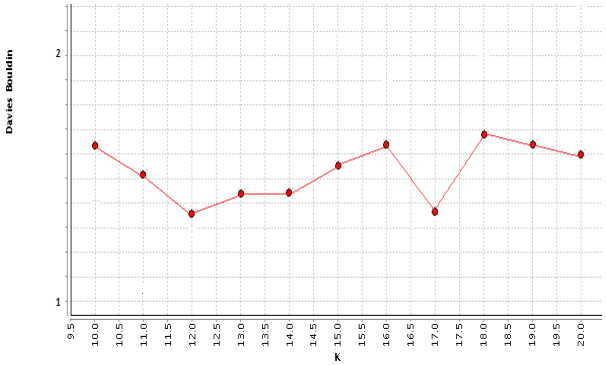
این شاخص در واقع میانگین شباهت بین هر خوشه با شبیه‌ترین خوشه به آن را محاسبه می‌کند.

۴-ارزیابی نتایج

به علت اینکه شاخص دیویس بولدین در نرم‌افزار رپیدماینر تنها برای ارزیابی روش K-Means به کار می‌روند و همچنین به دلیل اینکه مدل‌سازی اولیه با روش DBSCAN بسیار زمان‌بر بوده و روی حجم زیاد از داده‌ها امکان پذیر نمی‌باشد، در مدل‌سازی نهایی از روش DBSCAN استفاده نشده است. زیرا هدف ما استفاده از حجم نسبتا زیاد از داده به منظور افزایش دقت مدل‌سازی بوده است.

مدل‌سازی‌های اولیه انجام شده با روش K-Means با استفاده از شاخص دیویس بولدین ارزیابی شدند و از بین آن‌ها مدل با تعداد خوشه ۱۲ و نوع مقیاس "واگرایی برگمن" و نوع واگرایی " فاصله اقلیدسی" بهینه انتخاب شدند و روی مدل‌سازی نهایی لحاظ شدند.

شکل ۲ نمودار خطی حاصل از مقایسه شاخص دیویس بولدین در نرم‌افزار رپیدماینر با اعمال تعداد خوشه‌های متفاوت با استفاده از روش K-Means را نشان می‌دهد.



شکل ۲ : نمودار خطی مقایسه شاخص دیویس بولدین برای K های مختلف

هر چه شاخص دیویس بولدین کمتر باشد مدل‌سازی انجام شده بهینه‌تر می‌باشد. لذا تعداد خوشه ۱۲ و ۱۷ که شاخص دیویس بولدین آن‌ها از همه کمتر شدند از سایر تعداد خوشه‌ها نتیجه بهتری داشتند. در این پژوهش روی مدل‌سازی نهایی تعداد خوشه ۱۲ لحاظ شده است.

مدل‌سازی نهایی با روش K-Means و با در نظر گرفتن تعداد خوشه ۱۲ و واگرایی برگمن روی مجموعه داده حاصل از متن‌کاوی انجام شده است. این مجموعه داده خوشه‌بندی شده حاصل از متن کاوی(شامل ستونهای شماره ردیف، کلمات حاصل از متن‌کاوی، شماره خوشه) با مجموعه داده اولیه(شامل ستونهای شماره ردیف، عنوان پرس‌وجو و نتیجه حاصل از پرس‌وجو) بر اساس شماره ردیف، پیوند (Join) زده شده است. این شماره ردیف که یک شماره منحصر به فرد است و در فرآیند متن کاوی نیز لحاظ شده بود موجب می‌شود که مشخص گردد هر پرس‌وجوی ارسال شده به موتور جستجو به کدام خوشه تعلق دارد تا بتوان در پردازش‌های بعدی، برای هر خوشه تعداد کلیک کاربران روی هر سایت را محاسبه نموده و بر اساس این تعداد رتبه سایت در خوشه مورد نظر حاصل شود تا بتوان در پیشنهاد به کاربر از آن استفاده نمود. خوشه‌های حاصل از مدل‌سازی نهایی مورد تجزیه و تحلیل قرار گرفتند تا برچسبی به هر کدام از خوشه‌ها تخصیص داده شود. این برچسب‌ها به شرح زیر می‌باشند:

خوشه ۱ : خدمات عمومی آنلاین نظیر مشاهده آب‌وهوا، تقویم، مشاهده اماکن تفریحی نظیر پارک و موزه، پرسش و پاسخ‌های علمی

خوشه۲ : جستجو در خصوص فیلم و موسیقی و هنرپیشه‌ها و هنرمندان

خوشه ۳ : اخبار

خوشه ۴ : یاهو(فیلم، آهنگ، جستجو)

خوشه ۵ : مطالعه نظرات سایرین در خصوص اتومبیل، کتاب، هتل و ...

خوشه ۶ : مشاهده قیمت اتومبیل، خانه و همچنین مقایسه قیمت اجناس

خوشه ۷ : فروشگاه‌های آنلاین و ارسال ایمیل

خوشه ۸ : فیلم و عکس و کارت تبریک‌های الکترونیکی و بازی‌های آنلاین

خوشه ۹ : کتاب و مجلات آنلاین

خوشه ۱۰ : سرگرمی‌های اجتماعی

خوشه ۱۱ : دایره المعارف و اطلس‌ها

خوشه ۱۲ : خرید و فروش آنلاین

پس از تحلیل خوشه‌ها و استخراج برچسب هر خوشه، خروجی تولید شده در مرحله خوشه‌بندی به نرم‌افزار SQL

Server داده شـده و تعداد تکرار هر سـایت در هر خوشـه مورد محاسبه قرار گرفته است. سپس بر اساس این تعداد، یک رتبه به هر سـایت تخصـیص داده شـده است. این رتبه جهت پیشنهاد به کاربر و تسـهیل در جسـتجوی او به کار برده خواهد شد. شکل ۳ نحوه محاسـبه سـایت‌های با بهترین رتبه برای هر خوشـه در نرم‌افزار SQL Server را نشان می‌دهد.

```
select distinct  Clickurl,
ClusterRank,
UrlCount,
ISNULL(Title,'') AS Title,
ISNULL(KeyWords,'') AS Keywords,
ISNULL(Description,'') AS Description
FROM tblFinal
WHERE Cluster=N'Cluster_5'
ORDER BY ClusterRank DESC
```

0 %

Results | Messages

	Clickurl	ClusterRank	UrlCount	Title
	http://www.nextag.com	13	2094	Nextag | Compare Prices Before you
2	http://www.city-data.com	12	1268	City-Data.com - Stats about all US ci
3	http://www.kbb.com	11	818	Kelley Blue Book - New & Used Car
4	http://www.homedepot.com	10	640	Home Improvement Made Easy with

شکل ۳ : نحوه محاسبه سایت‌های با بهترین رتبه برای هر خوشه

۵- نتیجه

در مـراحـل اولیـه اجرای پژوهش، منـابع داده‌ای تحقیق جمع‌آوری شـده و با اسـتفاده از ابزارهای سـئو بر اسـاس سـتون آدرس سایت موجود در داده، سه ستون "عنوان"، "توضیحات" و "کلمات کلیدی" به داده مورد نظر اضـافه شـده است و روی این داده از طریق نرم‌افزار رپیدماینر عملیات متن‌کاوی صـورت گرفته اسـت. سپس داده‌های حاصل از متن‌کاوی وارد مدل‌سازی شده و روی آن‌ها عملیات خوشـه‌بندی انجام شـده اسـت. در مدل‌سازی این پژوهش روش K-Means با تعداد خوشـه ۱۲ بیشترین دقت را کسـب نموده اسـت. لذا تحلیل‌های نهایی روی خوشـه‌بندی با تعداد خوشـه ۱۲ صـورت گرفته است. لازم به ذکر است که برای سـنجش دقت خوشـه‌ها از شـاخص دیویس بولدین استفاده شده اسـت. سـپس با نرم‌افزار SQL Server خوشـه‌ها تحلیل شـده و برای هر آدرس سـایت در هر خوشه تعداد تکرار آن سایت در آن خوشـه (تعـداد کلیک‌های صـورت گرفته برای آن سـایت) و همچنین رتبه سـایت در آن خوشـه به ازای این تعداد کلیک، به دست آمده است تا با استفاده از روش‌های پیش‌بینی کننده بتوان برای پرس‌وجوهـای آتی کـه وارد موتور جسـتجو می‌شـونـد سایت‌های با رتبه بالاتر را به کاربر پیشنهاد نمود تا به هدف اصلی این پژوهش که اسـتفاده از روش‌های داده‌کاوی برای شـناسـایی الگوهـای رفتاری کاربران و کمک به آن‌ها در جسـتجوهای آتی است، دست یافت.

مراجع

[1] R. R. Yager, A. Rybalov, On the Fusion of Documents from Multiple Collection Information Retrieval Systems, Journal of the American Society for Information Science, 1997.

[2] R.R.Yager, V. Kreinovich, On How to Merge Sorted Lists Coming from Different Web Search Tools, AmericanAssociation for Artificial Intelligence (AAAI) Symposium on Frontiers in Soft Computing and Decision Systems, MIT, Boston, MA, 1997.

[3] http://www.civilica.com/Paper-ITPF02-ITPF02_063.html

[4] Yan Li, Xin-Zhong Chen and Bing-Ru Yang.Research on Web mining-based intelligent search engine. International onference on Machine Learning and Cybernetics, 2002

[5] Bruno M. Fonseca, Paulo Golgher, Bruno Possas, Berthier Ribeiro-Neto, and Nivio Ziviani. Concept-based interactive query expansion. In Proc. CIKM'05. ACM, 2005.

[6] Paul Alexandru Chirita, Claudiu S. Firan, and Wolfgang Nejdl. Personalized query expansion for the web. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '07, pages 7-14, New York, NY, USA, 2007. ACM.

[7] H. Cui, J.R. Wen, J.Y. Nie, and W.Y. Ma. Probabilistic query expansion using query logs. In Proceedings of the 11th international conference on World Wide Web, pages 325-332. ACM, 2002.

[8] Ranieri Baraglia, Fidel Cacheda, Victor Carneiro, Diego Fernandez, Vreixo For-moso, Raffaele Perego, and Fabrizio Silvestri. Search shortcuts: a new approach to the recommendation of queries. In Proceedings of the third ACM conference on Recommender systems, RecSys '09, pages 77-84, New York, NY, USA, 2009. ACM.

[9] Sakthi Priya T, Revathy P, Pradeesh T, Rene Robin C.R., Design and development of an ontology based personal web search engine, 2nd International Conference on Communication, Computing & Security, Elsevier,2012.

[10] http://www.civilica.com/Paper-ICIKT01-ICIKT01_068.html

[11] http://www.civilica.com/Paper-IDMC04-IDMC04_039.html

[12] http://www.civilica.com/Paper-ITPF02-ITPF02_063.html

[13] F.Kovacs,C.Legany,A.Babos,*"ClusterValiditi Measurement Techniques"*, Department Of Automation And Applied Informatics, Budapest University of Technology and Economics, 2003.

[14] D.L. Davies and D.W. Bouldin. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, (2):224{227, 1979.