مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره چهارم شماره سوم. ص ۲۱۵-۲۰۱، ۱۳۹۶

روشی ترکیبی به‌منظور توصیه پرس‌وجوهای پزشکی در سیستم‌های توصیه‌گر

الهام اسماعیلی گوهری[۱]، سجاد ظریف‌زاده[۲]*، سعید حسنوند[۳]

• دریافت مقاله: ۹۶/۷/۶ • پذیرش مقاله: ۹۶/۹/۶

مقدمه: رشد روزافزون اطلاعات موجود در اینترنت و سربار زیاد اطلاعاتی، چالش مهمی برای کاربران در جهت دسترسی به اطلاعات موردنیازشان ایجاد کرده است. امروزه توصیه‌گرهای پرس‌وجو به یک جزء جدایی‌ناپذیر سیستم‌های بازیابی اطلاعات تبدیل شده‌اند. یکی از کاربردهای این توصیه‌گرها در زمینه علوم پزشکی است. این سیستم‌ها با به کارگیری فرایندهای شخصی‌سازی سعی در تسکین مشکل سرریز اطلاعات در وب و سرعت بخشیدن به جستجوی اطلاعات پزشکی کاربران دارند.

روش: این پژوهش از نوع کاربردی و توصیفی است. در این پژوهش سعی شد با استفاده از ویژگی‌های محتوایی پرس‌وجوها و نتایج جستجو روشی ارائه شود که ضمن حفظ ارتباط معنایی با پرس‌وجوی اصلی، کاربران را سریع‌تر به نیازهای اطلاعاتی‌شان برساند. به منظور خوشه‌بندی پرس‌وجوها از الگوریتم K-means استفاده شد. پیاده‌سازی روش پیشنهادی با استفاده از زبان برنامه‌نویسی جاوا و نرم‌افزار NetBeans IDE صورت گرفت.

نتایج: با توجه به سیستم پیشنهادی، استفاده توامان از ویژگی‌های ساختاری پرس‌وجوها و نتایج جستجو حاوی اطلاعاتی مفیدی برای تشخیص پرس‌وجوهای مشابه است. از آن‌جا که امکان وجود کلمات چندمعنا در پرس‌وجوی کاربران وجود دارد، استفاده از نتایج جستجو می‌تواند در امر تشخیص هدف کاربر از پرس‌وجو مفید باشد.

نتیجه‌گیری: نتایج حاصل از ارزیابی روش پیشنهادی با دادگان واقعی مربوط به موتور جستجوی بومی پارسی‌جو، بیانگر مؤثر بودن این روش در بهبود دقت توصیه نسبت به سایر روش‌ها است. طبق ارزیابی‌های انجام‌شده، دقت سیستم پیشنهادی برابر با ۷۷/۲۴٪ است که در مقایسه با مطالعات مطرح در این زمینه، ۱۰٪ بهبود داشته است.

کلید واژه‌ها: سیستم توصیه‌گر پزشکی، توصیه پرس‌وجو، موتور جستجو، بازیابی اطلاعات

•ارجاع: اسماعیلی گوهری الهام، ظریف‌زاده سجاد، حسنوند سعید. روشی ترکیبی به‌منظور توصیه پرس‌وجوهای پزشکی در سیستم‌های توصیه‌گر. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۶؛ ۴(۲): ۲۱۵-۲۰۱.

۱. کارشناس ارشد مهندسی کامپیوتر، گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه یزد، یزد، ایران
۲. دکترای مهندسی کامپیوتر، استادیار، گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه یزد، یزد، ایران
۳. کارشناس ارشد مهندسی کامپیوتر، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران
*** نویسنده مسئول:** یزد، صفائیه، دانشگاه یزد، دانشکده فنی و مهندسی، گروه مهندسی کامپیوتر
•شماره تماس: ۰۳۵۳۸۲۰۰۱۴۴
• **Email:** szarifzadeh@yazd.ac.ir

مقدمه

فرآیند تصمیم‌گیری بخشی از زندگی روزمره افراد را تشکیل می‌دهد؛ اما گاهی اوقات به دلیل عدم وجود تجربه و یا وجود انتخاب‌های بسیار، این فرآیند دشوار می‌شود. امروزه با توجه به وجود حجم وسیع انتخاب‌ها و تصمیم‌گیری‌هایی که در پی گسترش روزافزون اینترنت به وجود آمده است، استفاده از روش ارتباطات شخصی به‌منظور دریافت توصیه و پیشنهاد، غیرعملی است. به همین دلیل به سیستمی نیاز است که بتواند با استفاده از سابقه و علائق کاربر توصیه‌های مناسبی را به وی ارائه دهد. چنین سیستمی از صرف وقت و هزینه کاربر برای بررسی تمام اطلاعات موجود جلوگیری می‌کند. این سیستم‌ها مؤظف هستند که برای پرس‌وجوهای کاربران که به زبان طبیعی ارسال شده‌اند، پرس‌وجوهای مرتبطی را توصیه کنند که کاربران را سریع‌تر به اطلاعات مورد نیازشان برسانند [۲، ۱]. این حجم زیاد اطلاعات، کاربران را در پیدا کردن اطلاعات پزشکی موردنیاز خود نیز دچار چالش می‌کند.

موتورهای جستجو به عنوان مهم‌ترین ابزار بازیابی اطلاعات، توجه زیادی را در صنعت و دانشگاه به خود جلب کرده‌اند. بررسی گزارش‌های موتورهای جستجو می‌تواند به منظور فهم رفتار جستجوی کاربران، دسته‌بندی نیازهای اطلاعاتی آن‌ها، بهبود رتبه‌بندی اسناد و توصیه پرس‌وجوهای مفید، به کار برده شود [۳]. بازیابی اطلاعات سلامت (Helath Information HIR(Retrieval از فضای وب، یک عمل رایج و مهم از سوی کاربران است. کاربران مختلف ممکن است جستجوهایی در حیطه پزشکی برای خود، خانواده و یا دوستان خود انجام دهند تا بتوانند برای پرسش‌های خود پاسخی پیدا کنند. به دلیل حجم زیاد اطلاعات موجود در زمینه پزشکی و تخصصی بودن اطلاعات مربوط به این حوزه که توسط کاربران جستجو می‌شوند، نیاز به سیستمی وجود دارد که بتواند به صورت کارا و مؤثر به کاربران در جهت جستجوی اطلاعات مورد نیازشان کمک کند. این سیستم‌ها نوعی سیستم پالایش اطلاعات هستند و اطلاعات اضافی را قبل از ارائه به کاربران حذف می‌کنند. به این طریق سربار اطلاعات کم می‌شود و اطلاعات شخصی‌سازی‌شده به کاربر ارائه می‌شود [۴].

به طور کلی، توصیه پرس‌وجو را می‌توان در قالب‌های مختلفی از جمله تکمیل خودکار پرس‌وجو [۵]، تصحیح املایی پرس‌وجو [۶] و توسعه و بسط پرس‌وجو [۷] نیز به کار برد. اولین تلاش‌ها برای ساخت سیستم‌های توصیه‌گر پرس‌وجو از اوایل دهه ۲۰۰۰ میلادی آغاز شد. بررسی رفتارهای جستجوی کاربران از طریق گزارش موتورهای جستجو به منظور توصیه پرس‌وجو تاکنون در پژوهش‌های زیادی مورد مطالعه قرار گرفته است، اما تلاش کمی در زمینه فرموله کردن رفتارهای جستجوی کاربران در حیطه پزشکی و بهداشتی انجام شده است. اگرچه تاکنون موتورهای جستجو در بازیابی اطلاعات خوب عمل کرده‌اند، اما هنوز هم برای یافتن اطلاعات موردنیاز کاربران، به کلمات موجود در پرس‌وجوها وابسته هستند. به عبارتی این موتورهای جستجو بر اساس تطابق عبارات موجود در پرس‌وجوها و اسناد (و نه بر اساس معنا)، به تشخیص شباهت بین آن‌ها می‌پردازند [۸]؛ بنابراین اگر پرس‌وجویی نیاز اطلاعاتی کاربران را به درستی بیان نکند، نتایج مناسبی نیز نمایش داده نمی‌شود. به عنوان مثال، فرض کنید کاربری به جای وارد کردن پرس‌وجو "آیا هیچ درمان طبیعی برای بیماری سندرم کارتاژنز وجود دارد؟"، پرس‌وجوی "بیماری PCD" را وارد کند، در این صورت موتور جستجویی که مبتنی بر کلمات موجود در پرس‌وجوها عمل کند، نمی‌تواند شباهت بین این دو پرس‌وجوها را به درستی تشخیص دهد. از سویی دیگر هنگامی که کاربران می‌خواهند پرس‌وجوهایی در حیطه پزشکی وارد کنند، به دلیل دانش محدودی که اغلب افراد در رابطه با واژگان پزشکی دارند، پرس‌وجوهای ساده‌ای وارد می‌کنند؛ به‌عبارت‌دیگر در بعضی مواقع برای کاربران دشوار است تا پرس‌وجوی مناسبی را برای بیان نیاز اطلاعاتی خود پیدا کنند و اغلب با چندین بار اصلاح پرس‌وجو و امتحان کردن پرس‌وجوهای مختلف به اطلاعات موردنیاز خود دسترسی پیدا می‌کنند [۱۰، ۹]. این موضوع بیشتر در مواردی اتفاق می‌افتد که کاربر قصد جستجوی مطلبی را دارد که نام دقیق آن را نمی‌داند و اقدام به توصیف خواسته خود می‌کند. برای مثال فرض کنید که کاربری اسم دقیق دارویی را که می‌خواهد جستجو کند، نمی‌داند و به جای آن از عبارت "داروی ضدافسردگی" استفاده می‌کند. همچنین، این روش‌ها کلمات چندمعنا و کلمات مخفف را نیز در نظر نمی‌گیرند. این مسئله به خصوص در زمانی که پرس‌وجوها حاوی تعداد کلمات کمی باشند، بیشتر باعث ایجاد خطا می‌شود. یک مطالعه نشان می‌دهد که کاربران HIR معمولاً پرس‌وجوهای عمومی و با طول خیلی کم وارد می‌کنند [۱۱]. به عبارتی کاربران پرس‌وجوهایی با طول میانگین بیش از ۲ یا ۳ وارد نمی‌کنند [۱۲]. از طرف دیگر نتایج جستجو اغلب حاوی اطلاعات مفیدی برای پیدا کردن پرس‌وجوهای مشابه هستند.

روش‌های متعددی تاکنون برای توصیه پرس‌وجو مطرح شده

که هر کدام از آن‌ها با رویکردی متفاوت به این مبحث پرداخته است. Boldi و همکاران برای اولین بار مفهومی به نام گراف جریان پرس‌وجو را معرفی کردند. در این گراف، گره‌ها، پرس‌وجوهای واردشده توسط کاربران هستند و هر یال از پرس‌وجوی q_a به پرس‌وجوی q_b نشان‌دهنده آن است که حداقل در یک نشست، پرس‌وجوی q_b بلافاصله بعد از پرس‌وجوی q_a آمده است. وزن هر یال نیز برابر با تعداد دفعاتی است که این زوج در جستجوی کاربران تکرار شده است [۱۳]. سپس از این گراف به منظور پیدا کردن پرس‌وجوهای مشابه استفاده می‌شود. عیب اصلی این روش این است که در این گراف بیش از نیمی از زوج‌های پرس‌وجو فقط یک بار در جستجوی کاربران آمده است. از طرف دیگر این گراف به شدت نامتقارن است، زیرا ۹۳٪ از یال‌های آن، یال معکوس ندارند. همچنین این روش تنها قادر به ارائه توصیه برای پرس‌وجوهایی است که قبلاً حداقل یک بار در گزارش‌های موتور جستجو آمده باشند.

در روش دیگری که Mei و همکاران انجام دادند، از گراف دوبخشی "پرس‌وجو-آدرس اینترنتی" برای توصیه پرس‌وجو استفاده شده است. در این روش به ازای هر پرس‌وجوی موجود در دادگان که منجر به کلیک بر روی یکی از نتایج جستجو شده است، یک زوج (پرس‌وجو، آدرس کلیک‌شده) استخراج می‌گردد. سپس با استفاده از این زوج‌ها، می‌توان یک گراف دوبخشی ساخت که گره‌های آن شامل پرس‌وجوها و آدرس‌های اینترنتی است و هر یال بین پرس‌وجوی i و آدرس اینترنتی j نشان‌دهنده این است که کاربر بعد از وارد کردن پرس‌وجوی i، بر روی آدرس اینترنتی j کلیک کرده است. سپس شباهت بین پرس‌وجوها با استفاده از زمان اصابت (Hitting time) الگوریتم گام تصادفی به دست می‌آید [۱۴].

Zhang و همکاران [۱۵] از توزیع اسناد مرتبط با کاربر برای توصیه پرس‌وجو استفاده کردند. ایده اصلی این مقاله این است که زمانی که اسناد مشاهده‌شده توسط کاربر را به ترتیب زمان کلیک آن‌ها مرتب کنیم، خواهیم دید که توزیع اسنادی که مرتبط با پرس‌وجوی کاربر هستند، متمرکز است، ولی توزیع اسناد غیرمرتبط نامتمرکز است. علت آن هم، یادگیری کاربر در طول زمان است. به همین خاطر در این مقاله، از وزن عبارات موجود در اسناد کلیک‌شده توسط کاربر برای توصیه پرس‌وجو استفاده شده است. وزن این عبارات به کمک تراکم اسناد کلیک‌شده‌ای که شامل این عبارات هستند، به دست می‌آید. بدین‌ترتیب تأثیر محتوای صفحاتی که ارتباط کمتری با پرس‌وجوی کاربر داشته باشند، کمتر می‌شود؛ بنابراین برای هر پرس‌وجو، برداری از کلمات مرتبط ساخته شده و شباهت بین پرس‌وجوها با استفاده از بردارهای آن‌ها محاسبه می‌شود.

Mitsui و همکاران [۱۶] سعی کرده‌اند تا به توصیه پرس‌وجوهایی که در حیطه موضوع پرس‌وجوی موردنظر هستند، بپردازند. ایده آن‌ها از این جهت مورد استفاده قرار گرفته است که مشکلی که در اکثر سیستم‌های توصیه وجود دارد این است که در اغلب مواقع پرس‌وجوهایی با کلمات مشابه با پرس‌وجوی کاربر توصیه می‌شوند. به همین دلیل در این مقاله، ابتدا پرس‌وجوهای کاربر و اسنادی که توسط وی مرور شده‌اند، بررسی می‌شوند. سپس بر اساس اسنادی که اخیراً مورد توجه کاربر بوده‌اند، موضوع اسناد استخراج می‌شود و پرس‌وجویی در آن موضوع تولید و به کاربر توصیه می‌شود. عیب اصلی این روش این است که پرس‌وجوی ارائه‌شده ممکن است کاربران را از هدف واقعی خود دور کند. برای مثال فرض کنید کاربری به دنبال روش درمان سرطان پوست باشند و سیستم پرس‌وجوهایی در رابطه با آمار مرگ‌ومیر سرطان پوست به وی توصیه کند.

در تحقیق دیگری که توسط Zeng و همکاران انجام شد، از منطق فازی برای به دست آوردن فاصله معنایی پرس‌وجوها استفاده شد. آن‌ها برای به دست آوردن فاصله معنایی پرس‌وجوها از سلسله مراتب پزشکی و ارتباط معنایی در واژگان پزشکی استفاده می‌کنند. کلیک‌های کاربران را نیز به منظور اصلاح پرس‌وجوها به کار می‌برند. عیب اصلی روش‌های مرتبط با سلسله مراتب مفهومی این است که ممکن است پرس‌وجوهایی به کاربران توصیه شود که کلمات موجود در آن‌ها به دلیل دانش کم کاربران در حیطه پزشکی، برای آن‌ها نامأنوس باشد [۱۱].

همان‌طور که گفته شد، هر کدام از این روش‌ها در کنار مزایای خود، معایبی نیز دارند. با این وجود، نیاز به ارائه روشی وجود دارد که ضمن ارائه توصیه‌های معتبر، پیچیدگی‌های زبان طبیعی از جمله زبان فارسی را نیز در نظر بگیرد. کارهای اشاره‌شده همگی در زبان انگلیسی انجام شده‌اند. طبق بررسی‌های انجام‌شده تاکنون توصیه پرس‌وجوهای فارسی در حیطه پزشکی مورد تحقیق قرار نگرفته است؛ بنابراین علاوه‌بر چالش‌های مطرح‌شده در این حوزه، چالش‌های دیگری از جمله نبود ابزار پیش‌پردازش دقیق و دادگان مناسب برای زبان فارسی نیز وجود دارد.

نظر به اهمیت بالای توصیه پرس‌وجوهای پزشکی و کمک به کاربران در جهت یافتن اطلاعات موردنیاز خود در سریع‌ترین زمان ممکن، در این پژوهش سعی شده است با ترکیب ویژگی‌های محتوایی پرس‌وجوها و نتایج جستجو، روشی ترکیبی با هدف افزایش دقت در توصیه پرس‌وجوهای پزشکی ارائه شود. در این زمینه شناسایی پرس‌وجوهای مرتبط به خوبی و با دقت قابل قبولی نسبت به روش‌های موجود صورت گرفته است. همچنین توصیه پرس‌وجوهای پزشکی برای پرس‌وجوهای زبان فارسی که تاکنون مورد بررسی قرار نگرفته بود، امکان‌پذیر شده است. نتایج ارزیابی روش پیشنهادی نشان‌دهنده دقت قابل قبول روش پیشنهادی در مقایسه با روش‌های موجود است.

روش

این مطالعه از نوع کاربردی-توصیفی است. در این مقاله ترکیب ویژگی‌های ساختاری و نتایج حاصل از جستجوی پرس‌وجوها به منظور بهبود دقت توصیه پرس‌وجو مورد استفاده قرار گرفته است. ساختار روش پیشنهادی از سه گام اصلی (۱) پیش‌پردازش دادگان، (۲) خوشه‌بندی پرس‌وجوها و (۳) توصیه پرس‌وجو تشکیل شده است. بر این اساس ابتدا پس از گام پیش‌پردازش دادگان که شامل اصلاح غلط‌های املایی موجود در مجموعه دادگان، برچسب‌گذاری کلمات مترادف، ریشه‌یابی، یکسان‌سازی نوشتاری و حذف کلمات توقفی است، به خوشه‌بندی پرس‌وجوها پرداخته می‌شود. برای خوشه‌بندی پرس‌وجوها از دو ویژگی "N-گرام بر روی پرس‌وجوها" و "مجموعه آدرس‌های اینترنتی نمایش داده شده در صفحه نتایج پرس‌وجو و رتبه آن‌ها" استفاده می‌شود. در گام سوم، بعد از به دست آمدن شباهت بین پرس‌وجوها، با اجرای الگوریتم K-means به خوشه‌بندی پرس‌وجوها می‌پردازیم. در نهایت، با ورود هر پرس‌وجو، سیستم قادر است، اقدام به توصیه پرس‌وجوهای معتبر نماید.

الف) پیش‌پردازش دادگان

در این مرحله کلیه پرس‌وجوهای موجود در پایگاه داده پیش‌پردازش شده و ساختار نوشتاری تمامی آن‌ها تا حد امکان به فرم مشابهی تبدیل می‌شود.

- در فاز اول غلط‌های املایی موجود در پرس‌وجوهای مربوط به دادگان تصحیح می‌شوند. طبق بررسی‌های انجام‌گرفته در موتور جستجوی بومی پارسی‌جو نرخ خطاهای املایی کاربران فارسی‌زبان در پرس‌وجوها برابر با ۲۰ درصد است. غلط‌های املایی موجود در دادگان مربوط به این پژوهش به صورت دستی از پرس‌وجوهای کاربران حذف شده است. برای مثال اگر کاربری پرس‌وجوی "درمان واریث" را وارد کرده باشد، پس از این فاز، این پرس‌وجو به "درمان واریس" تبدیل می‌شود.
- از آن‌جا که برخی از کلمات می‌توانند در شکل‌های نوشتاری مختلف ظاهر شوند، لذا یکسان‌سازی نوشتاری این کلمات در این فاز صورت می‌گیرد. علت این کار وجود برخی از حروف زبان عربی در مجموعه حروف زبان فارسی است. برای مثال حرف "ی" و "ئ" که موجب نگارش‌های متفاوت یک کلمه واحد می‌شوند، در حالی‌که این دو حرف دارای یونیکدهای متفاوتی هستند؛ بنابراین در این فاز لازم است این حروف نرمال شده و تنها با یک یونیکد نمایش داده شوند. برای مثال، پرس‌وجوی "آمنئوتیک" به "آمنیوتیک" تبدیل می‌شود. از طرف دیگر، برخی حروف مانند حرکت‌ها، همزه و علامت‌های نشانه‌گذاری مانند "! : . ؟" نگارش‌های یک کلمه را از یکدیگر متمایز می‌کنند، در این مرحله این حروف از داخل پرس‌وجوها حذف می‌شوند. همچنین در این فاز، علامت‌های جمع نیز از پرس‌وجوها حذف می‌شوند.
- در مرحله بعدی اعداد موجود در پرس‌وجوها نرمال‌سازی می‌شوند. در حالت کلی، اعداد به کار رفته در پرس‌وجوها به عنوان عبارات بی‌معنی شناخته شده و به جزء در مواردی که حاوی معنای خاصی هستند، حذف می‌گردند. برای مثال در پرس‌وجوی "آزمایش تیروئید T4" یا "ویتامین ب-۶" هر دو عدد وارد شده دارای معنی مشخصی هستند و نباید از پرس‌وجو حذف شوند. اعداد ممکن است به صورت حروفی، عددی به فارسی و انگلیسی ظاهر شوند که تمام آن‌ها تنها به یک شکل عددی به فارسی تبدیل می‌شوند.
- تبدیل تمامی کلماتی که تصریف یک ریشه لغوی هستند به ریشه خود، به گونه‌ای که معنای آن‌ها دست‌نخورده باقی بماند، باعث کاهش گوناگونی مجموعه کلمات به کار رفته در پرس‌وجوها می‌شود و در نتیجه، قدرت تشخیص کلمات افزایش می‌یابد. به عبارتی این فرآیند باعث می‌شود تا تصریف‌های را که به شکل‌های مختلف در داخل یک متن ظاهر شده‌اند، شناسایی شده و تمامی آن‌ها به عنوان یک کلمه در نظر گرفته شده و به یک کلمه نگاشت شوند. در این پژوهش ریشه‌یابی کلمات موجود در پرس‌وجوهای حیطه پزشکی به صورت دستی و در قالب یک جدول از پیش تعریف‌شده، صورت گرفته است.

- سپس به هر یک از کلمات مترادف، برچسب یکسانی زده می‌شود تا کلمات مشابه یک شکل در نظر گرفته شوند. این امر می‌تواند فاصله معنایی کلمات را از یکدیگر کاهش دهد. به همین خاطر از مجموعه دادگان کلمات مترادف زبان فارسی که شامل ۱۵۰۰۰ مدخل و ۱۳۵۰۰۰ واژه است [۱۷]، استفاده شده است؛ بنابراین هر کلمه با پرکاربردترین کلمه مترادف خود جایگزین می‌شود و در صورتی که خود آن کلمه پرکاربردتر از سایر مترادف‌هایش باشد، تغییری پیدا نمی‌کند.
- فاز آخر در پیش‌پردازش دادگان، حذف کلمات و عبارات توقفی از پرس‌وجوها است. کلمات و عبارات توقفی به مجموعه کلمات و عباراتی گفته می‌شود که در درصد بالایی از پرس‌وجوها ظاهر می‌شوند ولی حاوی معنای خاصی نمی‌باشند. این کلمات و عبارات نه تنها بار پردازشی سیستم را افزایش می‌دهند، بلکه در مواردی باعث کاهش دقت سیستم نیز می‌شوند؛ بنابراین، در این مرحله با استفاده از روش‌های مختلف از جمله استفاده از عبارت‌های منظم و جستجو در فهرست از پیش تعیین شده کلمات توقفی، کلمات و عبارات توقفی موجود در حیطه پزشکی شناسایی و از پرس‌وجوها حذف می‌شوند. برای مثال پرس‌وجویی با عنوان "کنترل فشارخون چگونه است"، بعد از حذف کلمات و عبارات توقفی که شامل "چگونه" و "است" می‌باشند، به صورت "کنترل فشارخون" تبدیل می‌شود؛ اما برخی از عبارات نیز با وجود اینکه شامل یک یا چند کلمه توقفی هستند، اما دارای معنی هستند، برای مثال "از بین بردن" که در آن "از" و "بین" کلمه توقفی هستند. در چنین مواردی در صورت حذف کلمات و عبارت توقفی، معنای آن پرس‌وجو از دست خواهد رفت که مطلوب نیست. برای جلوگیری از این موضوع، در این مقاله فهرستی از این کلمات و عبارات با عنوان فهرست استثنائات کلمات توقفی استخراج شده و از ورود آن کلمات به مرحله حذف کلمات و عبارت توقفی جلوگیری می‌شود. در جدول ۱، پیش‌پردازش یک پرس‌وجوی پزشکی آمده است.

جدول ۱: پیش‌پردازش یک پرس‌وجو در مرحله پیش‌پردازش دادگان

پرس‌وجو	اصلاح غلط‌های املایی	یکسان‌سازی نوشتاری	نرمال‌سازی اعداد	ریشه‌یابی	برچسب‌گذاری کلمات مترادف	حذف کلمات توقفی	پرس‌وجوی نهایی
سه مورد از روش‌هایی که می‌توان برای کاهس تپش قلب استفاده کرد؟	تبدیل "کاهس" به "کاهش"	حذف "؟"، حذف "ها" نشانه‌ی جمع	تبدیل سه به "۳" و سپس حذف "۳"	تبدیل "می‌توان" به "توانستن"، تبدیل "کرد" به "کردن"	تبدیل "تپش" به "ضربان"	حذف "مورد"، "از"، "که"، "توانستن"، "برای"، "استفاده" و "کردن"	روش کاهش ضربان قلب

ب) خوشه‌بندی پرس‌وجوها

همان‌طور که بیان شد، امروزه اکثر سیستم‌های توصیه‌گر مبتنی بر کلمات موجود در پرس‌وجوها هستند؛ اما کلمات به تنهایی نمی‌توانند هدف کاربر از جستجو را بیان کنند. این کار باعث می‌شود تا در بسیاری از مواقع توصیه‌هایی که ارائه می‌شوند، متناسب با نیازهای اطلاعاتی کاربران نباشد، اگر چه این پرس‌وجوها دارای کلمات مشابهی با پرس‌وجوهای کاربران باشند. از طرف دیگر، این رویکرد باعث می‌شود که کلمات چندمعنا در نظر گرفته نشوند. برای مثال کاربری واژه شیر را جستجو می‌کند تا از فواید شیر آگاهی پیدا کند، ولی سیستم به وی صفحاتی در رابطه با حیوان شیر برمی‌گرداند. به همین منظور در این مقاله برای رفع این مشکل، علاوه بر در نظر گرفتن ویژگی‌های ساختاری پرس‌وجو از نتایج جستجو نیز به عنوان یک ویژگی دیگر برای خوشه‌بندی پرس‌وجوها استفاده شده است تا دقت سیستم توصیه‌گر را بهبود دارد.

بعد از پیش‌پردازش دادگان، در این مرحله به خوشه‌بندی پرس‌وجوها پرداخته می‌شود. به منظور خوشه‌بندی پرس‌وجوها از دو ویژگی زیر استفاده شده است تا میزان شباهت بین آن‌ها محاسبه شود:

- N-گرام بر روی پرس‌وجوها
- مجموعه آدرس‌های اینترنتی نمایش داده شده در صفحه نتایج پرس‌وجو و رتبه آن‌ها

ویژگی اول: N-گرام بر روی پرس‌وجوها

به طور کلی اگر دو پرس‌وجو دارای کلمات مشابهی باشند، احتمالاً نیاز اطلاعاتی یکسانی را نمایش می‌دهند. بدین منظور، در این پژوهش برای بررسی میزان شباهت لغوی بین پرس‌وجوها، از N-گرام‌ها تا سقف ۳-گرام استفاده شده است. علت استفاده از ۳-گرام این است که بررسی‌هایی که در موتور جستجوی بومی پارسی‌جو انجام شده است، نشان می‌دهد که ۳-گرام برای عبارات فارسی می‌تواند به خوبی مفهوم جستجو

را نشان دهد. برای مثال برای پرس‌وجویی با عنوان "درمان بیماری واریس پا" داریم:
تک-گرام‌ها: "درمان"، "بیماری "واریس"، "پا"
دو-گرام‌ها: "درمان بیماری"، "بیماری واریس"، "واریس پا"
سه-گرام‌ها: "درمان بیماری واریس"، "بیماری واریس پا"

حال برای محاسبه میزان شباهت بین پرس‌وجوها بر اساس شباهت لغوی آن‌ها، ابتدا به ازای هر پرس‌وجو یک بردار از ۳/۲/۱-گرام‌های آن ساخته می‌شود. پس برای مثال بالا داریم:
{درمان، بیماری، واریس، پا، درمان بیماری، بیماری واریس، واریس پا، درمان بیماری واریس، بیماری واریس پا}
سپس برای محاسبه میزان شباهت بین دو پرس‌وجو بر اساس این ویژگی، از رابطه (۱) استفاده می‌شود:

$$Similarity_{Lexical}(q_x, q_y) = \frac{|KW(q_x) \cap KW(q_y)|}{|KW(q_x) \cup KW(q_y)|} \quad (۱)$$

که در آن $KW(q_x)$ و $KW(q_y)$ به ترتیب نشان‌دهنده بردار شامل ۳/۲/۱-گرام‌های مربوط به پرس‌وجوهای q_x و q_y هستند.

برای مثال جدول ۲ را در نظر بگیرید. همان‌طور که می‌بینید، در این جدول دو پرس‌وجو وجود دارد که به ترتیب برابر با "عوامل تشدید سرطان روده" و "علت ایجاد کلیت عصبی روده" است. حال بردار ۳/۲/۱-گرام‌های هر یک از این پرس‌وجوها را محاسبه می‌کنیم.

جدول ۲: محاسبه بردار ۳/۲/۱-گرام دو پرس‌وجوی نمونه

پرس‌وجو	عنوان پرس‌وجو	پیش‌پردازش پرس‌وجو	بردار ۳/۲/۱- گرام‌ها
q_1	عوامل تشدید سرطان روده	عوامل سرطان روده	(عوامل، سرطان، روده، عوامل سرطان، سرطان روده، عوامل سرطان روده)
q_2	علت ایجاد کلیت عصبی روده	علت کلیت عصبی روده	(علت، کلیت، عصبی، روده، علت کلیت، کلیت عصبی، عصبی روده، علت کلیت عصبی، کلیت عصبی روده)

بنابراین داریم:

$KW(q_1) = \{$عوامل، سرطان، روده، عوامل سرطان، سرطان روده، عوامل سرطان روده$\}$
$KW(q_2) = \{$علت، کلیت، عصبی، روده، علت کلیت، کلیت عصبی، عصبی روده، علت کلیت عصبی، کلیت عصبی روده$\}$

پس

$KW(q_1) \cap KW(q_2) = \{$روده$\}$
$KW(q_1) \cup KW(q_2) = \{$علت، کلیت، عصبی، روده، عوامل، سرطان، علت کلیت، کلیت عصبی، عصبی روده، عوامل سرطان، سرطان روده، علت کلیت عصبی، عوامل سرطان روده، کلیت عصبی روده$\}$

حال برای به دست آوردن میزان شباهت بین این دو پرس‌وجو بر اساس بردار ۳/۲/۱-گرام آن‌ها طبق رابطه (۱) داریم:

$$Similarity_{Lexical}(q_1, q_2) = \frac{1}{14}$$

ویژگی دوم: مجموعه آدرس‌های اینترنتی نمایش داده شده در صفحه نتایج پرس‌وجو و رتبه آن‌ها

ویژگی دومی که از آن برای به دست آوردن میزان شباهت بین پرس‌وجوها استفاده می‌شود، k نتیجه اول برگردانده شده در صفحه نتایج یک پرس‌وجو است که در این پژوهش مقدار k برابر با ۱۰ در نظر گرفته شد. علت انتخاب مقدار ۱۰ این است که تحقیقات نشان داده است که ترافیک صفحه اول نتایج در موتور جستجوی گوگل بیشتر از صفحات دیگر است. به طور دقیق‌تر صفحه اول ۹۱/۵٪ از ترافیک را به خود اختصاص داده است، در حالی که صفحات دوم و سوم به ترتیب ۴/۸٪ و ۱/۱٪

از ترافیک را به خود اختصاص می‌دهند [۱۸]. این امر باعث می‌شود تا موتورهای جستجو نتایجی را که به پرس‌وجوی کاربر مرتبط‌تر هستند، در صفحه اول نتایج ارائه دهند. به همین علت در این مقاله نیز از نتایج صفحه اول برای محاسبه میزان شباهت بین پرس‌وجوها استفاده شد. به طور کلی می‌توان گفت در صورتی که ۱۰ نتیجه اول برگردانده شده در پاسخ دو پرس‌وجو مشابه باشند و یا به عبارت بهتر دارای اشتراک باشند، به احتمال زیاد آن دو پرس‌وجو مشابه یکدیگر هستند. برای مثال اگر در ۱۰ نتیجه اول برگردانده شده در پاسخ دو پرس‌وجوی "سلاح اتمی" و "حادثه ناکازاکی" آدرس‌های اینترنتی یکسانی وجود داشته باشد، می‌توان گفت که این دو پرس‌وجو مشابه یکدیگر هستند، علی‌رغم اینکه این دو پرس‌وجو دارای کلمات یکسانی نیستند و از ویژگی اول ما، هیچ امتیازی به دست نمی‌آورند.

البته رتبه نتایج جستجو نیز مهم است. برای مثال فرض کنید فردی به دنبال پیدا کردن فواید میوه سیب باشد و پرس‌وجوی "apple" را در موتور جستجو گوگل وارد کند، آن موقع خواهد دید که، چند نتیجه اول جستجو در رابطه با شرکت اپل و آدرس‌های بعدی در رابطه با میوه سیب است، در حالی‌که اگر پرس‌وجوی "apples" را در گوگل وارد کند، نتایج اول جستجو در رابطه با میوه سیب است؛ بنابراین علی‌رغم این‌که این دو پرس‌وجو از نظر لغوی شبیه یکدیگر هستند، اما هدف کاربران از بیان این دو پرس‌وجو کاملاً متفاوت است. به همین دلیل نتایج جستجو و رتبه آن نتایج می‌تواند در تشخیص میزان شباهت پرس‌وجوها مفید باشد. برای محاسبه شباهت بین دو پرس‌وجو بر اساس ویژگی دوم، ابتدا باید نتایج جستجو بر اساس رتبه‌شان، وزن‌دهی شوند. هر چه که رتبه یک آدرس در صفحه نتایج بالاتر باشد، اهمیت آن و در نتیجه وزن آن هم بیشتر خواهد بود. علت هم این است که موتورهای جستجو، نتایج را از قبل بر اساس میزان ارتباطشان با پرس‌وجوی کاربر رتبه‌بندی می‌کند و اگر یک آدرس در رتبه بالاتری قرار بگیرد، به پرس‌وجوی کاربران مرتبط‌تر است [۱۹، ۱۵].

بدین‌ترتیب به آدرسی که در رتبه i قرار دارد، وزن $w(i) = \frac{1}{2^i}$ می‌دهیم. اکنون برای محاسبه میزان شباهت بین دو پرس‌وجو بر اساس ویژگی دوم بایستی به ازای تک‌تک آدرس‌هایی که در ۱۰ نتیجه اول هر دو پرس‌وجو مشترک هستند، مقدار کسر $\frac{w[r_x(i)]+w[r_y(i)]}{|r_x(i)-r_y(i)|+1}$ محاسبه و نرمال شود. بنابراین طبق رابطه (۲) داریم:

$$Similarity_{Top-K}(q_x, q_y) = \frac{1}{2\sum_{i=1}^{k} w(i)} \sum_{i=1}^{m} \frac{w[r_x(i)] + w[r_y(i)]}{|r_x(i) - r_y(i)| + 1} \quad (۲)$$

که در این‌جا m برابر با تعداد آدرس‌های اینترنتی مشترک در ۱۰ نتیجه اول دو پرس‌وجوی q_x و q_y است. $r_x(i)$ و $r_y(i)$ به ترتیب رتبه آدرس اینترنتی مشترک iام در نتایج پرس‌وجوی q_x و q_y هستند. $w[r_x(i)]$ و $w[r_y(i)]$ نیز نشان‌دهنده وزن آن آدرس‌های مشترک است. همان‌طور که پیش‌تر بیان شد، در این مقاله از ۱۰ نتیجه اول جستجو استفاده می‌شود، بنابراین مقدار k برابر با ۱۰ است و عبارت $\sum_{i=1}^{k} w(i)$ برابر با یک در نظر گرفته می‌شود.

برای مثال، دو پرس‌وجوی "بیماری CCHF" و "تب کریمه‌کنگو" را در نظر بگیرید که ۱۰ نتیجه اول جستجوی آن‌ها در جدول ۳ آمده است.

جدول ۳: ۱۰ نتیجه اول جستجوی دو پرس‌وجو

۱۰ نتیجه اول جستجوی دو پرس‌وجوی نمونه		آدرس‌های یکسان	رتبه آدرس‌های یکسان در صفحه نتایج q_1	رتبه آدرس‌های یکسان در صفحه نتایج q_2
بیماری CCHF (q_1)	تب کریمه‌کنگو (q_2)			
fa.wikipedia.org/wiki/تب_خونریزی‌دهنده_کریمه-کنگو	rooziato.com/1396101740-تب-کریمه-کنگو/	fa.wikipedia.org/…	۱	۲
doc.umsha.ac.ir/uploads/راهنمای_cchf.pdf	https://fa.wikipedia.org/wiki/تب_خونریزی‌دهنده_کریمه-کنگو	doc.umsha.ac.ir/….	۲	۵
bbc.com/persian/science-40077261	tabnak.ir/fa/news/698439/تب-کریمه‌کنگو-تاکنون-جان-سه-نفر-را-گرفت	bbc.com/….	۳	۸
pezeshk.us/?p=30055	zoomit.ir/2017/5/28/157017/what-is-crimean-congo-haemorrhagic-fever-iran/	rooziato.com/….	۵	۱
rooziato.com/1396101740/تب-کریمه-کنگو/	doc.umsha.ac.ir/uploads/راهنمای_cchf.pdf fardanews.com/fa/news/678384/علائم-تب-کریمه-کنگو-را-بشناسید	zoomit.ir/….	۷	۴
home.sums.ac.ir/~aliasghar_su/files/powerpoints/CCHF.ppt	infosalamat.com/blog/articles/همه-چیز-در-مورد-تب-کریمه-کنگو			
zoomit.ir/2017/5/28/157017/what-is-crimean-congo-haemorrhagic-fever-iran/	bbc.com/persian/science-40077261			
treatment.sbmu.ac.ir/uploads/kerime.pdf fararu.com/fa/news/113077/هشدار-جدی-درباره-تب-کریمه-کنگو ana.ir/news/37611	namnak.com/p17247/تب-کریمه-کنگو tasnimnews.com/fa/news/1396/03/28/1440044/آخرین-وضعیت-تب-کریمه-کنگو-در-تهران-تعداد-مبتلایان-در-ایران			

حال برای محاسبه شباهت بین این دو پرس‌وجو بر اساس ویژگی دوم، طبق رابطه (۲) داریم:

$$Similarity_{Top-K}(q_1,q_2) = \frac{1}{2\sum_{i=1}^{10} w(i)} \sum_{i=1}^{5} \frac{w[r_x(i)] + w[r_y(i)]}{|r_x(i) - r_y(i)| + 1}$$

$$= \frac{1}{2}\left[\left(\frac{\frac{1}{2^2}+\frac{1}{2^1}}{1+1}\right) + \left(\frac{\frac{1}{2^5}+\frac{1}{2^2}}{3+1}\right) + \left(\frac{\frac{1}{2^8}+\frac{1}{2^3}}{5+1}\right) + \left(\frac{\frac{1}{2^5}+\frac{1}{2^1}}{4+1}\right) + \left(\frac{\frac{1}{2^7}+\frac{1}{2^4}}{3+1}\right)\right] = 0.32$$

میزان شباهت بین دو پرس‌وجوی q_1 و q_2 بر اساس این ویژگی برابر با ۰/۳۲ است. این دو پرس‌وجو براساس ویژگی اول، هیچ امتیازی کسب نمی‌کنند.

حال برای به دست آوردن شباهت کلی بین دو پرس‌وجو، باید مقادیر به دست آمده از این دو ویژگی را طبق رابطه (۳) و به صورت خطی با یکدیگر ترکیب کرد. بنابراین:

$$Similarity_{Total}(q_x,q_y) = \alpha * Similarity_{Lexical}(q_x,q_y) + \beta * Similarity_{Top-K}(q_x,q_y) \quad (۳)$$

مقدار ضرایب این دو ویژگی در بخش نتایج ارزیابی محاسبه شده است. پس از محاسبه شباهت بین پرس‌وجوها با استفاده از رابطه (۳) با اجرای الگوریتم K-means به خوشه‌بندی پرس‌وجوها با استفاده از این دو ویژگی پرداخته می‌شود. الگوریتم خوشه‌بندی K-means یکی از پرکاربردترین الگوریتم‌های خوشه‌بندی است. از نقاط قوت این الگوریتم این است که علاوه‌بر سادگی و سرعت بالا، برای پایگاه‌داده‌های بزرگ، کارا و ارتقاپذیر است. علت بالا بودن سرعت الگوریتم K-means، کم بودن تعداد تکرارهای این الگوریتم نسبت به تعداد داده‌هایی است که باید خوشه‌بندی شوند؛ به عبارت دیگر پیچیدگی محاسباتی این الگوریتم برابر با $O(IKN)$ است که N تعداد کل پرس‌وجوها، K تعداد خوشه‌ها و I تعداد تکرارهای الگوریتم است [۲۰]. به عبارت دقیق‌تر K-means داده‌ها را درون K خوشه منحصر به فرد تقسیم می‌کند، به

طوری‌که داده‌های درون یک خوشه بیشترین شباهت ممکن را با یکدیگر و بیشترین تفاوت را با خوشه‌های دیگر داشته باشند.

ج) توصیه پرس‌وجو

حال بعد از خوشه‌بندی پرس‌وجو، با وارد شدن پرس‌وجوی کاربر به سیستم، ابتدا وجود یا عدم وجود آن در خوشه‌های پرس‌وجو، مورد بررسی قرار می‌گیرد. در صورتی که این پرس‌وجو قبلاً توسط کاربران وارد شده باشد و در گزارش‌های قبلی موتور جستجو موجود باشد، سیستم ابتدا خوشه‌ای را که پرس‌وجوی موردنظر به آن تعلق دارد، پیدا می‌کند. سپس برای توصیه پرس‌وجو به وی، k شبیه‌ترین پرس‌وجوی موجود در آن خوشه که بیشترین شباهت را به مرکز خوشه دارند و از قبل محاسبه و ذخیره شده‌اند، انتخاب و توصیه می‌شوند؛ اما اگر پرس‌وجوی کاربر از قبل در سیستم وجود نداشته باشد، باید نزدیک‌ترین خوشه به این پرس‌وجو پیدا شود. برای این منظور ابتدا با استفاده از رابطه (۳)، میزان شباهت این پرس‌وجو با مراکز خوشه‌ها به دست می‌آید. سپس این پرس‌وجو به خوشه‌ای که بیشترین شباهت را به مرکز آن دارد، اضافه می‌شود. اکنون مانند حالت قبل، K پرس‌وجویی که بیشترین شباهت را به مرکز خوشه دارند، به کاربر توصیه می‌شوند. در نهایت با توجه به اینکه پرس‌وجوی جدیدی به این خوشه اضافه شده است، این امکان وجود دارد که مرکز آن خوشه تغییر کند، بنابراین مرکز خوشه مجدداً محاسبه می‌شود. این کار به صورت دوره‌ای و برون‌خط و در ساعاتی که بار کاری سیستم کم است، انجام می‌شود. اگرچه هر یک از روش‌های عنوان‌شده در بخش‌های قبل با رویکرد متفاوتی به توصیه پرس‌وجو به کاربران می‌پردازند، اما عدم وجود مجموعه دادگان عمومی از رفتار کاربران به دلیل رقابت سایت‌ها، حفظ محرمانگی جستجوی کاربران و موضوعات تجاری، باعث دشوار شدن عملیات ارزیابی و مقایسه این روش‌ها می‌شود؛ بنابراین عموماً ارزیابی‌ها به بررسی نمونه‌های استخراج‌شده از گزارش‌های سایت‌ها توسط افراد خبره و محاسبه معیارهای ارزیابی، محدود می‌شود.

نتایج

در این بخش به ارزیابی روش پیشنهادی در این مقاله می‌پردازیم و نتایج به دست آمده از ارزیابی‌ها را بیان می‌کنیم.

الف) مدل ارزیابی

برای ارزیابی مدل ابتدا باید تأثیر هر کدام از ضرایب ویژگی‌ها را به دست آورد. برای این منظور، با استفاده از الگوریتم grid exhaustive search مقدار بهینه هر کدام از این ضرایب به دست می‌آید. حال با استفاده از روش اعتبارسنجی متقابل K بخشی به ارزیابی مدل پرداخته می‌شود. در این روش، نمونه‌های آموزشی به صورت تصادفی به K بخش مساوی تقسیم می‌شوند، سپس یک بخش به عنوان داده ارزیابی و $k-1$ بخش به عنوان داده آموزشی استفاده می‌شود. این فرایند k بار تکرار می‌شود، بدین‌صورت از تمامی نمونه‌ها برای آموزش مدل استفاده شده و هر نمونه نیز یک‌بار برای ارزیابی مورد استفاده قرار گرفته است. در نهایت، میانگین نتایج به دست آمده در هر دور به عنوان تخمین نهایی در نظر گرفته می‌شود. در این مقاله نیز همانند اکثر تحقیقات علمی از مقدار $k=10$ یعنی اعتبارسنجی متقابل ۱۰ بخشی استفاده شد [۲۱].

محققان برای ارزیابی روش‌های پیشنهادی خود در زمینه سیستم‌های توصیه‌گر از معیارهای متفاوتی استفاده می‌کنند. انتخاب این معیارها به ماهیت سیستم توصیه‌گر، مجموعه داده مورد استفاده آن و همچنین تکنیکی که از آن بهره می‌برد، بستگی دارد. به عنوان مثال در یک فروشگاه اینترنتی مثل آمازون، در صورتی که بازدید کاربران از یک کالا با خرید آن همراه باشد، یک موفقیت تلقی می‌شود؛ اما در یک سایت آموزشی مرور یک آموزش و یا دانلود آن به وسیله کاربر، می‌تواند به عنوان موفقیت سیستم در نظر گرفته شود. به طور کلی دو نوع معیار ارزیابی در سیستم‌های توصیه‌گر وجود دارد: ارزیابی برخط، ارزیابی برون‌خط. در رویکرد برخط سیستم به طور واقعی توسط کاربران مورد آزمایش قرار می‌گیرد و در نتیجه از دقت بالایی برخوردار است؛ اما در این روش سیستم احتیاج به اجتماع بزرگی از کاربران دارد که بتواند به نتایج قابل استنادی دست پیدا کند که این موارد به‌علاوه هزینه زیاد این رویکرد، می‌تواند به عنوان عیب بزرگ آن ذکر شود. در رویکرد برون‌خط از دادگان مربوط به رفتار گذشته کاربران برای ارزیابی سیستم استفاده می‌شود [۲۲]. در این مقاله نیز از رویکرد ارزیابی برون‌خط استفاده شده است. به منظور ارزیابی روش پیشنهادی خود در این مقاله از سه معیار ارزیابی دقت (Precision)، پوشش (Coverage) و کیفیت خوشه‌بندی استفاده شد.

معیار دقت: این معیار، توانایی سیستم توصیه‌گر را برای تولید توصیه‌های دقیق نشان می‌دهد که با استفاده از رابطه (۴) محاسبه می‌شود:

$$Precision = \frac{TP}{TP + FP} \quad (۴)$$

که در اینجا، معیار TP بیانگر تعداد پرس‌وجوهای مرتبطی است که به درستی مرتبط تشخیص داده شده‌اند. معیار FP نیز بیانگر تعداد پرس‌وجوهای نامرتبطی است که به اشتباه مرتبط تشخیص داده شده‌اند.

معیار پوشش: این معیار درصد توصیه‌هایی را نشان می‌دهد که سیستم توصیه‌گر قادر به ارائه حداقل یک توصیه برای آن‌ها لست.

معیار کیفیت خوشه‌بندی: برای ارزیابی کیفیت خوشه‌بندی پرس‌وجوها با استفاده از رابطه (۵)، نسبت فاصله هر پرس‌وجو تا مرکز خوشه‌ای که به آن تعلق دارد (d_{i,c_i}) به فاصله آن تا مراکز خوشه‌های دیگر (d_{i,c_j}) محاسبه شده است. به عبارتی این مقدار برای هر پرس‌وجو، به تعداد خوشه‌ها محاسبه می‌شود.

$$\delta = \frac{d_{i,c_i}}{d_{i,c_j}} \quad (۵)$$

ب) دادگان ارزیابی

دادگان مورد استفاده در این پژوهش، گزارش‌های مربوط به ۶ روز موتور جستجوی بومی پارسی‌جو (parsijoo.ir) در بازه ۱۳۹۵/۲/۲۳ تا ۱۳۹۵/۲/۲۸ است که شامل ۷۹۹۱۹۰ رکورد جستجو است. هر رکورد از دادگان که در یک سطر آمده است، نشان‌دهنده یک جستجوی کاربر است که شامل شناسه کاربر، زمان و تاریخ ارسال پرس‌وجو، پرس‌وجوی ارسال‌شده، آدرس IP دستگاه کاربر، ۱۰ نتیجه اول برگردانده‌شده در پاسخ به آن پرس‌وجو و آدرس اینترنتی کلیک‌شده توسط کاربر (در صورت کلیک) است. همان‌طور که در مرحله پیش‌پردازش اشاره شد، دادگان قبل از استفاده، جهت نرمال‌سازی و رفع برخی از نواقص، مورد پیش‌پردازش قرار گرفته‌اند. ۱۰۰۰ رکورد از این دادگان به صورت تصادفی برای مجموعه آزمایشی انتخاب شدند و مابقی رکوردها، دادگان آموزشی را شکل می‌دهند.

ج) نتایج ارزیابی

روش پیشنهادی با استفاده زبان برنامه‌نویسی جاوا نوشته شده است. برای ارزیابی تأثیر ضرایب ویژگی‌ها از الگوریتم exhaustive grid search در بازه‌های مختلف استفاده شده است. همان‌طور که در جدول ۴ آمده است، بهترین مقادیر پارامترها برای α و β به ترتیب برابر با ۰/۳ و ۰/۷ به دست آمده است.

جدول ۴: نتایج اجرای الگوریتم grid search برای تعیین ضرایب ویژگی‌ها

ضرایب	محدوده	گام	بهترین مقدار
α	[۱-۰]	۰/۱	۰/۳
β	[۱-۰]	۰/۱	۰/۷

به منظور ارزیابی تأثیر تعداد توصیه‌ها بر دقت سیستم، تعداد توصیه‌های ارائه‌شده برای دادگان مجموعه آزمایشی را در بازه [۵-۱۰] تغییر داده‌ایم. همان‌طور که در جدول ۵، مشاهده می‌شود، با افزایش تعداد توصیه‌ها، دقت سیستم کاهش پیدا کرد. علت آن هم این است که هر چه تعداد توصیه‌ها بیشتر شود، توصیه‌هایی ارائه می‌شود که شباهت کمتری با پرس‌وجوی کاربر دارند؛ بنابراین هر چه تعداد توصیه‌ها کمتر باشد، توصیه‌های با دقت بیشتری ارائه می‌شود. در اینجا بیشترین دقت با تعداد ۵ توصیه به دست آمد و بدترین نتیجه هنگامی است که تعداد توصیه‌ها برابر با ۱۰ باشد.

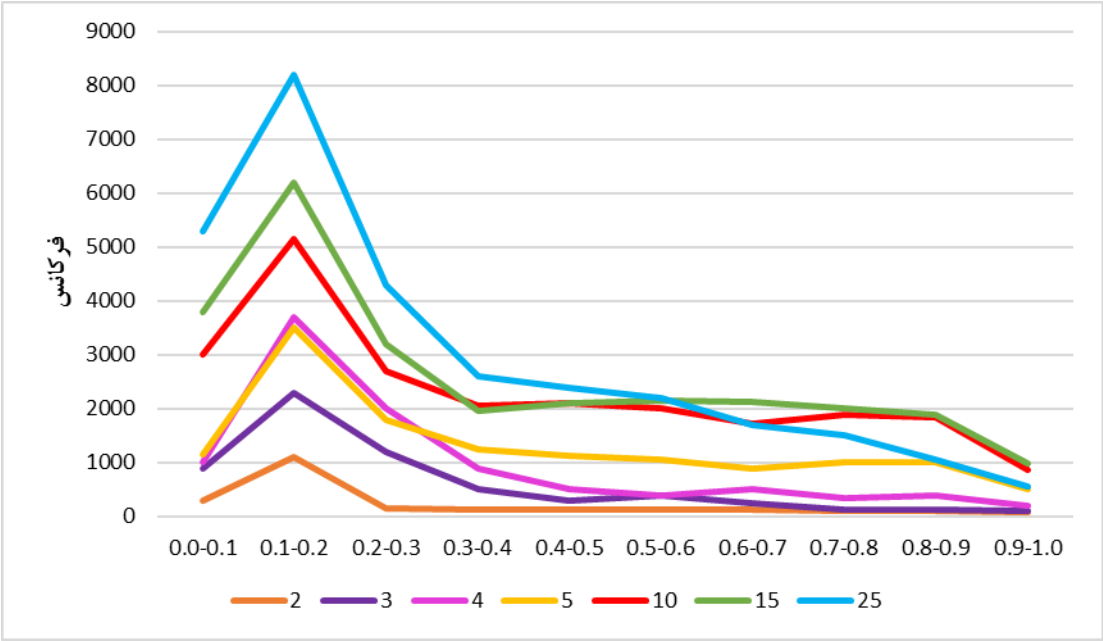
جدول ۵: ارزیابی مدل پیشنهادی با استفاده از معیار دقت و پوشش

تعداد توصیه / معیار	N=۵	N= ۶	N= ۷	N= ۸	N= ۹	N= ۱۰
TP	۳۸۶۲	۴۴۳۱	۵۰۳۳	۵۵۴۴	۵۹۱۲	۶۳۰۹
FP	۱۱۳۸	۱۵۶۹	۱۹۶۷	۲۴۵۶	۳۰۸۸	۳۶۹۱
دقت	٪۷۷/۲۴	٪۷۳/۸۵	٪۷۱/۹	٪۶۹/۳	٪۶۵/۶۸	٪۶۳/۰۹
پوشش	٪۱۰۰					

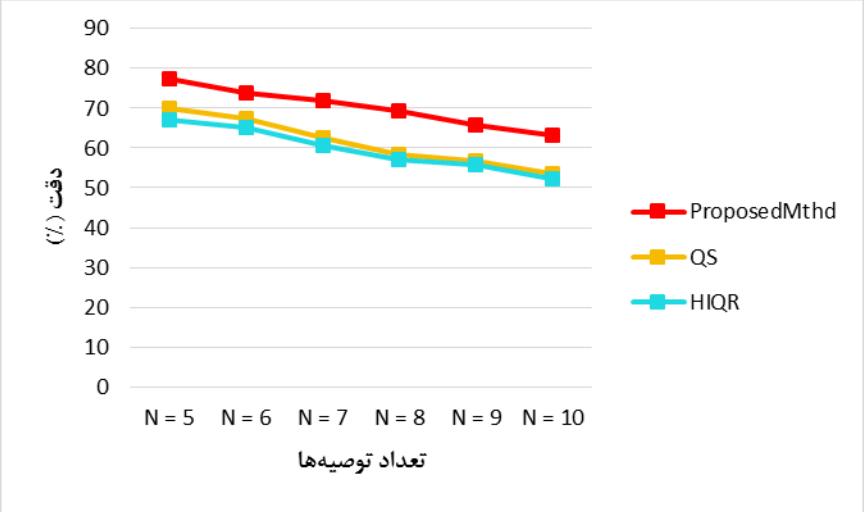
همان‌طور که در بخش قبل، بیان شد، هنگام ارائه توصیه به کاربر، نزدیک‌ترین پرس‌وجوها به مرکز دسته به وی توصیه می‌شوند، از آن‌جایی که همیشه حداقل یک پرس‌وجو برای وی برگردانده می‌شود، بنابراین پوشش مدل برابر با ۱۰۰٪ است.

برای محاسبه معیار کیفیت خوشه‌بندی با استفاده از رابطه (۵)، تعداد رخداد مقادیر محاسبه شده δ به فواصل ۰/۱ برای تمامی دادگان موجود در مجموعه آزمایشی به دست آمده است. توزیع رسم‌شده در شکل ۱ نشان‌دهنده این است که خوشه‌بندی پرس‌وجوها مناسب بوده است، زیرا بیشترین فرکانس در بازه‌های کمتر اتفاق افتاده است؛ به عبارت دیگر پرس‌وجوها به درستی به مرکز خوشه خود نسبت به خوشه‌های دیگر نزدیک‌تر هستند.

در شکل ۲ نیز خلاصه‌ای از نتایج مقایسه روش پیشنهادی با روش‌های مطرح در این زمینه آمده است که در بخش بحث و نتیجه‌گیری به تفصیل به توضیح آن پرداخته شده است.



شکل ۱: نسبت δ به ازای تعداد خوشه‌های متفاوت



شکل ۲: مقایسه دقت روش پیشنهادی با دیگر پژوهش‌ها

بحث و نتیجه‌گیری

گسترش روزافزون اطلاعات در وب، فرآیند تصمیم‌گیری و انتخاب را برای بسیاری از کاربران دشوار کرده است. برای رفع این مشکل و کمک به کاربران در انجام انتخاب‌هایشان، سیستم‌های توصیه‌گر به وجود آمدند. در سال‌های اخیر سیستم‌های توصیه‌گر به عنوان یکی از بخش‌های جدایی‌ناپذیر و حیاتی موتورهای جستجو، شبکه‌های اجتماعی و به طور کلی هر سیستم برخطی که با کاربران در ارتباط است، شده است. از این سیستم‌ها با هدف کمک به کاربران برای دسترسی هر چه سریع‌تر و بهتر به اطلاعات موردنیازشان بهره‌بری می‌شود. یکی از کاربردهای توصیه‌گرها، در حوزه جستجوی مفاهیم پزشکی است. رشدنمایی اطلاعات پزشکی و بهداشتی، یکی از چالش‌هایی است که کاربران با آن روبه‌رو هستند. از طرف دیگر دانش کم اغلب افراد در حیطه پزشکی آن‌ها را برای یافتن اطلاعات موردنیازشان با مشکل روبه‌رو کرده است. توصیه پرس‌وجوهای مناسب و معتبر به کاربران، می‌تواند باعث افزایش رضایتمندی کاربر از سیستم شود و کیفیت و کارایی کل سیستم را افزایش دهد، چرا که کاربران را آسان‌تر و سریع‌تر به نتایج مطلوبشان می‌رساند.

هدف پژوهش حاضر، ارائه روشی ترکیبی با بهره‌گیری از ویژگی‌های محتوایی پرس‌وجوها به همراه نتایج جستجو، به منظور توصیه پرس‌وجو در حیطه پزشکی است. در گام اول داده‌های پژوهش، مورد پیش‌پردازش قرار گرفتند تا نواقص موجود در آن‌ها رفع شده و یکسان‌سازی شوند. در گام دوم با استفاده از دو ویژگی N-گرام بر روی پرس‌وجوها و ۱۰ نتیجه اول نتایج جستجو به خوشه‌بندی پرس‌وجوها پرداخته شد و در نهایت، در گام آخر سیستم می‌تواند با ورود پرس‌وجوی جدید به توصیه پرس‌وجو بپردازد. دادگان مورد استفاده در این پژوهش شامل دادگان مربوط به شش روز جستجوی کاربران در موتور جستجوی بومی پارسی‌جو بوده است که قبل از استفاده مورد پیش‌پردازش قرار گرفته‌اند. نتایج ارزیابی نشان می‌دهد که روش پیشنهادی در مقایسه با روش‌های مطرح در این حیطه از دقت بالاتری برخوردار است. به طور دقیق‌تر، دقت روش پیشنهادی برای توصیه ۵ پرس‌وجو حدود ۷۸٪ و پوشش آن ۱۰۰٪ است که در مقایسه با روش ارائه شده توسط Zeng و همکاران [۱۱] بیش از ۱۰٪ و در مقایسه با روش Zhang و همکاران [۱۵] حدود ۹٪ بهبود داشته است.

دقت الگوریتم پیشنهادی با دو روش توصیه پرس‌وجوی مطرح‌شده در بالا مقایسه شده است. برای اشاره به این الگوریتم‌ها به ترتیب از الگوریتم HIQR و QS الگوریتم استفاده شد. برای مقایسه، ابتدا این روش‌ها را پیاده‌سازی و سپس آن‌ها را بر روی پایگاه داده به کار برده و دقت آن‌ها را محاسبه نمودیم. در پیاده‌سازی الگوریتم HIQR از آن‌جایی که سلسله‌مراتبی از مفاهیم پزشکی به فارسی وجود ندارد، ابتدا با استفاده از مشاوره‌های بالینی با افراد خبره، سلسله‌مراتبی از مفاهیم پزشکی استخراج شد و از کلیک کاربر بر روی نتایج نیز برای ساخت مدل استفاده شد. در پیاده‌سازی QS، وزن عبارات موجود در اسناد کلیک‌شده کاربران را به دست آوردیم. برای این منظور از کتابخانه آماده Jsoup برای خزش آدرس‌های کلیک‌شده کاربران استفاده شد.

خلاصه نتایج که در شکل ۲ آمده است، نشان می‌دهد که به طور میانگین روش پیشنهادی ما پرس‌وجوهای دقیق‌تری را در مقایسه با دو روش موجود ارائه می‌دهد. همان‌طور که مشخص است، به ازای تعداد توصیه‌های کم، روش پیشنهادی دقت بهتری دارد، علت آن هم این است که هر چه تعداد توصیه‌ها افزایش پیدا کند، فاصله توصیه‌هایی که ارائه می‌شوند تا مرکز خوشه نیز زیاد می‌شود و در نتیجه احتمال توصیه پرس‌وجوهایی که شباهت کمتری با پرس‌وجوی ورودی دارند، افزایش می‌یابد.

همچنین روش پیشنهادی با روش مطرح شده توسط Boldi و همکاران در [۱۳] نیز مقایسه شده است. برای پیاده‌سازی این روش ابتدا پرس‌وجوهایی که کاربران در مدت زمان ۳۰ دقیقه وارد می‌کنند، به عنوان پرس‌وجوهای نشست جاری کاربر در نظر گرفته می‌شوند. سپس گراف حاصل از توالی پرس‌وجوهای ارائه‌شده در نشست‌های مختلف کاربران رسم می‌شود. از توالی موجود در این گراف برای توصیه پرس‌وجو به پرس‌وجوهای ورودی استفاده می‌شود. دقت به دست آمده از این روش بر روی دادگان ارزیابی این پژوهش برابر با ۵۳٪ است. علت دقت کم این روش این است که این روش تنها قادر به ارائه توصیه برای پرس‌وجوهایی است که قبلاً حداقل یک بار در گزارش‌های موتور جستجو وارد شده باشند.

بنابراین نتایج نشان می‌دهند که استفاده توامان از ویژگی‌های ساختاری و محتوایی پرس‌وجوها در کنار بهره‌گیری از نتایج حاصل از جستجوی پرس‌وجوها می‌تواند میزان شباهت بین دو پرس‌وجو را بهتر محاسبه کرده و اهداف کاربران از بیان پرس‌وجو را بهتر تشخیص دهند. روش‌های قبلی فقط از یک ویژگی برای به دست آوردن میزان شباهت بین پرس‌وجوها استفاده می‌کردند، به همین خاطر هدف کاربر از بیان یک

پرس‌وجو را به درستی تشخیص نمی‌دادند، این موضوع بیشتر در زمانی که پرس‌وجوی کاربر حاوی کلمات چندمعنا باشد باعث ایجاد مشکل می‌شود؛ اما روش پیشنهادی با در نظر گرفتن ویژگی‌های N–گرام بر روی پرس‌وجوها و نتایج جستجو می‌تواند این مشکل را در توصیه پرس‌وجوها برطرف کند.

مطالعات آینده می‌توانند بر ایجاد، بهبود و ارتقای الگوریتم‌ها و تکنیک‌های موجود در زمینه توصیه پرس‌وجوها تمرکز کنند. همچنین می‌توان با ساخت و توسعه دادگان استاندارد برای زبان فارسی، حجم عملیات محاسباتی برای پیش‌پردازش دادگان را به حداقل رساند. استفاده از ابزارهای معنایی و مجموعه داده‌های مرتبط با مفاهیم، مخفف‌ها و علائم پزشکی بر دقت سیستم توصیه پرس‌وجو می‌افزاید. علاوه‌بر این می‌توان از دیگر الگوریتم‌های یادگیری ماشین از جمله ماشین بردار پشتیبان چند هسته‌ای، یادگیری تقویتی و شبکه‌های عصبی نیز برای ساخت مدل استفاده کرد.

References

1. Zhu X, Guo J, Cheng X, Lan Y. More than relevance: high utility query recommendation by mining users' search behaviors. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management; 2012 Oct-Nov 2-2; USA: ACM; 2012. p. 1814-18.

2. Huang Z, Cautis B, Cheng R, Zheng Y. KB-Enabled Query Recommendation for Long-Tail Queries. Proceedings of the 25th ACM International on Conference on Information and Knowledge Management; 2016 Oct 24– 28; Indiana, USA: ACM; 2016.

3. Yang L, Mei Q, Zheng K, Hanauer DA. Query Log Analysis of an Electronic Health Record Search Engine. AMIA Annu Symp Proc 2011; 2011: 915–24.

4. Melville P, Sindhwani V. Recommender Systems. In Encyclopedia of Machine Learning. USA: Springer; 2011.

5. Shokouhi M. Learning to personalize query auto-completion. Proceedings of the 36th Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR); 2013 Jul-Aug 28 -1; Dublin, Ireland: ACM; 2013. p. 103-12.

6. Soo J. A non-learning approach to spelling correction in web queries. Proceedings of the 22th International Conference on World Wide Web; 2013 May 13-17; Rio de Janeiro, Brazil: ACM; 2013. p. 101-2.

7. Xu J, Croft WB. Query expansion using local and global document analysis. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; 1996 Aug 18–22; Zurich, Switzerland ACM; 1996. p. 4-11.

8. Li L, Yang Z, Liu L, Kitsuregawa M. Query-URL bipartite based approach to personalized query recommendation. Proceedings of the 23th National Conference on Artificial Intelligence; 2008 Jul 13-17; Chicago, Illinois: Association for the Advancement of Artificial Intelligence; 2008. p. 1189-94.

9. Wang J, Huang JZ, Guo J, Lan Y. Query ranking model for search engine query recommendation. International Journal of Machine Learning and Cybernetics 2017;8(3):1019-38.

10. Guo J, Zhu X, Lan Y, Cheng X. Modeling users' search sessions for high utility query recommendation. Information Retrieval Journal 2017;20(1):4-24.

11. Zeng QT, Crowell J, Plovnick RM, Kim E, Ngo L, Dibble E. Assisting consumer health information retrieval with query recommendations. J Am Med Inform Assoc 2006;13(1):80-90.

12. Spink A, Wolfram D, Jansen MB, Saracevic T. Searching the web: The public and their queries. Journal of the Association for Information Science and Technology 2001; 52(3): 226-34.

13. Boldi P, Bonchi F, Castillo C, Donato D, Gionis A, Vigna S. The query-flow graph: model and applications. In Proceedings of the 17th ACM Conference on Information and Knowledge Management; 2008 Oct 26-30; Napa Valley, California, USA: ACM; 2008. p. 609-18.

14. Mei Q, Zhou D, and Church K. Query suggestion using hitting time. In Proceedings of 17th ACM Conference on Information and Knowledge Management; 2008 Oct 26–30; Napa Valley, California, USA: ACM Press; 2008. p. 469–78.

15. Zhang B, Zhang B, Zhang S, Ma C. Query recommendation based on irrelevant feedback analysis. 8th International Conference on Biomedical Engineering and Informatics (BMEI); 2015 Oct 14-16. Shenyang, China IEEE; 2015. p. 644-8.

16. Mitsui M, Shah C. Multi-word generative query recommendation using topic modeling. In Proceedings of 10th ACM International Conference on Recommender Systems; 2016 Sep 15–19; Boston, Massachusetts, USA: ACM; 2016. p. 27-30.

17. Khodaparasti F. Comprehensive dictionary of synonymous and opposite vocabulary of Persian language. [cited 2017 Sep 1]. Available from: http://dadegan.ir/catalog/D3911124a.

18. Hodgdon M. Value of organic first-page results. [cited 2017 Sep 5]. Available from: https://www.infront.com/blogs/the-infront-blog/2015/6/17/value-of-first-page-google-results.

19. Hong Y, Vaidya J, Lu H. Search engine query clustering using top-k search results. IEEE/WIC/ACM International Conferences on Web Intelligence and

Intelligent Agent Technology; 2011 Aug 22-27; Lyon, France: IEEE; 2011.p. 112-9.
20. Gaur D, Gaur S. Comprehensive Analysis of Data Clustering Algorithms. In: Jung HK, Kim JT, Sahama T, Yang CH, editors. Future Information Communication Technology and Applications: ICFICE 2013. Dordrecht: Springer Netherlands; 2013. p. 753-62.
21. Parra-Santander D, Brusilovsk P. Improving Collaborative Filtering in Social Tagging Systems for the Recommendation of Scientific Articles. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology; 2010 Aug-Sept 31-3; 2010 Toronto, ON, Canada: IEEE 2010; p.136-42.
22. De Moor K, De Pessemier T, Mechant P, Courtois C, De Marez AJL, Martens L. Users' (Dis)satisfaction with the PersonalTV Application: Combining Objective and Subjective Data. ACM Computers in Entertainment 2011; 9(3): 1-22.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2017; 4(2): 201-215

A Hybrid Method for Query Recommendation in Recommender Systems

Esmaeeli Gohari Elham[1], Zarifzadeh Sajjad[2]*, Hasanvand Saeed[3]

• **Received:**, 28 Sep 2017 • **Accepted:** 27 Nov 2017

Introduction: The rapid growth of information in the Internet and high informational overload has created an important challenge for users in accessing their needed information. Nowadays, query recommendation systems have become a major part of information retrieval systems. One of the applications of these recommendation systems is in medical sciences. Through applying personalization approach, these systems attempt to decrease the problem of informational overload in Web and to accelerate users' medical search.

Method: In this applied and descriptive study, by using the lexical features and search results of queries, we tried to propose a method that helps users to access their desired information in a short time while maintaining the lexical relationship with the original query. The popular k-means algorithm was used to cluster queries. The implementation of the proposed method was done by using java programming language and in NetBeans IDE software.

Results: According to the proposed method, the combined use of the lexical features and search results of queries leads to useful information for detecting similar queries. Since there is a possibility that a query contains multi-meaning words, using search results can be useful in identifying the user's intent of a query.

Conclusion: Evaluation of the proposed model with the real search log of the Parsijoo search engine indicated the precision rate of 77.4% for this method that in comparison to other methods shows 10% improvement of precision.

Keywords: Medical recommendation system, query recommendation, Search engine, Information retrieval

• **Citation:** Esmaeeli Gohari E, Zarifzadeh S, Hasanvand S. A Hybrid Method for Query Recommendation in Recommender Systems. Journal of Health and Biomedical Informatics 2017; 4(2): 201-215.

1. M. Sc. in Computer Engineering, Computer Engineering Dept., Technical and Engineering Campus, Yazd University, Yazd, Iran
2. Ph. D. in Computer Engineering, Assistant Professor of Computer Engineering, Computer Engineering Dept., Technical and Engineering Campus, Yazd University, Yazd, Iran
3. M. Sc. in Computer Engineering, School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran

***Correspondence**: Faculty Electrical and Computer Engineering, Yazd University, Yazd, Iran.

• **Tel:** 03538200144 • **Email:** szarifzadeh@yazd.ac.ir