اثربخشی بسط پرس‌وجو مبتنی بر خوشه‌بندی اسناد شبه‌بازخورد با الگوریتم K-NN

رضا خدایی[۱]، دانشجوی کارشناسی ارشد؛ محمدعلی بالافر[۲]، استادیار؛ سیدناصر رضوی[۳]، استادیار

۱- دانشکده مهندسی برق و کامپیوتر - دانشگاه تبریز - تبریز - ایران - r.khodaei91@ms.tabrizu.ac.ir
۲- دانشکده مهندسی برق و کامپیوتر - دانشگاه تبریز - تبریز - ایران - balafarila@tabrizu.ac.ir
۳- دانشکده مهندسی برق و کامپیوتر - دانشگاه تبریز - تبریز - ایران - razavi@iust.ac.ir

چکیده: بسط پرس‌وجو یکی از روش‌های مؤثر در بهبود اثربخشی نتایج بازیابی اطلاعات است. روش بازخورد شبه‌مرتبط (PRF) فرض می‌کند که اسناد رتبه‌بالا از نتایج اولیه بازیابی‌شده مرتبط به پرس‌وجو است و تعدادی کلمه مرتبط را از اسناد رتبه‌بالا برای بسط انتخاب می‌کند. وجود اسناد نامرتبط در بین اسناد رتبه بالا محققان را به ارائه روش‌هایی برای انتخاب بهترین اسناد به‌عنوان منبع برای انتخاب کلمه بسط سوق می‌دهد که انتخاب بهترین اسناد برای استخراج کلمات مرتبط برای بسط، موضوع مهمی در روش‌های بسط پرس‌وجو هست. در این مقاله، از خوشه‌بندی اسناد شبه‌بازخورد (CPRF) حاصل از نتایج اولیه، بر اساس شباهت مبتنی بر کلمه برای قرار دادن شبیه‌ترین اسناد کنار هم استفاده می‌شود. تعدادی از خوشه‌ها طبق محتوایشان به‌عنوان خوشه‌های بازخورد انتخاب می‌شوند و از خوشه‌های بازخورد، اسناد رتبه‌بالا به‌عنوان اسناد بازخورد انتخاب می‌شوند. سپس، یک سند ترکیبی از روی اسناد انتخابی تشکیل می‌شود و کلمات سند ترکیبی بر اساس تابع رتبه‌بندی TF-IDF مرتب می‌شوند. بعد، کلمات رتبه بالا برای بسط انتخاب می‌شوند. آزمایش‌های انجام‌گرفته روی مجموعه داده پزشکی MED نشان می‌دهد روش پیشنهادی معیار متوسط میانگین دقت (MAP) بالاتری نسبت به روش بازخورد شبه‌مرتبط (PRF) دارد.

واژه‌های کلیدی: بازیابی اطلاعات، بسط پرس‌وجو، بازخورد شبه‌مرتبط، کلمات بسط، اسناد بازخورد.

Effectiveness of Query Expansion based on Clustering of Pseudo-Feedback Documents with K-NN Algorithm

Reza Khodaei[1], MSc Student; Mohammad Ali Balafar[2], Assistant Professor; Seyyed Naser Razavi[3], Assistant Professor

1- Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran, Email: r.khodaei91@ms.tabrizu.ac.ir
2- Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran, Email: balafarila@tabrizu.ac.ir
3- Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran, Email: razavi@iust.ac.ir

Abstract: Query expansion is one of the effective approaches for improving effectiveness of information retrieval. Pseudo-relevance feedback (PRF) supposes that top-ranked documents of retrieval from primary retrieved results are relevant to the query and selects some relevant terms from top-ranked documents for expansion. Existence of noisy documents in the top ranked documents drives researchers toward inventing approaches for selecting best documents as source for selecting expansion terms. The selection of the best documents for extraction of relevant terms for expansion is the most important issue in query expansion. In this paper, we propose clustering of pseudo feedback documents (CPRF), selected from primary results, based on cosine similarity to place the most similar documents beside each other. Some clusters are selected as feedback ones based on their inner content and some top ranked documents of them are selected as feedback documents. A combined document is constructed from selected documents and terms of combined document are ranked by term frequency-inverse document frequency (TF-IDF) schema. High ranked terms are selected for query expansion. Experimental results over MED collection shows postulated approach overcome pseudo relevance feedback approach respect to average mean accuracy (MAP).

Keywords: Information retrieval, Query expansion, Pseudo-relevance feedback, Expansion terms, Relevant documents.

تاریخ ارسال مقاله: ۱۳۹۳/۰۸/۱۴
تاریخ اصلاح مقاله: ۱۳۹۴/۰۲/۱۷، ۱۳۹۴/۰۲/۲۹ و ۱۳۹۴/۰۴/۲۹
تاریخ پذیرش مقاله: ۱۳۹۴/۰۵/۱۱
نام نویسنده مسئول: محمدعلی بالافر
نشانی نویسنده مسئول: ایران - تبریز - بلوار ۲۹ بهمن - دانشگاه تبریز - دانشکده مهندسی برق و کامپیوتر.

۱- مقدمه

بازیابی اطلاعات به‌عنوان بخش مهمی از علم کامپیوتر، پاسخ‌گوی نیازهای اطلاعاتی کاربران است. امروزه موتور جستجوها به‌صورت فزاینده‌ای برای جستجوی اطلاعات توسط کاربران اینترنت مورد استفاده قرار می‌گیرند. هدف موتور جستجوها برآورده کردن نیازهای اطلاعاتی کاربران هست. نیاز اطلاعاتی کاربر، به‌صورت ساختاری از کلمات که پرس‌وجو نام دارد به موتور جستجوها ارائه می‌شود. کاربران، فرآیند جستجو را با ارسال پرس‌وجو شروع می‌کنند و انتظار دریافت مرتبط‌ترین اطلاعات را دارند.

جستجو بر اساس پرس‌وجوی ارسالی کاربر انجام می‌شود و این امکان وجود دارد که پرس‌وجوی ارسالی کاربر دقیق و عاری از ابهام نباشد. معمولاً مشکلات زبان طبیعی، انتخاب کلمات نامناسب در پرس‌وجو، وجود کلمات مبهم با چند معنی [۱] و کوتاهی پرس‌وجوها [۲] مشکلات موجود در پرس‌وجوها هستند که معمولاً کاربران مبتدی با این مشکلات روبرو می‌شوند. تقریباً ۷ تا ۲۳ درصد از پرس‌وجوهای موتور جستجوها کمتر از ۳ کلمه دارند [۳]. بسط پرس‌وجو یکی از روش‌های انطباق پرس‌وجو است که سعی در برطرف‌کردن مشکلات مذکور را دارد. در بسط پرس‌وجو تعدادی کلمه مرتبط به موضوع پرس‌وجو اضافه می‌شوند به‌طوری‌که بازیابی با پرس‌وجوی بسط داده‌شده کارایی را افزایش می‌دهد.

در بازیابی اطلاعات، روش بازخورد شبه‌مرتبط[۱] (PRF) [۴] چارچوبی برای بهبود کارایی سامانه‌های بازیابی اطلاعات است. در این روش نتایج بازیابی‌شده اولیه، مرتبط به موضوع پرس‌وجو در نظر گرفته شده و تعدادی از اسناد رتبه بالا به‌منظور استخراج کلمات بسط انتخاب می‌شوند. درحالی‌که ممکن است موتور جستجو اسناد نامرتبط را به‌عنوان اسناد مرتبط تشخیص دهد. فرض مرتبط بودن نتایج اولیه در روش بازخورد شبه‌مرتبط ممکن است اسناد نامرتبط را در فرآیند بسط دخیل کند و کلمات نامرتبط به پرس‌وجو اضافه شود. برای جلوگیری از این مشکل روش‌هایی به‌منظور انتخابِ بهتر اسناد بازخورد برای بسط انجام شده است [۷-۵]. به‌طورکلی انتخاب بهترین اسناد، هدف کلیه روش‌های بسط پرس‌وجوی مبتنی بر روش بازخورد شبه‌مرتبط است تا کلمات بسط مرتبط‌تری انتخاب شود.

برای به‌دست آوردن اسناد بازخورد مناسب، طبقه‌بندی اسناد بازخورد در [۵] ارائه شده است. در این روش، از یادگیری همکار برای به‌دست آوردن داده‌های آموزشی و طبقه‌بندی اسناد نتایج اولیه استفاده شده است. اسناد با برچسب مثبت به‌عنوان منبع برای بسط پرس‌وجو استفاده می‌شوند.

خوشه‌بندی اسناد حاصل از نتایج اولیه بر اساس شباهت مبتنی بر کلمه در [۶] انجام گرفته است که در آن از اسناد موجود در هم‌پوشانی خوشه‌های رتبه‌بالا به‌عنوان منبع برای بسط پرس‌وجو استفاده شده است. برای انتخاب کلمات بسط از مدل ربط لاورنکو[۲] [۸] استفاده شده است. نتایج بازیابی این روش، بهبود را در معیار متوسط میانگین دقت[۳] (MAP) نشان می‌دهد. همچنین، باز رتبه‌بندی اسناد بازیابی‌شده با خوشه‌بندی اسناد بازیابی‌شده با استفاده از مدل فضای برداری برای بازیابی، نتایج موفقیت‌آمیزی را داشته است [۹، ۱۰].

هدف خوشه‌بندی قرار دادن شبیه‌ترین اسناد کنار هم است. شباهت کسینوسی با ویژگی‌های مبتنی بر کلمه اسناد، معیار مناسبی برای شباهت بین اسناد هست و برای بازیابی اسناد و خوشه‌بندی اسناد بازخورد استفاده شده است [۶، ۱۱]. شباهت کسینوسی، شباهت ضریب جاکارد[۴]، شباهت ضریب دایس[۵]، [۱۲] KL-دیورژانس[۶]، [۱۳] برای محاسبه شباهت بین دو سند و یا شباهت بین سند و پرس‌وجو به‌کار می‌روند.

در این مقاله، از خوشه‌بندی اسناد شبه‌بازخورد (CPRF) برای به‌دست آوردن منبع برای انتخاب کلمات بسط استفاده می‌شود که شباهت‌هایی به راهکار ارائه‌شده در مرجع [۶] دارد. در روش ارائه‌شده در این مقاله، برای خوشه‌بندی اسناد، از الگوریتم نزدیک‌ترین K همسایه[۷] (K-NN) استفاده شده است. قطعیت الگوریتم K-NN از تولید نتایج تصادفی جلوگیری می‌کند زیرا در هیچ‌یک از مراحل الگوریتم عملی تصادفی اتفاق نمی‌افتد. اسناد بازخورد برای بسط از اسنادِ رتبه‌بالایِ خوشه‌هایِ رتبه‌بالا انتخاب می‌شوند در حالی‌که اسناد بازخورد برای بسط در [۶] از هم‌پوشانی خوشه‌ها با افزونگی انتخاب می‌شوند. کلمات بسط با تابع رتبه‌بندی TF-IDF[۸] رتبه‌بندی می‌شوند. نتایج آزمایش‌های انجام‌گرفته روی مجموعه‌داده پزشکی MED [۱۴] نشان می‌دهد که خوشه‌بندی اسناد بازخورد به‌منظور انتخاب اسناد بهتر، معیار متوسط میانگین دقت (MAP) را بهبود می‌دهد.

در ادامه این مقاله کارهای پیشین آورده شده است. در قسمت ۳ روش ارائه‌شده در این مقاله برای بسط بیان شده و در قسمت ۴ مراحل انجام آزمایش‌ها و چگونگی انجام آن بیان شده و نتایج آزمایش‌های انجام‌گرفته روی مجموعه‌داده MED آورده شده است و در بخش ۵ نتیجه‌گیری مقاله بیان شده است.

۲- کارهای پیشین

بازخورد مرتبط[۹] (RF) [۱۵] روشی مؤثر برای بهبود دقت بازیابی با بازساختاری پرس‌وجوی اولیه با استفاده از اسناد بازخورد است. در بازخورد مرتبط، کاربر صریحاً با سامانه بازیابی تعامل می‌کند و اسناد را با عنوان مرتبط و یا نامرتبط برچسب می‌زند و بازیابی بعدی با اطلاعات بازخورد انجام می‌گیرد. بازخورد شبه‌مرتبط (PRF) [۴]، شکل خودکار بازخورد مرتبط است و فرض می‌کند که تعدادی از اسناد رتبه‌بالا در بازیابی اولیه مرتبط به موضوع پرس‌وجو هستند و از این اسناد رتبه‌بالا به‌عنوان منبعی برای انتخاب کلمات بسط استفاده می‌کند. تعدادی از کلمات بسط به پرس‌وجو اضافه شده و بازیابی نهایی با پرس‌وجوی بسط داده‌شده انجام می‌گیرد. آزمایش‌های این روش روی مجموعه‌داده‌های مختلف بهبود نتایج را نشان می‌دهد.

برای انتخاب بهترین اسناد، خوشه‌بندی اسناد نتایج اولیه می‌تواند راه‌حل مناسبی باشد. اسناد رتبه‌بالا به‌صورت بردار با وزن‌دهی TF-IDF نشان داده می‌شوند. اسناد با الگوریتم K-NN بر اساس ویژگی‌های مبتنی بر کلمه، با شباهت کسینوسی خوشه‌بندی شده‌اند و اسناد موجود در هم‌پوشانی خوشه‌ها به‌عنوان اسناد چیره انتخاب شده و با برگزینی چندباره این اسناد، مدل ربط کلمات با روش لاورنکو ساخته شده و مرتبط‌ترین کلمات برای بسط پرس‌وجو انتخاب می‌شوند [۶]. آزمایش‌های این روش روی مجموعه داده‌های استاندارد TREC، مثل AP ،WSJ و مجموعه‌داده‌های حجیم مثل WT10g بهتر شدن نتایج جستجو را نسبت به مدل زبان[۱۰] نشان می‌دهد.

همچنین خوشه‌بندی اسناد بازخورد برای بسط پرس‌وجو در حوزه جستجوی اسناد با اندازه کوچک که تعداد کلمه‌های متمایز آن‌ها کم هست، استفاده شده است [۷]. خوشه‌های ایجادشده بررسی می‌شوند و بعضی از خوشه‌ها ممکن هست ادغام و یا حذف شوند. پس از تحلیل خوشه‌ها، اسناد مرتبط با ویژگی چیرگی رتبه‌بندی شده و برای بسط انتخاب می‌شوند. کلمات با مدل ربط لاورنکو رتبه‌بندی شده و رتبه‌بالاترین کلمه‌ها به پرس‌وجو اضافه می‌شود. در آزمایش‌های این روش روی مجموعه داده‌های patent مثل TREC-CRT، ChemAppPat، DentPat مؤثر بودن این روش و افزایش پیداشوندگی اسناد را نشان می‌دهد.

برای انتخاب مجموعه‌ای از اسناد بازخورد متنوع، اعضای خوشه‌ها با معیار دارا بودن زیرمجموعه‌ای یکسان از کلمه‌های پرس‌وجو، انتخاب می‌شوند [۱۶]. دلیل این خوشه‌بندی وجود اسناد خیلی شبیه و زائد در اسناد رتبه‌بالا هست و سعی در نادیده گرفتن این اسناد را دارد. این نوع برگزینی اسناد برای خوشه‌ها، منجر به نتایج چندان مطلوبی روی مجموعه‌داده‌های NTCIR نشد.

از خوشه‌بندی با تمایل[۱۱] برای خوشه‌بندی اسناد نتایج اولیه در [۱۷] استفاده شده است. تعدادی از اسناد که بیشترین تفاوت را نسبت به یکدیگر داشته باشند به‌عنوان مرکز خوشه‌ها انتخاب می‌شوند. برای مرکزهای انتخاب‌شده خوشه‌ها ساخته می‌شوند و بسط پرس‌وجو با کلمات نماینده هر خوشه انجام می‌گیرد. همچنین برای پیشنهاد پرس‌وجو به کاربر، از خوشه‌بندی اسناد نتایج اولیه در [۱۸] استفاده شده است. نتایج اولیه خوشه‌بندی می‌شوند. برای خوشه‌های ساخته‌شده کلمات بسط استخراج شده و پرس‌وجوهای جدید برای هر خوشه به کاربر پیشنهاد می‌شوند.

در روش دیگری برای مسئله انتخاب بهترین اسناد برای بسط، نتایج اولیه با ویژگی‌های مبتنی بر کلمه طبقه‌بندی می‌شوند. داده‌های آموزشی طبقه‌بندها، تعدادی از اسناد رتبه بالا به‌عنوان طبقه مثبت و تعدادی از اسناد رتبه پایین به‌عنوان طبقه منفی در نظر گرفته شده‌اند [۵]. طبقه‌بند ساخته‌شده، سایر اسناد رتبه‌بالای اولیه را طبقه‌بندی می‌کند. اسناد برچسب‌گذاری‌شده با برچسب مثبت به‌عنوان اسناد بازخورد انتخاب می‌شوند و کلمه‌های بسط از این اسناد استخراج می‌شوند. آزمایش‌های این روش روی مجموعه‌داده‌های استاندارد TREC نتایج بهتری را نشان می‌دهد.

روش‌های بازیابی اطلاعات بسیاری فرض خوشه‌بندی را برای بهبود اثربخشی بازیابی پذیرفته‌اند [۶]. فرض خوشه‌بندی بیان می‌کند که اسناد مرتبط و شبیه به هم متقابلاً به پرس‌وجو نیز مرتبط هستند [۱۹]. خوشه‌بندی‌های انجام‌شده در کارهای مذکور، روی نتایج اولیه، بر اساس شباهت بین دو سند و یا دارا بودن زیرمجموعه‌ای یکسان از کلمه‌های پرس‌وجو انجام گرفته است.

برای انتخاب کلمات بسط، کلمات اسناد انتخاب‌شده رتبه‌بندی می‌شوند تا رتبه بالاترین کلمات به پرس‌وجو اضافه شود. برای انتخاب کلمات بسط، مدل ربط لاورنکو بر اساس مدل زبان نتایج مطلوبی را نشان داده است [۶، ۷]. تابع رتبه‌بندی KL-دیورژانس در [۵] استفاده شده و نتایج بهتری در بازیابی به‌دست آمده است. تابع رتبه‌بندی TF-IDF، Okapi BM24 و مدل زبان لاورنکو برای رتبه‌دهی کلمات بسط به‌کار برده شده‌اند [۷].

۳- خوشه‌بندی اسناد شبه‌بازخورد با شباهت کسینوسی

در این بخش، ابتدا شباهت کسینوسی بین اسناد بیان شده و سپس الگوریتم بسط پرس‌وجو آورده شده است. اسناد به‌صورت بردار با ویژگی‌های مبتنی بر کلمه نشان داده می‌شوند. کلمه‌های اسناد با تابع وزن‌دهی TF-IDF وزن‌دهی شده و به‌صورت بردار برای هر سند نشان داده می‌شود.

۳-۱- اندازه‌گیری شباهت کسینوسی بین دو سند

اسناد حاصل از نتایج اولیه به‌صورت برداری از وزنTF-IDF کلمات نشان داده می‌شوند. برای محاسبه شباهت کسینوسی بین دو سند، ضرب داخلی بردارشان محاسبه می‌شود و به‌صورت فرمول (۱) بیان می‌شود.

$$\cos\_sim(d_1,d_2)=\sum\frac{W(w,d_1)}{\|d_1\|}\frac{W(w,d_2)}{\|d_2\|} \quad (۱)$$

که در آن d_1 و d_2 دو سندی هستند که شباهتشان اندازه‌گیری می‌شود. $W(w,d_1)$ وزن کلمه w در سند d_1 و $W(w,q)$ وزن w در سند d_2 است که با تابع وزن‌دهی TF-IDF در فرمول (۲) محاسبه شده است.

$$TF-IDF(w)_d=tf(w)_d\ .\ idf(w) \quad (۲)$$

$$idf(w)=\log_{10}(\frac{N}{df(w)}) \quad (۳)$$

که در آن $f(w)_d$ ، فراوانی کلمه w در سند d را نشان می‌دهد. N تعداد اسناد مجموعه‌داده و $df(w)$ تعداد اسنادی از مجموعه‌داده که کلمه w را دارا می‌باشند. $\|d_1\|$ اندازه سند ۱ و $\|d_1\|$ اندازه سند ۲ هست که به‌صورت رابطه (۴) به‌دست می‌آید.

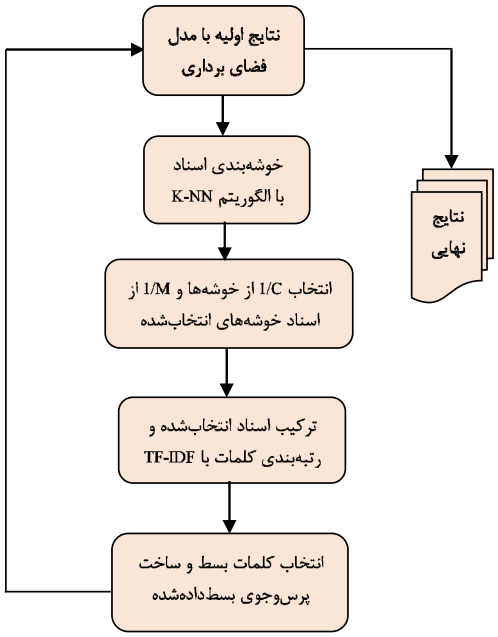
$$\|d\|=\sqrt{\sum_w W(w,d)^2} \quad (۴)$$

۳-۲- خوشه‌بندی اسناد شبه‌بازخورد با شباهت کسینوسی (CPRF)

در این مقاله، بازیابی پایه با مدل فضای برداری انجام می‌شود. پرس‌وجوها و اسناد با تابع وزن‌دهی TF-IDF وزن‌دهی شده و به‌صورت بردار نشان داده می‌شوند. میزان شباهت اسناد به پرس‌وجو با شباهت کسینوسی نرمال‌سازی‌شده محاسبه شده و رتبه‌بالاترین اسناد به‌عنوان نتایج مرتبط برگردانده می‌شود. مراحل انتخاب اسناد بازخورد و بسط پرس‌وجو در این مقاله در پنج مرحله بیان شده است و به‌صورت کلی در شکل ۱ نشان داده شده است.

- بازیابی اسناد با مدل فضای برداری: $|R|$ سند برای خوشه‌بندی در نظر گرفته می‌شود.
- خوشه‌بندی اسناد رتبه‌بالا: اسناد به‌صورت بردار بر اساس تابع وزن‌دهی TF-IDF نشان داده می‌شوند. شباهت کسینوسی برای محاسبه شباهت بین اسناد استفاده شده است. شباهت کسینوسی بین ۲ سند با فرمول (۱) محاسبه می‌شود.
- رتبه‌بندی خوشه‌ها: خوشه‌ها بر اساس شباهت اعضایشان نسبت به مرکز خوشه رتبه‌بندی می‌شوند. $1/C$ از خوشه‌ها انتخاب شده و $1/M$ از اسناد خوشه‌ها برای بسط انتخاب می‌شوند.
- ترکیب اسناد انتخاب‌شده: اسناد انتخاب‌شده با هم ترکیب شده و کلمات حاصل از آن با تابع وزن‌دهی TF-IDF که در فرمول (۲) آورده شده، رتبه‌بندی شده و برای بسط استخراج می‌شوند. فراوانی کلمه‌ها در سند ترکیب‌شده برابر با تعداد اسنادی از نتایج اولیه که شامل آن کلمه هستند محاسبه می‌شود.

ساخت پرس‌وجوی بسط: کلمات بسط انتخاب‌شده به پرس‌وجو اضافه می‌شوند.



شکل ۱: فلوچارت مراحل راهکار ارائه‌شده در مقاله

۳-۲-۱- بازیابی اسناد با مدل فضای برداری

اسناد و پرس‌وجو به‌صورت بردار نشان داده می‌شوند. درایه‌های بردار، وزن TF-IDF کلمات موجود در اسناد و پرس‌وجو است. شباهت اسناد با پرس‌وجو با شباهت کسینوسی محاسبه می‌شود. برای سند d و پرس‌وجوی q، شباهت آن‌ها با شباهت کسینوسی نرمال‌سازی‌شده به‌صورت فرمول (۵) محاسبه می‌شود.

$$sim(d,q) = \sum \frac{W(w,d)}{\| d \|} \frac{W(w,q)}{\| q \|} \quad (۵)$$

$W(w,q)$ و $W(w,d)$ به‌ترتیب وزن TF-IDF کلمه w در پرس‌وجوی q و سند d است. شباهت تمام اسناد با فرمول (۵) نسبت به پرس‌وجو محاسبه می‌شود و با مرتب کردن نزولی شباهت آن‌ها تعدادی از اسناد به کاربر به‌عنوان بازیابی اولیه ارائه می‌شود.

۳-۲-۲- خوشه‌بندی اسناد رتبه‌بالا

$|R|$ تعداد از اسناد رتبه‌بالا به‌عنوان اسناد شبه‌مرتبط انتخاب می‌شود. اسناد به‌صورت بردار با وزن‌دهی TF-IDF کلمات نشان داده می‌شوند. از الگوریتم خوشه‌بندی K-NN برای قرار دادن اسناد مشابه در کنار هم استفاده می‌شود. استفاده از الگوریتم K-NN نتایج قطعی تولید می‌کند و برخلاف الگوریتم‌های خوشه‌بندی که از مرکزهای تصادفی برای خوشه‌بندی استفاده می‌کنند، نتایج یکسانی را در اجراهای مختلف تولید می‌کند. معیار خوشه‌بندی شباهت کسینوسی هست که در فرمول (۱) آورده شده است.

۳-۲-۳- رتبه‌بندی خوشه‌ها

خوشه‌ها با محاسبه مجموع شباهت کسینوسی اعضای آن‌ها نسبت به پرس‌وجو امتیازدهی می‌شوند و خوشه‌ها به‌صورت نزولی رتبه‌بندی می‌شوند. $1/C$ از خوشه‌ها به‌عنوان خوشه‌های بازخورد برای بسط انتخاب می‌شوند و $1/M$ از اسناد خوشه‌های انتخاب‌شده برای بسط بدون افزونگی برگزیده می‌شوند (در الگوریتم خوشه‌بندی K-NN هم‌پوشانی اتفاق می‌افتد و در انتخاب اسناد ممکن است افزونگی اتفاق بیافتد. در این مقاله، اسناد بدون افزونگی انتخاب می‌شوند در حالی‌که در [۶]، اسناد چیره با افزونگی انتخاب می‌شوند).

۳-۲-۴- ترکیب اسناد انتخاب‌شده

اسناد انتخاب‌شده با هم ترکیب می‌شوند به‌طوری‌که فراوانی کلمه w در آن برابر با تعدادی از اسناد انتخاب‌شده از خوشه‌ها است که کلمه w را دارا هستند. برای انتخاب کلمات بسط از تابع وزن‌دهی TF-IDF استفاده کرده‌ایم که در بازیابی اسناد کلمات با این تابع رتبه‌بندی شده‌اند و رتبه‌بالاترین کلمه‌ها برای بسط انتخاب می‌شوند (از مدل ربط لاورنکو برای انتخاب کلمات بسط استفاده شده است [۶] که بازیابی با مدل زبان انجام می‌گیرد).

۳-۲-۵- ساخت پرس‌وجوی بسط داده‌شده

همانند فرمول پرس‌وجوی بسط در [۶]، پرس‌وجوی بسط داده‌شده به‌صورت ترکیبی از پرس‌وجوی اصلی و کلمات بسط به‌صورت فرمول (۶) ساخته می‌شود.

$$\lambda(q) + (1-\lambda)(t_1,\ t_2,\ ...,\ t_e) \qquad (۶)$$

q پرس‌وجوی اولیه است که به‌صورت مجموعه‌ای از کلمات مستقل از هم نشان داده می‌شود. λ وزن پرس‌وجوی اولیه را در پرس‌وجوی بسط داده‌شده نشان می‌دهد. $t_1, t_2, ..., t_e$ کلمات انتخاب‌شده برای بسط هستند که با مقدار $(1-\lambda)$ نسبت به پرس‌وجوی بسط‌داده‌شده وزن‌دهی می‌شوند. پس از تشکیل این پرس‌وجو، بازیابی نهایی با این پرس‌وجو اجرا می‌شود.

۴- آزمایش‌ها

۴-۱- پیکربندی آزمایش‌ها

۴-۱-۱- مجموعه‌داده آزمایش

راهکار ارائه‌شده روی مجموعه‌داده پزشکی MED [۱۴] اعمال می‌شود. اسناد این مجموعه‌داده، چکیده‌ای از مقالات پزشکی به‌صورت فشرده است که خلاصه‌ای از جزئیات این مجموعه‌داده در جدول ۱ آورده شده است. اسناد مرتبط به پرس‌وجوها نیز در مجموعه‌داده مشخص شده است. همچنین نمایه‌زنی اسناد و پرس‌وجوهای مجموعه‌داده با موتور جستجوی متن‌باز indri [۲۰] انجام شده است. پس از نمایه‌زنی، اسناد و پرس‌وجوها به‌صورت مجموعه‌ای از کلمات مستقل از هم نشان داده می‌شوند.

جدول ۱: خلاصه‌ای از جزئیات مجموعه‌داده

تعداد اسناد	۱۰۳۳
متوسط اندازه اسناد	۱۶۲
تعداد نشانه‌ها	۹۷۶۰
تعداد پرس‌وجوها	۳۰
متوسط تعداد اسناد مرتبط به هر پرس‌وجو	۲۳

۴-۱-۲- یادگیری و ارزیابی پارامترها

کل مجموعه‌داده به دو قسمت آموزشی و آزمایش تقسیم شده است. یادگیری پارامترها با ۱۰ پرس‌وجوی اول به‌عنوان داده آموزشی انجام می‌گیرد. تعداد اسناد شبه‌بازخورد انتخاب‌شده برای بسط می‌تواند مقادیر متفاوتی داشته باشد که از مجموعه مقادیر ($|R| \in \{10, 15, 20, 25, 30, 50\}$) انتخاب می‌شود. تعداد کلمات بسط انتخاب‌شده برای اضافه شدن به پرس‌وجو از مجموعه مقادیر ($e \in \{2, 5, 7, 10, 15\}$) انتخاب می‌شود. کلمات بسط در پرس‌وجوی بسط داده‌شده با وزنی نسبت به پرس‌وجوی اولیه ترکیب می‌شود که وزن پرس‌وجوی اولیه از مجموعه مقادیر ($\lambda \in \{0.2, 0.4, 0.6, 0.8\}$) انتخاب می‌شود. با توجه به این‌که کلمات بسط از اسناد انتخاب‌شده از خوشه‌ها انتخاب می‌شوند لذا نسبت تعداد خوشه‌های انتخاب‌شده به تعداد همه خوشه‌ها در قسمت ۳ مقاله از مجموعه مقادیر ($1/C \in \{1/2, 1/3, 1/4\}$) انتخاب می‌شود. تعداد اسناد انتخاب‌شده از خوشه‌های انتخاب‌شده نیز با نسبت‌های ($1/M \in \{1/2, 1/3, 1/4\}$) مورد ارزیابی قرار گرفته است. مقادیر پارامترهای یادگیری‌شده در جدول ۲ آورده شده است. سعی شده است تا مقادیر مختلف برای پارامترها طوری انتخاب شود تا رفتار و کارایی راهکار ارائه‌شده برای مقادیر مختلف نشان داده شود. بعد از یادگیری پارامترها، از پرس‌وجوهای ۱۱ تا ۳۰ برای انجام آزمایش‌ها و مقایسه روش ارائه‌شده با روش‌های مبنا استفاده شده است.

۴-۲- معیار مقایسات

برای مقایسه روش‌های بیان‌شده در مقاله، کارایی را با معیار متوسط میانگین دقت (MAP) اندازه می‌گیریم که به‌صورت فرمول (۷) است.

$$MAP = \frac{1}{|Q|}\sum_{q\in Q} ap(q) \qquad (۷)$$

که در آن Q مجموعه پرس‌وجوهای مورد آزمایش است. $|Q|$ تعداد پرس‌وجوهای مورد آزمایش است. $ap(q)$ متوسط دقت برای پرس‌وجوی q هست که در فرمول (۸) آورده شده است.

$$ap(q) = \frac{\sum_{k}^{|R|} p@k}{|R|} \qquad (۸)$$

$p@k$ دقت در k سند بازیابی شده است و $|R|$ تعداد اسناد بازیابی شده برای پرس‌وجوی q است.

۴-۲-۱- روش‌های مورد مقایسه مبنا

روش بسط پرس‌وجوی ارائه‌شده در این مقاله با روش‌های زیر به‌عنوان مبنا مقایسه شده است:

- مدل فضای برداری (VSM): مدل فضای برداری به‌عنوان روش بازیابی پایه استفاده شده است.
- روش بازخورد شبه‌مرتبط (PRF): در این روش $|R|$ تعداد از رتبه‌بالاترین اسناد برای بسط انتخاب می‌شوند. اسناد با هم ترکیب می‌شوند و کلمات سند ترکیبی با تابع رتبه‌بندی TF-IDF رتبه‌بندی شده و رتبه بالاترین کلمات برای بسط انتخاب می‌شوند.
- روش بازخورد ارتباطی کامل (TrueRF[۱۲]): در این روش اسناد مرتبط از بین $|R|$ سند بازیابی‌شده اولیه، برای بسط انتخاب می‌شوند. اسناد با هم ترکیب شده و کلمات سند ترکیب‌شده با تابع رتبه‌بندی TF-IDF رتبه‌بندی می‌شوند و رتبه بالاترین کلمات برای بسط انتخاب می‌شوند. این روش به‌عنوان حد بالا در ارزیابی در نظر گرفته می‌شود چرا که تمام اسناد مرتبط در $|R|$ سند بازیابی شده برای بسط انتخاب می‌شوند.
- روش ارائه‌شده در مقاله، خوشه‌بندی اسناد شبه‌بازخورد با شباهت کسینوسی (CPRF).

جدول ۲: یادگیری پارامترها

روش بسط پرس‌وجو	پارامتر وزن پرس‌وجوی اولیه، λ	اسناد بازخورد، $\|R\|$	تعداد کلمات بسط، e	نسبت تعداد خوشه‌های بازخورد، $1/C$	نسبت اسناد انتخاب‌شده، $1/M$
PRF	۰/۸	۲۰	۱۰	–	–
CPRF	۰/۶	۳۰	۱۰	۰/۵	۰/۵

۴-۲-۲- نتایج آزمایش‌ها

نتایج ارزیابی روش‌های مورد مقایسه، روی مجموعه داده MED در جدول ۳ آورده شده است. هر کدام از روش‌ها با بهترین مقادیر به‌دست‌آمده برای پارامترهای خودشان، روی پرس‌وجوهای شماره ۱۱ تا ۳۰ مورد آزمایش قرار گرفته‌اند.

با توجه به جدول ۳ روش بسط پرس‌وجو با خوشه‌بندی اسناد شبه‌بازخورد (CPRF) نسبت به روش ارتباط شبه‌بازخورد (PRF)، کارایی (متوسط میانگین دقت) را بهبود می‌دهد. روش خوشه‌بندی اسناد بازخورد، ۹/۴۸ درصد نسبت به ارتباط شبه‌بازخورد و ۱۴/۹۹ درصد نسبت به مدل فضای برداری کارایی را افزایش می‌دهد. اختلاف ۱۷/۰۷ درصدی بین روش CPRF و TrueRF نشان می‌دهد که می‌توان اسناد بهتری را از بین اسناد شبه‌بازخورد برای بسط انتخاب کرد و از حضور اسناد نامرتبط برای بسط خودداری کرد.

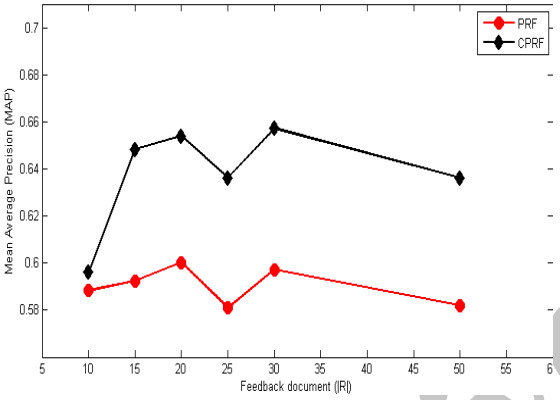
جدول ۳: کارایی (متوسط میانگین دقت) روش‌های VSM، PRF، CPRF و TrueRF

روش	VSM	PRF	CPRF	TrueRF
کارایی (MAP)	۰/۵۷۱۴	۰/۶۰۰۲	۰/۶۵۷۱	۰/۷۶۹۳

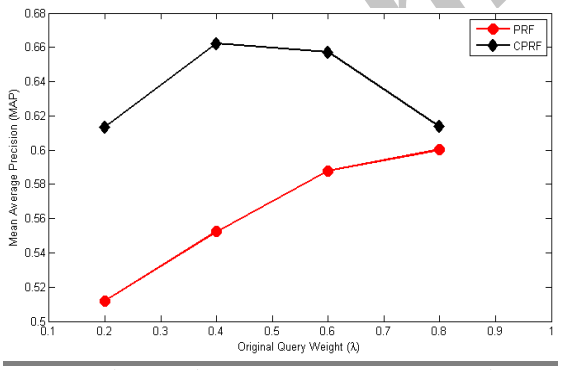
رفتار روش‌های PRF و CPRF بر تعداد مختلف اسناد شبه‌بازخورد، در شکل ۲ آورده شده است. با توجه به این‌که متوسط تعداد اسناد مرتبط به هر پرس‌وجو در مجموعه‌داده، ۲۳ سند مرتبط هست، در تعداد اسناد بازخورد بیشتر از ۲۰، کارایی (متوسط میانگین دقت) کاهش می‌یابد. کاهش کارایی به‌دلیل وجود اسناد خطا بین اسناد انتخاب‌شده برای بسط است که کلمات خطا را به پرس‌وجو اضافه می‌کند. سایر پارامترهای این دو روش در مقادیر یادگیری‌شده قرار داده شده‌اند. همان‌طور که در شکل ۲ مشخص است هر دو روش PRF و CPRF در تعداد ۲۰ سند شبه‌بازخورد بیشترین کارایی را به‌دست آورده‌اند.

شکل ۳ تأثیر وزن پرس‌وجوی اصلی نسبت به کلمات بسط را نشان می‌دهد. با قرار دادن سایر پارامترها در مقادیر یادگیری‌شده، کارایی برای مقادیر مختلف λ به‌صورت شکل ۳ به‌دست می‌آید. با توجه به شکل ۲ اختصاص وزن ناچیز به پرس‌وجوی اصلی و یا خیلی زیاد به پرس‌وجوی اصلی و کلمات بسط، کارایی را کاهش می‌دهد. وزن ۰/۸ برای PRF و ۰/۴ در روش CPRF بیشترین کارایی را به‌دست می‌آورد.

با اختصاص وزن ناچیز به کلمات بسط، اهمیت کلمه‌های بسط کاهش پیدا می‌کند و در بازیابی نقش کمتری را خواهد داشت و به بازیابی بر اساس پرس‌وجوی اولیه تمایل خواهد داشت. درحالی‌که اختصاص وزن بیشتر به کلمات بسط، بازیابی را به بازیابی با کلمات بسط سوق می‌دهد و منجر به تشکیل پرس‌وجویی جدید و نادیده گرفتن پرس‌وجوی اولیه که منظور اصلی کاربر را نشان می‌دهد، می‌شود. با توجه به این‌که در روش PRF، برای انتخاب اسناد بازخورد فقط رتبه بالا بودن اسناد مدنظر هست احتمال استفاده از اسناد خطا برای بسط افزایش می‌یابد و کلمات بسط نامناسب استخراج می‌شوند. کارایی در روش PRF با اختصاص وزن بیشتر به پرس‌وجوی اصلی افزایش می‌یابد تا اثر کلمات بسط نامناسب را کاهش دهد.



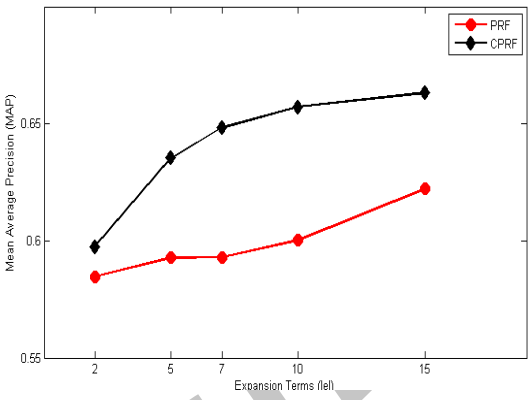
شکل ۲: کارایی (MAP) نسبت به تعداد اسناد بازخورد مختلف ($|R|$) برای PRF و CPRF



شکل ۳: کارایی (MAP) نسبت به وزن پرس‌وجوی اولیه (λ) برای CPRF و PRF

به‌طورکلی افزودن کلمات مرتبط به پرس‌وجو، کارایی را افزایش می‌دهد. تغییرات کارایی نسبت به تعداد کلمات بسط اضافه‌شده به پرس‌وجو در شکل ۴ نشان داده شده است. با قرار دادن مقادیر یادگیری‌شده برای سایر پارامترها، استفاده از کلمات بسط بیشتر، کارایی را برای هر دو روش PRF و CPRF افزایش می‌دهد. افزودن ۱۵ کلمه بسط بیشترین کارایی را به‌دست آورده است. با توجه به شکل ۴

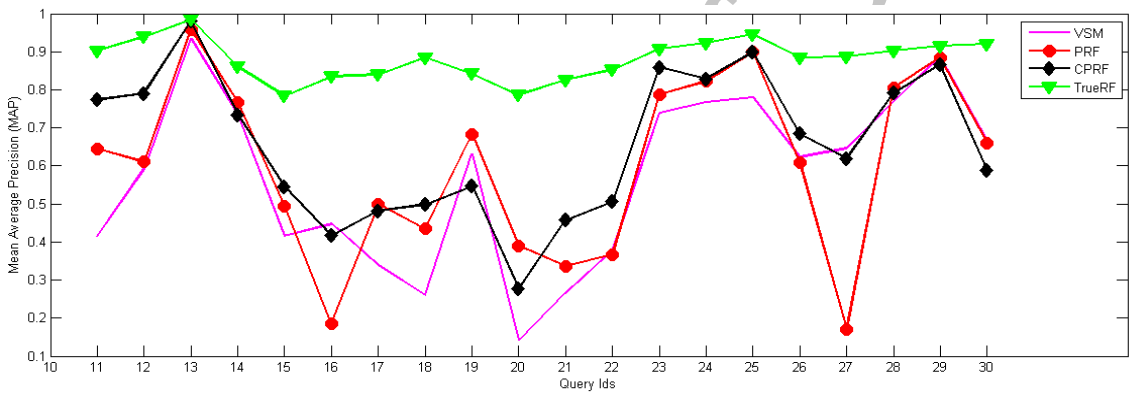
افزودن کلمات بسط تا ۵ کلمه کارایی را به طرز سریع‌تری افزایش می‌دهد و این نشان می‌دهد که پرس‌وجوها کامل نیستند و افزودن تعداد حداقلی از کلمات بسط، می‌تواند منظور دقیق‌تری از نیاز اطلاعاتی کاربر را نشان دهند.



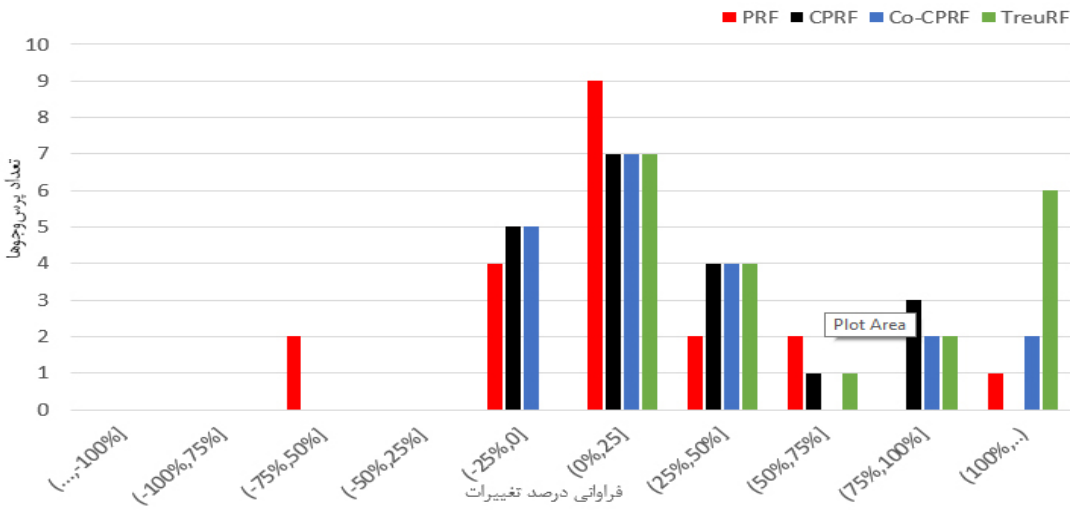
شکل ۴: کارایی (MAP) نسبت به تعداد کلمات بسط (e) برای روش CPRF و PRF

کارایی (میانگین دقت[۱۳]) روی تک‌به‌تک پرس‌وجوها ۱۱ تا ۳۰ از مجموعه‌داده در شکل ۵ آورده شده است. همه روش‌ها در مقادیر یادگیری‌شده پارامترها آزمایش شده‌اند. روش TrueRF از اسناد مرتبط در ۳۰ سند رتبه بالا برای بسط استفاده می‌کند. روش CPRF هم از ۳۰ سند رتبه بالا برای خوشه‌بندی استفاده می‌کند و سعی در استفاده از اسناد مرتبط برای بسط را دارد. تغییرات کارایی در پرس‌وجوی شماره ۱۳ و ۲۳ نشان می‌دهد که اسناد مرتبط مشخص‌شده برای این پرس‌وجوها کیفیت لازم برای بسط پرس‌وجو را ندارند. در سایر پرس‌وجوها می‌توان اسناد بهتری را انتخاب کرد و کارایی تا ۲ برابر افزایش داد.

شکل ۶ فراوانی درصد تغییر کارایی روش‌ها را نسبت به مدل فضای برداری روی تک‌تک پرس‌وجوها نشان می‌دهد. همان‌طور که مشخص است روش CPRF نسبت به PRF تعداد بیشتری از پرس‌وجوها را بهبود می‌دهد. همچنین PRF در ۲ مورد ۵۰ تا ۷۵ درصد کارایی را کاهش داده است درحالی‌که CPRF برای هیچ پرس‌وجویی تا این مقدار کارایی را کاهش نداده است.



شکل ۵: کارایی (MAP) برای پرس‌وجوهای شماره ۱۱ تا ۳۰ برای VSM، PRF، CPRF و TrueRF



شکل ۶: مقاومت‌پذیری CPRF نسبت به TrueRF و PRF

جدول ۴: تأثیر خوشه‌بندی اسناد شبه‌بازخورد (CPRF) در انتخاب اسناد مرتبط برای بسط، تعداد اسناد شبه‌بازخورد ۳۰ (۳۰ = $|R|$)

	شماره پرس‌وجوها																			
	۳۰	۲۹	۲۸	۲۷	۲۶	۲۵	۲۴	۲۳	۲۲	۲۱	۲۰	۱۹	۱۸	۱۷	۱۶	۱۵	۱۴	۱۳	۱۲	۱۱
تعداد اسناد مرتبط به پرس‌وجو	۱۴	۳۷	۳۹	۱۸	۲۸	۲۴	۲۲	۳۹	۲۵	۲۷	۳۹	۲۷	۱۵	۲۱	۱۳	۲۹	۱۶	۲۱	۹	۱۸
تعداد اسناد مرتبط بازیابی‌شده	۸	۲۳	۱۹	۹	۱۲	۲۱	۱۷	۱۹	۷	۹	۳	۱۱	۱۲	۹	۹	۱۲	۱۱	۱۹	۸	۱۳
تعداد اسناد انتخاب‌شده	۲۶	۲۶	۲۳	۲۳	۲۷	۲۱	۲۶	۲۶	۲۸	۲۸	۲۲	۲۶	۲۸	۲۷	۲۱	۲۲	۲۸	۲۷	۲۶	۲۶
تعداد اسناد مرتبط انتخاب‌شده	۷	۱۹	۱۹	۹	۱۲	۱۴	۱۵	۱۶	۷	۹	۰	۹	۱۱	۸	۸	۱۰	۹	۱۷	۸	۱۱

۴-۲-۳- تأثیر خوشه‌بندی بر استفاده از اسناد مرتبط برای بسط

روش خوشه‌بندی اسناد شبه‌مرتبط که در این مقاله ارائه شده است، اسناد نتایج اولیه را خوشه‌بندی می‌کند و تعدادی از اسناد خوشه‌ها را برای بسط انتخاب می‌کند. در بین اسناد انتخاب‌شده برای بسط تعدادی سند نامرتبط هم وجود خواهد داشت. تعداد اسناد مرتبط و نامرتبط که برای بسط پرس‌وجوهای ۱۱ تا ۳۰ انتخاب می‌شوند در جدول ۴ آورده شده است. با توجه به جدول ۴ در ۶ پرس‌وجو که ستون مربوطه آن‌ها پررنگ شده است تمامی اسناد مرتبط بازیابی شده، برای بسطشان انتخاب شده است. با توجه به مقادیر تعداد اسناد انتخاب‌شده در جدول، روش خوشه‌بندی اسناد شبه‌بازخورد، تعدادی از اسناد نامرتبط را از فرآیند بسط خارج می‌سازد.

۵- نتیجه‌گیری

با توجه به نتایج آزمایش‌ها روی مجموعه‌داده MED، روش خوشه‌بندی اسناد شبه‌بازخورد (CPRF) نسبت به روش ارتباط شبه‌بازخورد (PRF)، کارایی بهتری را در بازیابی اسناد پزشکی دارد. کارایی روش CPRF نسبت به ارتباط شبه‌بازخورد (PRF) ۹/۵ درصد بیشتر است. روش ارائه‌شده (CPRF) به‌دلیل اضافه کردن کلمات بسط، کارایی را افزایش می‌دهد و برای استخراج کلمات بسط مناسب باید اسناد مرتبط و مناسبی انتخاب شوند. انتخاب اسناد مناسب برای بسط اصلی‌ترین موضوع در روش‌های بسط پرس‌وجو هست و روش‌های بسط پرس‌وجو سعی در انتخاب بهترین اسناد به‌عنوان اسناد بسط را دارند. خوشه‌بندی اسناد شبه‌بازخورد (CPRF) نشان داد که می‌توان اسناد بهتری را برای بسط انتخاب کرد تا کلمات بهتری را برای بسط استخراج کرد. اختلاف کارایی روش مقاله (CPRF) نسبت به حد بالای این روش (TrueRF)، نشان می‌دهد که می‌توان با بررسی بیشتر اسناد شبه‌بازخورد و انتخاب اسناد بهتر، کارایی را باز هم افزایش داد.

برای کارهای آتی، استفاده از روش‌های کلاس‌بندی اسناد، تحلیل محتویات خوشه‌ها، استفاده از سایر روش‌های رتبه‌بندی کلمات پیشنهاد می‌شود تا بتوان کارایی را افزایش داد. می‌توان از تحلیل زبانی و روش‌های معنایی برای استخراج زمینه و مفهوم اسناد و پرس‌وجو استفاده کرد تا خوشه‌های بهتری را به‌دست آورد.

مراجع

[1] R. Krovetz, "Homonymy and polysemy in information retrieval," *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 72-79, 1997.

[2] A. Spink, and B.J. Jansen, *A Study of Web Search Trends*, Available online at: http://www.webology.ir/2004/v1n2/-a4.html/.

[3] M. Sanderson, "Ambiguous queries: test collections need more sense," *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 499-506, 2008.

[4] J. Xu, and W.B. Croft, "Query expansion using local and global document analysis.," *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4-11, 1996.

[5] J.X. Huang, J. Miao, and B. He, "High performance query expansion using adaptive co-training," *Information Processing and Management*, vol. 49, pp. 441-453, 2013.

[6] K.S. Lee, and W.B. Croft, "A deterministic resampling method using overlapping document clusters for pseudo-relevance feedback," *Information Processing and Management*, vol. 49, pp. 792-806, 2013.

[7] S. Bashir, "Improving retrievability with improved cluster-based pseudo-relevance feedback selection," *Expert Systems with Applications*, vol. 39, pp. 7495-7502, 2012.

[8] V. Lavrenko, and W.B. Croft, "Relevance based language models," *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 120-127, 2001.

[9] K.S. Lee, Y.C. Park, and K.S. Choi, "Re-ranking model based on document clusters," *Information Processing and Management*, vol. 37, pp. 1-14, 2001.

[10] K.S. Lee, K. Kageura, and K.S. Choi, "Implicit ambiguity resolution using incremental clustering in cross-language information retrieval," *Information Processing and Management*, vol. 40, pp. 145-159, 2004.

[11] T. Anastasios, and C.J. Van-Rijsbergen, "Query-sensitive similarity measures for information retrieval," *Knowledge and Information Systems*, vol. 6, no. 5, pp. 617-642, 2004.

[12] G. Salton, A. Wong, and C.S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, pp. 613-620, 1975.
[13] D. Hiemstra, "A linguistically motivated probabilistic model of information retrieval," *Research and Advanced Technology for Digital Libraries*, pp. 569-584, 1998.
[14] U.O. Glascow, (2014/03), *Medline collection*, Available online at: http://ir.dcs.gla.ac.uk/resources/test_collections-/medl/.
[15] J.J. Rocchio, *Relevance Feedback in Information Retrieval*, 1971.
[16] T. Sakai, T. Manabe, and M. Koyama, "Flexible pseudo-relevance feedback via selective sampling," *ACM Transactions on Asian Language Information Processing*, vol. 4, pp. 111-135, 2005.
[17] A.L. Kaczmarek, "Interactive query expansion with the use of clustering-by-directions algorithm," *IEEE Transactions on Industrial Electronics*, vol. 58, pp. 3168-3173, 2011.
[18] G. Bordogna, A. Campi, G. Psaila, and S. Ronchi, "Disambiguated query suggestions and personalized content-similarity and novelty ranking of clustered results to optimize web searches," *Information Processing and Management*, vol. 48, pp. 419-437, 2012.
[19] N. Jardine, and C.J. Van-Rijsbergen, "The use of hierarchic clustering in information retrieval," *Information Storage and Retrieval*, vol. 7, pp. 217-240, 1971.
[20] T. Strohman, D. Metzler, H. Turtle, and W.B. Croft, "Indri: a language model-based search engine for complex queries," *Proceedings of the International Conference on Intelligent Analysis*, pp. 2-6, 2005.

زیرنویس‌ها

[1] Pseudo Relevance Feedback
[2] Lavrenko's Relevance Model
[3] Mean Average Precision
[4] Jacard's Coefficient
[5] Dice Coefficient
[6] KL-Divergence
[7] K-Nearest Neighbor
[8] Term Frequency-Inverse Document Frequency
[9] Relevance Feedback
[10] Language Model
[11] Clustering-by-Direction
[12] True Relevance Feedback
[13] Average Precision