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

سولماز لطفی آذری داریان 1، رضا جاویدان 2

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

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

**چکيده - 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

# 1- 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]1و2[.

تکنيک­هاي کاوش در وب، مي­توانند براي شخصي­سازي استفاده کاربران از وب به کار روند. براي مثال مي­توان رفتار کاربر را از طريق مقايسه الگوي پيمايش فعلي وي با الگوهاي پيمايش استخراج شده از فايل­هاي ثبت­شده، به صورت بلادرنگ پيش­بيني 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]3[.

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 2 پژوهش­های مرتبط در این زمینه بررسی شده است. در بخش 3 روش پیشنهادی توضیح داده شده است. در بخش 4 نتایج حاصل از روش بیان شده و نهایتا در بخش 5 نتیجه­گیری ارائه شده است.

# 2- مرور کارهای مرتبط

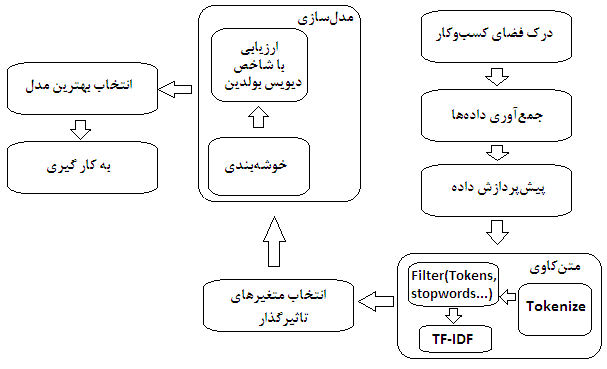
داده­کاوی به فرآیند تجزیه و تحلیل پایگاه­ داده­های بزرگ به منظور یافتن الگوهایی مفید، گفته می­شود. روش­های داده­کاوی شامل دو دسته عمده روش­های پیش­بینی کننده و روش­های تشریحی می­باشند. هدف این روش­ها جداسازی و قرار دادن اشیا در تعدادی از کلاس­هاست که در روش­های پیش­بینی کننده مانند روش رده­بندی، این 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 2002، 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]4[. Foncesa و همکارانش در سال 2007 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.]5[. Chirita و همکارانش در سال 2007 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]6[. Cui و همکارانش در سال 2009 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 30 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]7[. Braglia و همکارانش در سال 2009 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]8[. پیریا و پویتی در مقاله­ای در سال 2012 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]9[. آقایان کاووسی و مشیری در اولين كنفرانس بين المللي فناوري اطلاعات و دانش در سال 1382، مقاله‌ای با عنوان استخراج مدل رفتاري موتورهاي جستجوي مورد استفاده در يك فراجويشگر هوشمند ارائه دادند که درآن مقاله یك فراجويشگر ارائه شده است، که علایق كاربر را دريافت مي‌کند و با توجه به آن پرس­وجوهاي متناسب با چند موتور جستجو تنظيم مي­گردد و به موتورهاي جستجو ارسال مي­شود. سپس نتايج بازگشتي از موتورهاي جستجو پالايش مي­شوند و پس از تعيين اولويت در اختيار كاربر قرار مي­گيرند. در اين مقاله تلاش شده است تا بخشي از معماري يك عامل هوشمند سفارشي مورد بررسي قرار گيرد كه قادر است به مرور زمان و با توجه به بازخوردي كه از كاربر دريافت مي­كند مدل رفتاري موتورهاي جستجو را در خوشه­هاي موضوعي مختلف استخراج نمايد.]10[. آقایان حسنپور و دهقانی و کنشلو در چهارمين كنفرانس داده­كاوي در سال 1389، مقاله‌ای تحت عنوان مرتب­سازي نتايج موتورهاي جستجو بر اساس تاريخچه رفتاري كاربر ارائه دادند که در آن راهكار جديدي برای موتورهای جستجو ارائه شده است. دراين راهكار به منظور شناسايي كاربران صفحاتي كه پس از جستجو مورد توجه كاربر قرار مي­گيرند پردازش معنايي شده و به عنوان تاريخچه ملاقات­هاي قبلي كاربر مورد توجه قرار مي­گيرد، كه در شكل­گيري الگوي رفتاري وي موثر مي­باشد]11[. آقایان امیر پناه، امید پناه و امین پناه در دومين همايش فناوري اطلاعات، حال، آينده در سال 1390، مقاله­ای با عنوان بررسي وب­كاوي و پيشنهاد روشي جديد براي افزايش سرعت جستجو در وب ارائه دادند. در اين مقاله، روش جديدي براي جستجوی اطلاعات موجود در وب به وسيله موتورهاي جستجو پيشنهاد شده است. اين روش با تركيب دو روش موجود و به كاربرده شده در دو موتور جستجوي مشهور سبب كاهش زمان جستجو و بالا بردن كارائي جستجو مي­شود]12[.

# 3- روش پیشنهادی

معماری روش پیشنهادی و روند ارائه روش در شکل 1 نشان داده شده است و قسمت­های بعد بر اساس معماری روش توضیح داده شده است.



شکل 1 : 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.

## 3-1- جمع­آوری و پردازش داده

در این پژوهش داده­ مورد استفاده، فایل رویدادنگاری پرس­وجوهای ثبت شده توسط 650000 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.

## 3-2- 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 200 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.

## 3-3- 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:

1-تعیین K یا تعداد خوشه­ها

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

3-محاسبه فاصله تمامی نمونه­ها با مراکز تعیین شده اولیه

4-تخصیص نمونه­ها به مراکز نزدیک­تر

5-محاسبه مرکز هندسی خوشه­های تشکیل­شده

6-تکرار مراحل 3 تا 5 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.

## 3-4- 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]13[ :

1- 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.

2- 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 :

1-معیارهای خروجی (External Criteria)

2-معیارهای درونی (Internal Criteria)

3-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]14[ از معیار شباهت بین دو خوشه(Rij) استفاده می­کند که بر اساس پراکندگی یک خوشه(si) و عدم شباهت بین دو خوشه(dij) تعریف می­شود. شباهت بین دو خوشه را می­توان به صورت­های مختلفی تعریف کرد ولی بایستی شرایط زیر را دارا باشد:

* Rij ≥ 0
* Rij = Rji
* اگر si و sj هر دو برابر صفر باشند آنگاه Rij نیز برابر صفر باشد.
* اگر sj > sk و dij = dik آنگاه Rij > Rik
* اگر sj = sk و dij < dik آنگاه Rij > Rik

معمولا شباهت بین دو خوشه به صورت زیر تعریف می­شود :

Rij= (1)

که در آن dij و si با روابط زیر محاسبه می­شوند :

dij=d(νi, νj) (2)

si= (3)

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

DB=(4)

که Ri در آن به صورت زیر محاسبه می­شود :

Ri= max(Rij),i=1…nc (5)

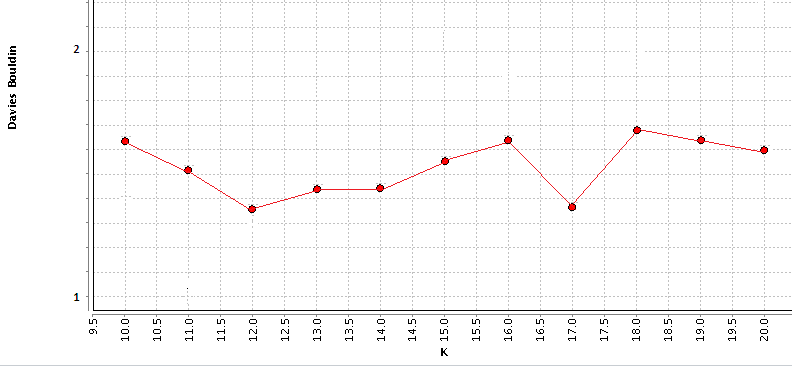
این شاخص در واقع میانگین شباهت بین هر خوشه با شبیه­ترین خوشه به آن را محاسبه می­کند.

# 4-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 12 و نوع مقیاس "واگرایی برگمن" و نوع واگرایی " فاصله اقلیدسی" بهینه انتخاب شدند و روی مدل­سازی نهایی لحاظ شدند.

شکل 2 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 2 : 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 12 و 17 که شاخص دیویس بولدین آن­ها از همه کمتر شدند از سایر تعداد خوشه­ها نتیجه بهتری داشتند. در این پژوهش روی مدل­سازی نهایی تعداد خوشه 12 لحاظ شده است.

مدل­سازی نهایی با روش K-Means و با در نظر گرفتن تعداد خوشه 12 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 1 : خدمات عمومی آنلاین نظیر مشاهده آب­وهوا، تقویم، مشاهده اماکن تفریحی نظیر پارک و موزه، پرسش و پاسخ­های علمی

خوشه2 : جستجو در خصوص فیلم و موسیقی و هنرپیشه­ها و هنرمندان

خوشه 3 : اخبار

خوشه 4 : یاهو(فیلم، آهنگ، جستجو)

خوشه 5 : مطالعه نظرات سایرین در خصوص اتومبیل، کتاب، هتل و ...

خوشه 6 : مشاهده قیمت اتومبیل، خانه و همچنین مقایسه قیمت اجناس

خوشه 7 : فروشگاه­های آنلاین و ارسال ایمیل

خوشه 8 : فیلم و عکس و کارت تبریک­های الکترونیکی و بازی­های آنلاین

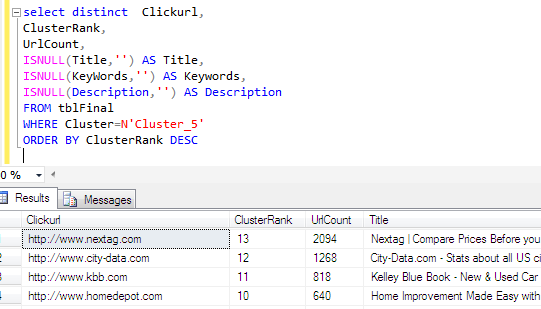
خوشه 9 : کتاب و مجلات آنلاین

خوشه 10 : سرگرمی­های اجتماعی

خوشه 11 : دایره المعارف و اطلس­ها

خوشه 12 : خرید و فروش آنلاین

پس از تحلیل خوشه­ها و استخراج برچسب هر خوشه، خروجی تولید شده در مرحله خوشه­بندی به نرم­افزار SQL Server داده شده و تعداد تکرار هر سایت­ در هر خوشه مورد محاسبه قرار گرفته است. سپس بر اساس این تعداد، یک رتبه به هر سایت تخصیص داده شده است. این رتبه جهت پیشنهاد به کاربر و تسهیل در جستجوی او به کار برده خواهد شد. شکل 3 نحوه محاسبه سایت­های با بهترین رتبه برای هر خوشه در نرم­افزار SQL Server را نشان می­دهد.



**شکل 3 : نحوه محاسبه سایت­های با بهترین رتبه برای هر خوشه**

# 5- نتيجه

در مراحل اولیه اجرای پژوهش، منابع داده­ای تحقیق جمع­آوری شده و با استفاده از ابزارهای سئو بر اساس ستون آدرس سایت موجود در داده، سه ستون "عنوان"، "توضیحات" و "کلمات کلیدی" به داده مورد نظر اضافه شده است و روی این داده از طریق نرم­افزار رپیدماینر عملیات متن­کاوی صورت گرفته است. سپس داده­های حاصل از متن­کاوی وارد مدل­سازی شده و روی آن­ها عملیات خوشه­بندی انجام شده است. در مدل­سازی این پژوهش روش K-Means با تعداد خوشه 12 بیشترین دقت را کسب نموده است. لذا تحلیل­های نهایی روی خوشه­بندی با تعداد خوشه 12 صورت گرفته است. لازم به ذکر است که برای سنجش دقت خوشه­ها از شاخص دیویس بولدین استفاده شده است. سپس با نرم­افزار SQL Server خوشه­ها تحلیل شده و برای هر آدرس سایت در هر خوشه تعداد تکرار آن سایت در آن خوشه (تعداد کلیک­های صورت گرفته برای آن سایت) و همچنین رتبه سایت در آن خوشه به ازای این تعداد کلیک، به دست آمده است تا با استفاده از روش­های پیش­بینی کننده بتوان برای پرس­وجوهای آتی که وارد موتور جستجو می­شوند سایت­های با رتبه بالاتر را به کاربر پیشنهاد نمود تا به هدف اصلی این پژوهش که استفاده از روش­های داده­کاوی برای شناسایی الگوهای رفتاری کاربران و کمک به آن­ها در جستجوهای 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](http://www.civilica.com/Paper-ITPF02-ITPF02_063=%D8%A8%D8%B1%D8%B1%D8%B3%DB%8C-%D9%88%D8%A8-%DA%A9%D8%A7%D9%88%DB%8C-%D9%88-%D9%BE%DB%8C%D8%B4%D9%86%D9%87%D8%A7%D8%AF-%D8%B1%D9%88%D8%B4%DB%8C-%D8%AC%D8%AF%DB%8C%D8%AF-%D8%A8%D8%B1%D8%A7%DB%8C-%D8%A7%D9%81%D8%B2%D8%A7%DB%8C%D8%B4-%D8%B3%D8%B1%D8%B9%D8%AA-%D8%AC%D8%B3%D8%AA%D8%AC%D9%88-%D8%AF%D8%B1-%D9%88%D8%A8.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](%20http://www.civilica.com/Paper-IDMC04-IDMC04_039.html)

[12] [http://www.civilica.com/Paper-ITPF02-ITPF02\_063.html](http://www.civilica.com/Paper-ITPF02-ITPF02_063=%D8%A8%D8%B1%D8%B1%D8%B3%DB%8C-%D9%88%D8%A8-%DA%A9%D8%A7%D9%88%DB%8C-%D9%88-%D9%BE%DB%8C%D8%B4%D9%86%D9%87%D8%A7%D8%AF-%D8%B1%D9%88%D8%B4%DB%8C-%D8%AC%D8%AF%DB%8C%D8%AF-%D8%A8%D8%B1%D8%A7%DB%8C-%D8%A7%D9%81%D8%B2%D8%A7%DB%8C%D8%B4-%D8%B3%D8%B1%D8%B9%D8%AA-%D8%AC%D8%B3%D8%AA%D8%AC%D9%88-%D8%AF%D8%B1-%D9%88%D8%A8.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.