بسط پرس‌وجو با خوشه‌بندی اسناد شبه بازخورد با شباهت حساس به پرس‌وجو

رضا خدائی[۱*]، محمدعلی بالافر[۲]، سیدناصر رضوی[۳]

اطلاعات مقاله

دریافت مقاله: ۱۳۹۳/۰۸/۱۴

پذیرش مقاله: ۱۳۹۴/۰۷/۰۵

واژگان کلیدی:

بازیابی اطلاعات،

بسط پرس‌وجو،

شباهت حساس به پرس‌وجو،

بازخورد شبه مرتبط.

چکیده

بسط پرس‌وجو به عنوان یکی از روش‌های انطباق پرس‌وجو، اثربخشی جستجو را در بازیابی اطلاعات افزایش می‌دهد. بازخورد شبه مرتبط (PRF) روشی برای بسط پرس‌وجو است که فرض می‌کند اسناد رتبه بالا از نتایج اولیه مرتبط به موضوع پرس‌وجو هستند و کلمات بسط را از این اسناد انتخاب می‌کند. درحالی‌که ممکن است اسناد نامرتبط به پرس‌وجو در اسناد رتبه‌بالا وجود داشته باشد. روش‌هایی برای انتخاب اسناد مرتبط و نادیده گرفتن اسناد خطا از اسناد رتبه‌بالا ارائه‌شده است که از خوشه‌بندی و یا طبقه‌بندی اسناد استفاده کرده‌اند. مهم‌ترین موضوع در بسط پرس‌وجو، انتخاب کلمات بسط از مرتبط‌ترین اسناد است. در این مقاله ما خوشه‌بندی اسناد شبه بازخورد را براساس شباهت حساس به پرس‌وجو ارائه می‌کنیم که در قرار دادن شبیه‌ترین اسناد در کنار هم مؤثر است. شباهت حساس به پرس‌وجو که نسبت به شباهت مبتنی بر کلمه نتایج بهتری را در بازیابی اسناد بدست آورده است، دلیل استفاده در این مقاله است. خوشه‌ها را مطابق با شباهت درونی‌شان رتبه‌بندی کرده و تعدادی از خوشه‌های رتبه‌بالا را برای بسط انتخاب می‌کنیم. کلمات بسط را از اسناد خوشه‌های انتخاب‌شده، براساس تابع رتبه‌بندی TF-IDF استخراج می‌کنیم. آزمایش‌های انجام‌شده روی مجموعه‌داده‌ی پزشکی MED نشان می‌دهد که نتایج جستجو برای پرس‌وجوهای بسط‌داده‌شده با اسناد انتخاب‌شده از خوشه‌ها، نسبت به روش بازخورد شبه مرتبط (PRF) و بازیابی اولیه (VSM) بهتر است و اثربخشی جستجو را افزایش می‌دهد.

۱- مقدمه

سامانه‌های بازیابی اطلاعات سعی در برطرف کردن نیازهای اطلاعاتی کاربران دارند. موتور جستجوها از معروف‌ترین سامانه‌های بازیابی اطلاعات هستند، طوری‌که روزانه توسط میلیون‌ها کاربر استفاده می‌شوند. شش میلیارد بازدید روزانه از موتور جستجوی Google نشان‌دهنده‌ی اهمیت فزاینده‌ی سامانه‌های بازیابی اطلاعات است [۱].

کاربران نیاز اطلاعاتی‌شان را در قالب مجموعه‌ای از کلمات (پرس‌وجو) به سامانه‌ی بازیابی ارسال می‌کنند و انتظار دریافت مرتبط‌ترین اطلاعات را دارند. جستجو براساس پرس‌وجوی ارسالی کاربر انجام می‌شود و این امکان وجود

*. پست الکترونیک نویسنده مسئول: r.khodaei۹۱@ms.tabrizu.ac.ir

۱. کارشناس ارشد، مهندسی کامپیوتر (نرم‌افزار)، دانشگاه تبریز

۳،۲. استادیار، دانشکده برق و کامپیوتر، دانشگاه تبریز

دارد که پرس‌وجوی ارسالی کاربر کوتاه، مبهم و بی‌معنی باشد و نتواند نیاز اطلاعاتی کاربر را توصیف کند. معمولاً مشکلات زبان طبیعی، انتخاب کلمات نامناسب برای پرس-وجو، وجود کلمات مبهم با چند معنی[۲] و کوتاهی پرس-وجوها [۳] 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 [۴]. بسط پرس‌وجو یکی از روش‌های انطباق پرس‌وجو است که سعی در برطرف کردن مشکلات مذکور را دارد. اضافه کردن کلمات مرتبط به موضوع پرس-وجو می‌تواند پرس‌وجو را از لحاظ وجود کلمات مرتبط بهبود دهد و نقص معنایی و مفهومی پرس‌وجو را برطرف کند. بسط پرس‌وجو معمولاً اثربخشی نتایج جستجو را افزایش می‌دهد.

در بازیابی اطلاعات، روش بازخورد شبه مرتبط[۳] (PRF) [۵] معمولاً نتایج جستجو را بهبود می‌دهد. در این روش فرض می‌شود که اسناد رتبه‌بالا مرتبط به موضوع پرس‌وجو هستند و تعدادی از اسناد رتبه‌بالا به‌منظور استخراج کلمات بسط انتخاب می‌شوند. حال‌آنکه ممکن است طبق فرض مرتبط بودن اسناد رتبه‌بالا، تعدادی از اسناد نامرتبط برای بسط پرس‌وجو انتخاب شوند و اثر منفی روی بازیابی داشته باشند. برای جلوگیری از این مشکل روش‌هایی به‌منظور انتخاب بهتر اسناد بازخورد برای بسط پیشنهاد شده است [۶-۸]. به‌طورکلی انتخاب بهترین اسناد، هدف کلیه‌ی روش‌های بسطِ پرس‌وجوی محلی است تا منجر به استخراج کلمات بهتری شود.

طبقه‌بندی اسناد شبه بازخورد می‌تواند اسناد بهتری را برای بسط انتخاب کند [۶]. در این روش، تعدادی از اسناد رتبه-بالا و رتبه‌پایین به‌عنوان داده‌های آموزشی به ترتیب با برچسب مثبت و منفی در نظر گرفته می‌شوند. بقیه اسناد بازخورد با ویژگی‌های مبتنی بر کلمه‌ی داده‌های آموزشی طبقه‌بندی‌شده و کلمه‌های بسط از اسناد با برچسب مثبت استخراج می‌شوند.

خوشه‌بندی اسناد رتبه‌بالا حاصل از نتایج اولیه با شباهت مبتنی بر کلمه در [۷] انجام گرفته است. خوشه‌ها براساس شباهت اعضایشان به پرس‌وجو رتبه‌بندی می‌شوند و خوشه‌های رتبه‌بالا به‌عنوان خوشه‌های بازخورد انتخاب می‌شوند. مرتبط‌ترین اسناد، از خوشه‌های بازخورد با حفظ افزونگی انتخاب می‌شوند. برای انتخاب کلمات بسط از مدل ربط لاورنکو[۴] [۹] استفاده شده است. نتایج جستجوی این روش، افزایش کارایی را با معیار متوسط میانگین دقت (MAP) نشان می‌دهد. همچنین، خوشه‌بندی اسناد بازیابی‌شده برای بازرتبه‌بندی اسناد که از مدل فضای برداری برای بازیابی اولیه استفاده شده است، نتایج موفقیت‌آمیزی را داشته است [۱۰ و ۱۱].

هدف خوشه‌بندی قرار دادن شبیه‌ترین اسناد کنار هم است. شباهت مبتنی بر کلمه معیار مناسبی برای شباهت بین اسناد است و برای بازیابی اسناد و خوشه‌بندی اسناد بازخورد استفاده شده است [۷ و ۱۲]. اسناد بازیابی‌شده مستقلاً و به‌صورت ضمنی به پرس‌وجو مرتبط هستند اما در شباهت مبتنی بر کلمه، پرس‌وجو صریحاً در نظر گرفته نمی‌شود. می‌توان برای محاسبه‌ی شباهت اسناد، صریحاً پرس‌وجو را نیز در نظر گرفت. اسناد حاصل از بازیابی اولیه، می‌توانند به زیرمجموعه‌ای از کلمات پرس‌وجو مرتبط باشند اما به تمام موضوع پرس‌وجو مرتبط نباشند. شباهت بین اسناد با در نظر گرفتن پرس‌وجو، شباهت حساس به پرس‌وجو است. در شباهت حساس به پرس‌وجو، شباهت بین اسناد با در نظر گرفتن اشتراکشان با پرس‌وجو بدست می‌آید [۱۲].

ما در این مقاله برای انتخاب بهتر اسناد بازخورد، روش خوشه‌بندی اسنادِ شبه بازخورد را مبتنی بر شباهت حساس به پرس‌وجو[۵] ارائه کرده‌ایم. اسناد حاصل از بازیابی اولیه که به‌صورت برداری از کلمه‌ها نشان داده می‌شود با شباهت

[۲] Pseudo-Relevance Feedback
[۴] Lavrenko's Relevance model
[۵] Query sensitive similarity

حساس به پرس‌وجو [۱۲] خوشه‌بندی می‌شود. خوشه‌ها با الگوریتم نزدیک‌ترین k همسایه (K-NN) ساخته می‌شوند. سپس خوشه‌ها براساس شباهت اسنادشان رتبه‌بندی می‌شوند و اسناد موجود در خوشه‌های رتبه بالا برای بسط انتخاب می‌شوند و کلمات بسط با تابع رتبه‌بندی TF-IDF[۶] رتبه‌بندی می‌شوند که رتبه‌بالاترین کلمه‌ها برای بسط انتخاب می‌شوند. نتایج آزمایش‌های ما نشان می‌دهد که کارایی سامانه افزایش پیدا می‌کند که نشان می‌دهد استفاده از شباهت حساس به پرس‌وجو در خوشه‌بندی اسناد بازخورد به استخراج کلمات بسط مرتبط کمک می‌کند.

در ادامه‌ی این مقاله، کارهای پیشین در بخش بعدی 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 مجموعه داده‌ی MED بیان شده است. همچنین، نتیجه‌گیری مقاله در بخش ۵ ذکر شده است.

۲- کارهای پیشین

روش‌های بازخورد مرتبط[۷] (RF) [۱۳] و بازخورد شبه مرتبط (PRF) [۵] به ترتیب از اسناد مرتبط و شبه مرتبط برای بازساختاری پرس‌وجو استفاده می‌کنند و نتایج بهتری را بدست می‌آورند. در بازخورد مرتبط، کاربر صریحاً با سیستم بازیابی تعامل می‌کند و بازیابی بعدی با اطلاعات بازخورد انجام می‌گیرد. بازخورد شبه مرتبط، شکل خودکار بازخورد مرتبط است و اسناد بازیابی‌شده‌ی اولیه مرتبط به موضوع پرس‌وجو در نظر گرفته می‌شود. تعدادی از اسناد رتبه بالا برای بسط پرس‌وجو انتخاب می‌شوند و تعدادی از کلمات این اسناد برای اضافه شدن به پرس‌وجو استخراج می‌شوند. آزمایش‌های این دو روش بهبود نتایج بازیابی را نشان می‌دهد.

از خوشه‌بندی اسناد رتبه‌بالا برای انتخاب اسناد بهتر و نادیده گرفتن اسناد خطا در [۷] استفاده شده است. اسناد رتبه‌بالا به‌صورت بردار با وزن‌دهی TF-IDF نشان داده می‌شوند اسناد با الگوریتم K-NN براساس ویژگی‌های مبتنی بر کلمه، با شباهت کسینوسی خوشه‌بندی می‌شوند. اسناد چیره در هم‌پوشانی خوشه‌ها با برگزینی چندباره به‌عنوان اسناد بازخورد برای بسط انتخاب می‌شوند. آزمایش‌های این روش روی مجموعه داده‌های استاندارد TREC، مثل AP، WSJ و مجموعه داده‌های حجیم مثل WT۱۰G بهتر شدن نتایج جستجو را نسبت به مدل زبان (LM)[۸] نشان می‌دهد.

برای بازیابی اسناد با اندازه‌ی کوچک، از خوشه‌بندی اسناد بازخورد برای بسط پرس‌وجو استفاده شده است [۸]. خوشه‌های ایجادشده از نتایج اولیه بررسی می‌شوند و بعضی از خوشه‌ها ممکن است ادغام و یا حذف شوند. پس از تحلیل و پردازش خوشه‌ها، اسناد چیره از بین هم‌پوشانی خوشه‌ها با افزونگی انتخاب می‌شوند. اسناد مرتبط با ویژگی چیره‌گی رتبه‌بندی شده و برای بسط انتخاب می‌شوند. آزمایش‌های این روش روی مجموعه داده‌های patent مثل TREC-CRT، ChemAppPat، DentPat مؤثر بودن این روش و افزایش پیدا شوندگی اسناد را نشان می‌دهد.

از خوشه‌بندی براساس معیار دارا بودن زیرمجموعه‌ای یکسان از کلمه‌های پرس‌وجو برای اعضای خوشه‌ها، به‌منظور انتخاب اسناد بازخورد متنوع در [۱۴] استفاده‌شده است. دلیل این خوشه‌بندی وجود اسناد خیلی شبیه و زائد است و سعی در نادیده گرفتن این اسناد را دارد. این نوع برگزینی اسناد برای خوشه‌ها، منجر به نتایج چندان مطلوبی را روی مجموعه داده‌های NTCIR نشد.

در روش دیگری برای مسئله‌ی انتخاب بهترین اسناد برای بسط، نتایج اولیه با ویژگی مبتنی بر کلمه طبقه‌بندی می‌شوند. تعدادی از اسناد رتبه‌بالا با برچسب مثبت و تعدادی از اسناد رتبه‌پایین با برچسب منفی برای داده‌های آموزشی

[۶] Term frequency-inverse document frequency
[۷] Relevance Feedback
[۸] Language Model

طبقه‌بندها، در نظر گرفته شده‌اند [۶]. طبقه‌بند ساخته-شده، سایر اسناد رتبه‌بالای اولیه را برچسب‌گذاری می‌کند. اسناد برچسب‌گذاری شده با برچسب مثبت به‌عنوان اسناد بازخورد انتخاب می‌شوند و کلمه‌های بسط از این اسناد استخراج می‌شوند. آزمایش‌های این روش روی مجموعه داده‌های TREC افزایش اثربخشی جستجو را نشان می-دهد.

روش‌های بازیابی اطلاعاتِ بسیاری فرض خوشه‌بندی را برای بهبود اثربخشی بازیابی پذیرفته‌اند [۷]. فرض خوشه-بندی بیان می‌کند که اسناد مرتبط و شبیه به هم، متقابلاً به پرس‌وجو نیز مرتبط هستند [۱۵]. خوشه‌بندی‌های انجام‌شده در کارهای مذکور روی نتایج اولیه، براساس شباهت بین دو سند و یا دارا بودن زیرمجموعه‌ای یکسان از کلمه‌های پرس‌وجو انجام گرفته است. معیار دیگری برای شباهت بین اسناد، شباهت حساس به پرس‌وجو است که شباهت بین دو سند را نسبت به پرس‌وجو اندازه می‌گیرد، در [۱۲] بررسی شده است. بازیابی با مدل فضای برداری[۹] (VSM)، با اندازه‌گیری شباهت، بدون در نظر گرفتن پرس-وجو و شباهت حساس به پرس‌وجو، نتایج بهتری را برای شباهت حساس به پرس‌وجو نشان می‌دهد [۱۲]. همچنین اعمال شباهت حساس به پرس‌وجو به روش بازیابیِ مدل زبان مبتنی بر خوشه‌بندی، نتایج بهتری را بدست آورده است [۱۶].

تابع رتبه‌بندی TF-IDF، Okapi BM۲٤ و مدل زبان لاورنکو، برای رتبه‌دهی کلمات بکار برده شده‌اند [۸]. استفاده از مدل ربط لاورنکو برای استخراج کلمات مرتبط در تحقیقاتی که مرتبط به مدل زبان بودند آزمایش شده است [۷، ۸]. تابع رتبه‌بندی KL-دیورژانس در [۶] برای انتخاب کلمات بسط استفاده شده و نتایج بهتری در بازیابی بدست آمده است.

۳- خوشه‌بندی اسناد شبه بازخورد مبتنی بر شباهت حساس به پرس‌وجو[۱۰] (QS-CPRF)

این بخش ابتدا شباهت حساس به پرس‌وجوی بین دو سند را با مثالی عملی بیان می‌کند. فرمول‌بندی شباهت حساس به پرس‌وجو و مدل فضای برداری در بخش بعدی آورده شده است. سپس الگوریتم ارائه‌شده برای بسط پرس‌وجو با استفاده از خوشه‌های ساخته‌شده با شباهت حساس به پرس‌وجو آورده شده است.

۳-۱- شباهت حساس به پرس‌وجو

برای بازیابی اسناد مرتبط به پرس‌وجو، شباهت پرس‌وجو با تک‌تک اسناد مجموعه محاسبه شده و شبیه‌ترین اسناد به‌عنوان نتایج جستجو به کاربر برگردانده می‌شوند. اسناد بازگردانده‌شده ممکن است به بخش‌هایی از پرس‌وجو شبیه باشند اما نسبت به موضوع کلی پرس‌وجو شبیه نباشند و اسناد با مفهوم‌های متفاوت بازیابی شوند. مثال زیر با الهام از مثال بیان شده در [۱۶] برای مفهوم شباهت حساس به پرس‌وجو آورده شده است.

پرس‌وجو۱: a ,b

پرس‌وجو۲: d ،b

سند۱: a, c, d

سند۲: b, d, e

سند۱ به پرس‌وجو۱ شبیه است زیرا در حرف a مشترک هستند. سند۲ به پرس‌وجو۱ شبیه است زیرا در حرف b مشترک هستند. سند۱ و سند۲ به هم شبیه هستند زیرا دارای حرف مشترک d هستند اما با در نظر گرفتن پرس-وجوی۱ هیچ شباهتی به هم ندارند. زیرا در هیچ یک از حرف‌های پرس‌وجو مشترک نیستند. می‌توان گفت سند۱ و سند۲ به‌صورت ایستا و بدون در نظر گرفتن پرس‌وجوی۱ به هم شبیه هستند اما به‌صورت پویا و در نظر گرفتن پرس-وجوی۱ به هم شبیه نیستند. از طرفی سند۱ و سند۲ نسبت

[۹] Vector space model

[۱۰] Query Sensitive similarity Clustering Pseudo-Relevance Feedback

به پرس‌وجوی ۲ به هم شبیه هستند زیرا در حرف d با پرس‌وجو مشترک هستند. با توجه به رابطه‌ی سند۱ و سند۲ نسبت به هر دو پرس‌وجو، می‌توان فهمید که تحت شرایطی این دو سند می‌توانند به هم شبیه باشند و یا هیچ شباهتی نداشته باشند. سند۱ و سند۲ نسبت به موضوع پرس‌وجوی۱ مرتبط نیستند، اما نسبت به موضوع پرس‌وجو۲ مرتبط هستند. درواقع شباهت اسناد به‌صورت پویا می‌تواند متفاوت شود.

جستجوی اسناد مرتبط به پرس‌وجوی "روزنامه آفتاب" می‌تواند اسنادی بازیابی کند که یا فقط به کلمه‌ی "روزنامه" و یا فقط به کلمه‌ی " آفتاب" شبیه باشند. درحالی‌که هدف کاربر از ارسال این پرس‌وجو بازیابی اسناد مرتبط به "روزنامه آفتاب" است. شباهت حساس به پرس‌وجو می‌تواند از بازیابی اسنادی که فقط به بخشی از پرس‌وجو شبیه باشند ممانعت کند و تعداد اسنادی که به موضوع کلی پرس‌وجو شبیه باشند را افزایش دهد.

۳-۲- اندازه‌گیری شباهت حساس به پرس‌وجو بین دو سند

برای محاسبه شباهت بین دو سند d و d'، با شباهت حساس به پرس‌وجو، شباهت کسینوسی نرمال‌سازی‌شده‌ی اشتراک دو سند با پرس‌وجو اندازه‌گیری می‌شود و به‌صورت فرمول (۱) بیان می‌شود [۱۲].

$$Qs\_sim(d,d',q)=\sum\frac{W(w,d\otimes d')}{\|d\otimes d'\|}\frac{W(w,q)}{\|q\|} \quad (۱)$$

که در آن $d\otimes d'$ سند مجازی دو سند d و d' است که از اشتراک کلمات دو سند بدست می‌آید.

فراوانی کلمات در سند مجازی، میانگین هندسی فراوانی آن کلمه در دو سند است که به‌صورت فرمول (۲) محاسبه می‌شود.

$$c(w,d\otimes d')=\sqrt{c(w,d)c(w,d')} \quad (۲)$$

$c(w,d)$ فراوانی کلمه w در سند d است. $W(w,d)$ وزن کلمه‌ی w در سند مجازی و $W(w,q)$ وزن w در پرس‌وجو است که با فرمول‌های (۳) و (۴) محاسبه می‌شود.

$$W(w,d)=\begin{cases}\log(c(w,d))+1 & c(w,d)>0\\ 0 & c(w,d)=0\end{cases} \quad (۳)$$

$$W(w,q)=c(w,q)\log(\frac{N+1}{df(w)}) \quad (۴)$$

$\|d\otimes d'\|$ اندازه‌ی سند مجازی و $\|q\|$ اندازه‌ی پرس‌وجو است و به‌صورت فرمول (۵) و (۶) محاسبه می‌شود.

$$\|d\otimes d'\|=\sqrt{\sum_w W(w,d\otimes d')^2} \quad (۵)$$

$$\|q\|=\sqrt{\sum_w W(w,q)^2} \quad (۶)$$

برای انتخاب کلمات بسط، از تابع رتبه‌بندی TF-IDF استفاده می‌کنیم که به‌صورت فرمول (۷) بیان می‌شود.

$$TF-IDF(w)_d=tf(w)_d\ .\ idf(w) \quad (۷)$$

$$idf(w)=\log_{10}(\frac{N}{df(w)}) \quad (۸)$$

که در آن $tf(w)_d$، فراوانی کلمه‌ی w در سند d را نشان می‌دهد. N تعداد اسناد مجموعه‌داده و $df(w)$ تعداد اسنادی از مجموعه‌اسناد است که کلمه‌ی w را دارا می‌باشند.

۳-۳- خوشه‌بندی اسناد شبه بازخورد مبتنی بر شباهت حساس به پرس‌وجو (QS-CPRF)

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

- بازیابی اسناد مرتبط به پرس‌وجو با بازیابی مدل فضای برداری: $|R|$ سند رتبه بالا برای خوشه‌بندی در نظر گرفته می‌شود.
- خوشه‌بندی نتایج اولیه با الگوریتم K-NN مبتنی بر معیار شباهت حساس به پرس‌وجو.
- رتبه‌بندی خوشه‌ها و انتخاب اسناد بازخورد.

- ترکیب اسناد انتخاب‌شده برای بسط.
- اضافه کردن کلمات انتخاب‌شده به پرس‌وجو و ساخت پرس‌وجوی جدید.

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

در مدل فضای برداری، اسناد و پرس‌وجو به‌صورت برداری از کلمات مجموعه نشان داده می‌شوند. برای یک سند و یا پرس‌وجو، درایه‌های بردار، وزن TF-IDF کلمات موجود در سند و یا پرس‌وجو است. شباهت کسینوسی سند d با پرس‌وجوی q به‌صورت فرمول (۹) محاسبه می‌شود.

(۱)

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

$W(w,q)$ و $W(w,d)$ به ترتیب وزن TF-IDF کلمه‌ی w در پرس‌وجوی q و سند d است. $\|q\|$ و $\|d\|$ اندازه‌ی پرس‌وجو و سند است که مشابه فرمول (۵) و (۶) محاسبه می‌شوند.

شباهت تمام اسناد با فرمول (۹) نسبت به پرس‌وجو محاسبه می‌شود و با مرتب کردن نزولی شباهتشان تعدادی از اسناد رتبه‌بالا به کاربر ارائه می‌شود.

۳-۳-۲- خوشه‌بندی اسناد رتبه‌بالا

$|R|$ سند رتبه‌بالا به‌عنوان اسناد شبه مرتبط انتخاب می‌شوند. از الگوریتم K-NN برای خوشه‌بندی اسناد استفاده می‌شود. برای هر سند، خوشه‌اش مطابق با شباهت حساس به پرس‌وجو بین مرکز خوشه و سایر اسناد ساخته می‌شود. اعضای خوشه‌ها مطابق با شباهتشان نسبت به مرکز خوشه رتبه‌بندی می‌شوند. معیار شباهت حساس به پرس‌وجو در فرمول (۱) آورده شده است.

۳-۳-۳- رتبه‌بندی خوشه‌ها

خوشه‌ها با محاسبه مجموع شباهت حساس به پرس‌وجوی اعضایشان نسبت به مرکز خوشه رتبه‌بندی می‌شوند. از $\frac{1}{C}$ خوشه‌های رتبه‌بالا به‌عنوان خوشه‌های بازخورد برای بسط انتخاب می‌شوند و $\frac{1}{M}$ از اسناد هر خوشه‌ی انتخاب‌شده برای بسط برگزیده می‌شوند و اسناد تکراری فقط یک‌بار انتخاب می‌شوند.

۳-۳-۴- ترکیب اسناد انتخاب‌شده

اسناد انتخاب‌شده باهم ترکیب می‌شوند به‌طوری‌که فراوانی کلمه w در سندِ ترکیبی، برابر با تعداد اسناد انتخاب‌شده از خوشه‌ها است که کلمه w را دارا باشند. سپس کلمات این سند ترکیبی براساس تابع وزن‌دهی TF-IDF رتبه‌بندی می‌شوند و رتبه‌بالاترین کلمه‌ها برای بسط انتخاب می‌شوند.

۳-۳-۵- ساخت پرس‌وجوی بسط داده‌شده

پرس‌وجوی بسط داده‌شده به‌صورت ترکیبی از پرس‌وجوی اصلی و کلمات بسط به‌صورت فرمول (۱۰) ساخته می‌شود [۷].

(۱)

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

که در آن q پرس‌وجوی اصلی است. λ وزن پرس‌وجوی اصلی را نسبت به کلمات بسط نشان می‌دهد و $t_1,\ t_2,\ ...,\ t_e$ها کلمات بسط انتخاب‌شده هستند.

۴- آزمایش‌ها

برای ارزیابی روش ارائه‌شده، ما آزمایش‌ها را روی مجموعه‌داده پزشکی MED [۱۷] انجام داده‌ایم (پرس‌وجوهای ۲۱ تا ۳۰). کارایی راه‌کار ارائه‌شده را با روش بازیابی پایه‌ی مدل فضای برداری (VSM)، روش بازخورد شبه مرتبط (PRF) و حد بالای روش ارائه‌شده (TrueRF) مقایسه کرده‌ایم.

۴-۱- پیکربندی آزمایش‌ها

۴-۱-۱- مجموعه‌داده آزمایش

ما روش ارائه‌شده را روی پرس‌وجوهای مجموعه‌داده‌ی پزشکی MED آزمایش کردیم. 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 پرس‌جوها نیز در مجموعه‌داده آورده شده است. همچنین نمایه‌زنی اسناد و پرس‌وجوهای

مجموعه‌داده با موتور جستجوی متن‌باز indri [۱۸] انجام شده است.

جدول ۱-خلاصه‌ای از جزئیات مجموعه‌داده MED

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

۴-۱-۲- پارامترهای مسئله

تعداد اسناد شبه مرتبط انتخاب‌شده برای بسط می‌تواند مقادیر متفاوتی داشته باشد که از مجموعه مقادیر ($\{۱۰$ ، ۱۵ ، ۲۰ ، ۲۵ ، ۳۰ ، ۳۵ ، ۵۰$\} \in |R|$) انتخاب می‌شود. تعداد کلمات بسط انتخاب‌شده برای پرس‌وجو از مجموعه مقادیر ($\{۲$ ، ۵ ، ۷ ، ۱۰، ۱۵$\} \in e$) انتخاب می‌شود. وزن پرس‌وجوی اولیه نسبت به کلمات بسط از مجموعه مقادیر ($\{۰/۲$ ، ۰/۴ ، ۰/۶ ، ۰/۸$\} \in \lambda$) انتخاب شده است. نسبت تعداد خوشه‌های انتخاب‌شده به‌کل خوشه‌ها در قسمت ۳، از مجموعه مقادیر ($\frac{1}{C} \in \left\{\frac{۱}{۲}, \frac{۱}{۳}, \frac{۱}{۴}\right\}$) انتخاب می‌شود. تعداد اسناد انتخاب‌شده از هر خوشه با نسبت ($\left\{\frac{۱}{۲}، \frac{۱}{۳}، \frac{۱}{۴}\right\}$) به تعداد اعضای خوشه آزمایش شده است. سعی شده است مقادیر انتخاب‌شده برای پارامترها به شکلی باشد تا روند تغییر در کارایی روش ارائه‌شده را نسبت به سایر مقادیر پارامترها نشان دهد. از پرس‌وجوهای (۱ تا ۲۰) برای بدست آوردن مقادیر پارامترها استفاده شده است. مقادیر بدست آمده برای پارامترها در جدول ۲ آورده شده است. از پرس‌وجوهای ۲۱ تا ۳۰ برای ارزیابی کارایی روش ارائه‌شده و مقایسه با روش‌های مبنا استفاده شده است.

۴-۲- مقایسات

برای ارزیابی کارایی و مقایسه روش‌های بیان‌شده در مقاله از معیار متوسط میانگین دقت (MAP) استفاده شده است که در فرمول (۱۱) آورده شده است.

(۱)

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

که در آن Q مجموعه پرس‌وجوهای مورد آزمایش است. $|Q|$ تعداد پرس‌وجوهای مورد آزمایش است. $ap(q)$ متوسط دقت برای پرس‌وجوی q است که در فرمول (۱۲) آورده شده است.

$$ap(q) = \frac{\sum_{k}^{|R|} p@k}{|R|} \qquad (۲)$$

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

جدول ۲- مقادیر بدست آمده برای پارامترها

مدل	پارامتر وزن پرس-وجوی اولیه، λ	اسناد بازخورد، $|R|$	تعداد کلمات بسط، e	نسبت تعداد خوشه‌های بازخورد، $\frac{1}{C}$	نسبت اسناد انتخاب شده، $\frac{1}{M}$
PRF	۰/۴	۲۵	۱۰	-	-
True RF	۰/۴	تعداد اسناد مرتبط بازیابی‌شده	۱۵	-	-
QS-CPRF	۰/۴	۲۵	۱۵	$\frac{۱}{۴}\|R\|$	$\frac{۱}{۳}\|R\|$

۴-۲-۱- روش‌های مورد مقایسه

روش بسط پرس‌وجوی ارائه‌شده در این مقاله با روش‌های زیر مقایسه شده است.

- مدل فضای برداری(VSM): مدل فضای برداری که به‌عنوان روش بازیابی پایه استفاده شده است.
- روش بازخورد شبه مرتبط (PRF): در این روش $|R|$ تعداد از رتبه‌بالاترین اسناد برای بسط انتخاب می‌شود. اسناد باهم ترکیب می‌شوند به‌طوری‌که فراوانی کلمات برابر با تعداد اسناد شبه بازخوردی که آن کلمه را دارا می‌باشند. کلمات سند ترکیبی بدست آمده با تابع رتبه‌بندی TF-IDF رتبه‌بندی شده و رتبه‌بالاترین کلمات برای بسط انتخاب می‌شوند.
- روش بازخورد ارتباطی کامل[۱۱] (TrueRF): این روش حد بالای روش خوشه‌بندی اسناد شبه مرتبط را نشان می‌دهد. در این روش اسناد مرتبط از بین $|R|$ سند رتبه بالا برای بسط انتخاب می‌شوند ($|R|=25$). اسناد باهم 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، زیرا از همه‌ی اسناد مرتبط در $|R|$ سند رتبه‌بالا برای بسط استفاده شده است.

۴-۲-۲- نتایج آزمایش‌ها

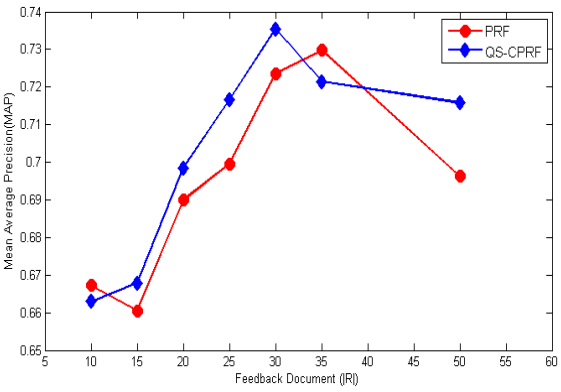
هرکدام از روش‌های مورد مقایسه با مقادیر بدست آمده برای پارامترهای خودشان روی پرس‌وجوهای شماره ۲۱ تا ۳۰ آزمایش شده‌اند. نتایج آزمایش‌ها روی مجموعه داده‌ی پزشکی MED در جدول ۳ آورده شده است.

جدول ۲ -کارایی(MAP) روش‌های VSM، PRF، QS-CPRF و TrueRF

روش	VSM	PRF	QS-CPRF	TrueRF
MAP	۰/۶۶۶۱	۰/۶۹۹۶	۰/۷۱۶۶	۰/۷۷۶۶

روش‌های بازخورد شبه مرتبط و خوشه‌بندی اسناد شبه مرتبط نسبت به روش بازیابی پایه VSM به ترتیب ۵ و ۷/۶ درصد کارایی (متوسط میانگین دقت) را افزایش داده-اند. با توجه به جدول ۳، روش خوشه‌بندی اسناد بازخورد با شباهت حساس به پرس‌وجو (QS-CPRF) نسبت به روش بازخورد شبه مرتبط (PRF) اثربخشی بیشتری روی بازیابی دارد. بهبود ۲/۵ درصدی نسبت به PRF عملکرد بهتر روش خوشه‌بندی را نسبت به فرض مرتبط بودن اسناد رتبه بالا، نشان می‌دهد. اختلاف کارایی ۸/۳ درصدی بین روش QS-CPRF و TrueRF نشان می‌دهد که می‌توان کارایی را افزایش داد.

تغییرات کارایی روش‌های PRF و QS-CPRF نسبت به تعداد اسناد بازخورد در شکل ۱ آورده شده است. بیشترین کارایی روش‌های PRF و QS-CPRF به ترتیب با ۳۵ و ۳۰ سند رتبه‌بالا بدست آمده است. بااین‌حال بیشترین کارایی روش PRF نسبت به بیشینه‌ی کارایی روش QS-CPRF کمتر است. کارایی بیشتر روش QS-CPRF نشان می‌دهد که این روش اسناد خطا را از فرآیند بسط خارج می‌کند تا کلمات مرتبط‌تری به پرس‌وجو اضافه شود.



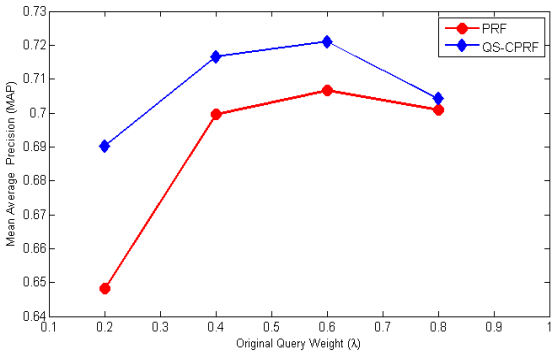
شکل ۱- بررسی رفتار روش‌های بسط پرس‌وجو نسبت به تعداد اسناد بازخورد

در شکل ۲ تغییرات کارایی نسبت به وزن پرس‌وجوی اصلی نشان داده شده است. با توجه به شکل ۲، بیشترین کارایی

[۱۱] True Relevance Feedback

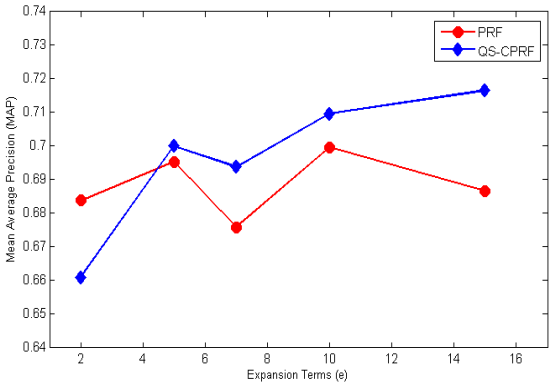
روش PRF در اختصاص وزن ۰/۶ به پرس‌وجوی اصلی بدست آمده است. برای روش QS-CPRF بیشترین کارایی در وزن ۰/۶ بدست آمده است.

همچنین اختصاص وزن بیشتر و یا کمتر به پرس‌وجوی اصلی کارایی را کاهش می‌دهد که نشان می‌دهد افزودن کلمات بسط به پرس‌وجو با وزن بیشتر کارایی را کاهش خواهد داد.



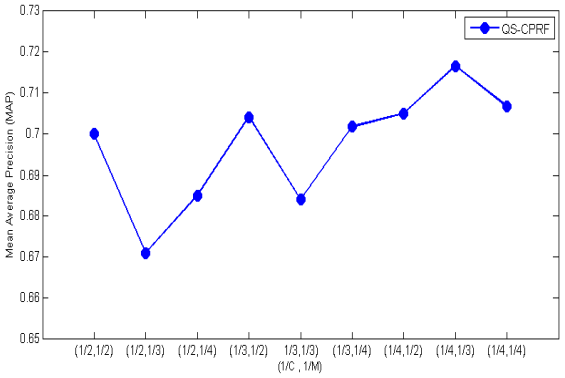
شکل ۲- کارایی نسبت به وزن پرس‌وجوی اصلی

تأثیر تعداد کلمات بسط بر کارایی روش‌های PRF و QS-CPRF در شکل ۳ بررسی شده است. به توجه به شکل ۳، هر دو روش بیشترین کارایی را در ۱۵ کلمه‌ی بسط بدست آورده‌اند. تغییر کارایی روش QS-CPRF نشان می‌دهد که اضافه کردن ۵ کلمه‌ی اول بیشترین تغییر را بدست می‌آورد و این ۵ کلمه بسیار مرتبط به موضوع پرس‌وجو هستند به‌طوری‌که کلمات بعدی تغییر چندانی در کارایی ندارند. همچنین روند نمودار نشان می‌دهد که کلمات بسط بیشتر متوسط میانگین دقت (MAP) را افزایش می‌دهد.



شکل ۳- کارایی نسبت به تعداد مختلف کلمات بسط

تغییرات کارایی برای مقادیر مختلف پارامترهای $\frac{1}{C}$ و $\frac{1}{M}$ در شکل ۶ نشان داده شده است. مقادیر مختلف برای این دو پارامتر به مقدار اسناد شبه بازخورد ($|R|$) بستگی دارد. نمودار شکل ۶ در مقدار ۲۵ برای پارامتر $|R|$ بدست آمده است. تعداد اسناد انتخاب‌شده برای بسط به مقادیر این دو پارامتر بستگی دارد. همچنین مقادیر مختلف برای پارامترها در تعداد اسناد خطا و تعداد اسناد مرتبط برای بسط اثرگذار است.



شکل ۴ - تغییرات کارایی (MAP) نسبت به مقادیر مختلف پارامترهای $\frac{1}{C}$ و $\frac{1}{M}$

کارایی (میانگین متوسط دقت) روش‌های مدل فضای برداری (VSM)، بازخورد شبه مرتبط (PRF)، خوشه‌بندی اسناد شبه بازخورد (QS-CPRF) و حد بالای روش خوشه-بندی (TrueRF) روی تک‌تک پرس‌وجوهای شماره ۲۱ تا ۳۰ در شکل ۴ آورده شده است. در بعضی از پرس‌وجوها روش‌های بسط کارایی را نسبت به بازیابی پایه (مدل فضای برداری) کاهش داده‌اند که عدم توانایی روش بسط در انتخاب اسناد مرتبط را نشان می‌دهد. حتی حد بالای روش QS-CPRF نیز برای ۲ پرس‌وجو کارایی را نسبت به سایر روش‌ها کمتر کرده است.

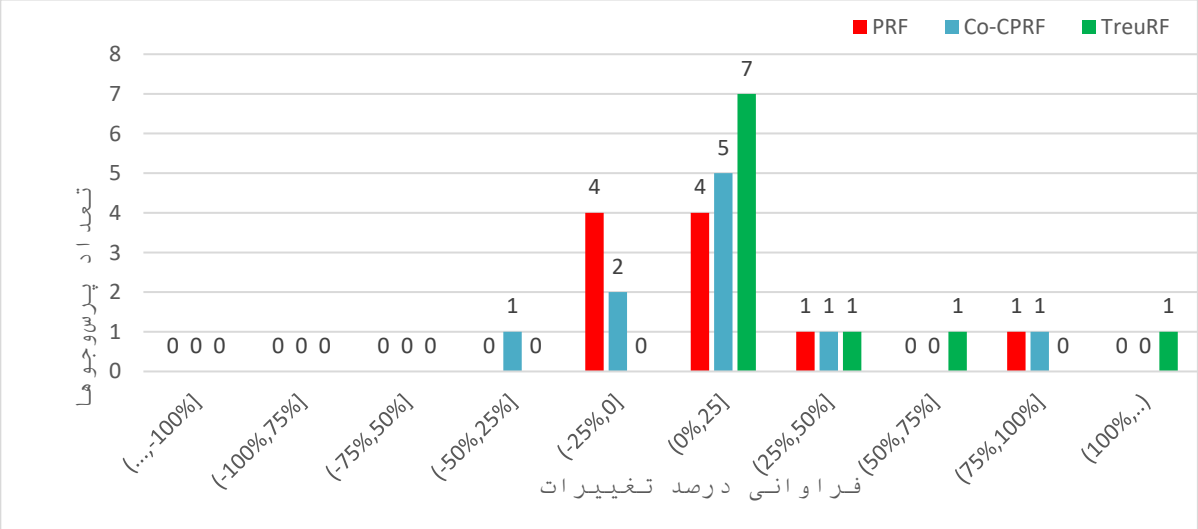
۴-۲-۳- مقاومت‌پذیری

فراوانی درصد بهبود کارایی روش‌های مورد مقایسه روی تک‌تک پرس‌وجوها در شکل ۵ نشان می‌دهد که روش PRF در ۴ پرس‌وجو کارایی را کاهش داده است درحالی‌که

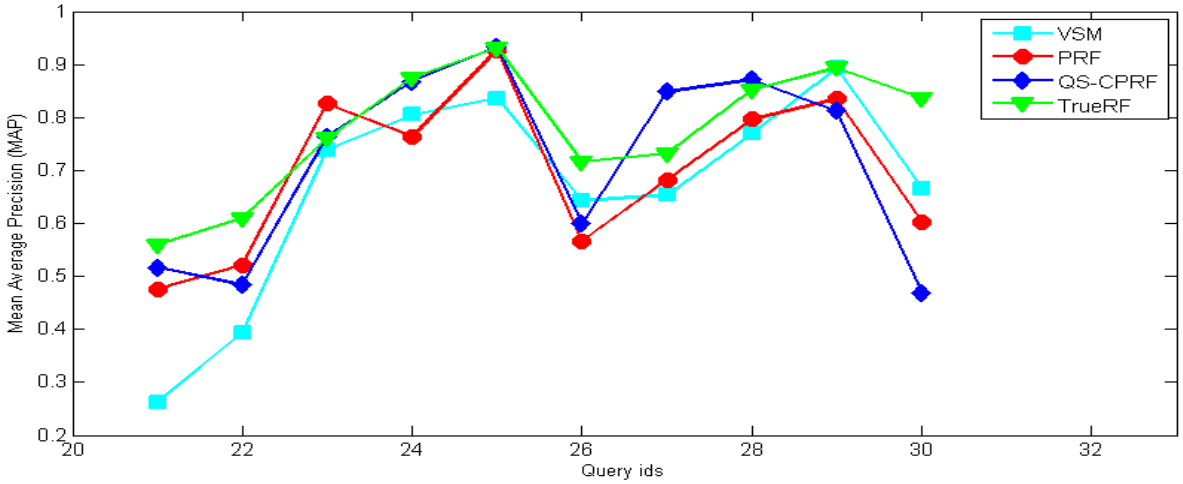
روش خوشه‌بندی اسناد شبه بازخورد در ۳ پرس‌وجو کارایی را کاهش داده است و عملکرد بهتر خوشه‌بندی اسناد بازخورد را نشان می‌دهد. همچنین روش خوشه‌بندی همانند بازخورد شبه مرتبط ۱ پرس‌وجو را بالای ۵۰ درصد بهبود داده است.

۴-۳- بررسی عملکرد خوشه‌بندی در انتخاب اسناد مرتبط و نادیده گرفتن اسناد نامرتبط

عملکرد روش خوشه‌بندی اسناد شبه بازخورد در انتخاب اسناد مرتبط برای بسط و نادیده گرفتن اسناد نامرتبط، در جدول ۴ آورده شده است. با توجه به جدول ۴، روش خوشه‌بندی از ۲۵ سند رتبه بالا، به‌طور متوسط ۱۳/۴ سند را برای بسط انتخاب می‌کند. که از این تعداد، به‌طور متوسط ۹/۶ سند مرتبط و تقریباً کمتر از ۴ سند نامرتبط به پرس‌وجو است. درحالی‌که در بین ۲۵ سند رتبه‌بالا به‌طور متوسط ۱۲ سند مرتبط وجود دارد و ۱۳ سند دیگر نامرتبط به پرس‌وجو هستند. نسبت تعداد اسناد مرتبط به همه‌ی اسناد انتخاب‌شده برای بسط در روش خوشه‌بندی اسناد شبه بازخورد (QS-CPRF) ۰/۷۲ است درحالی‌که برای روش بازخورد شبه مرتبط (CPRF) برابر ۰/۴۸ است. نسبت تعداد اسناد مرتبط و اسناد نامرتبط به اسناد انتخاب‌شده برای بسط نشان می‌دهد که روش خوشه‌بندی اسناد بازخورد (QS-CPRF)، سعی می‌کند از حضور اسناد خطا در فرآیند بسط جلوگیری کند.



شکل ۵- مقاومت‌پذیری QS-CPRF نسبت به TrueRF و PRF



شکل ۶- کارایی پرس‌وجوهای شماره ۱۱ تا ۳۰ مورد آزمایش قرار گرفته است.

جدول ۳- بررسی خوشه‌بندی اسناد شبه بازخورد در انتخاب اسناد مرتبط و نادیده گرفتن اسناد خطا- تعداد اسناد شبه بازخورد، ۲۵

	متوسط	شماره پرس‌وجوها									
		۳۰	۲۹	۲۸	۲۷	۲۶	۲۵	۲۴	۲۳	۲۲	۲۱
تعداد اسناد مرتبط به پرس‌وجو	۲۳	۱۴	۳۷	۳۹	۱۸	۲۸	۲۴	۲۲	۳۹	۲۵	۲۷
تعداد اسناد مرتبط بازیابی‌شده	۱۲	۸	۲۲	۱۷	۷	۱۱	۱۸	۱۷	۱۶	۷	۸
تعداد اسناد انتخاب‌شده-	۱۳/۴	۱۵	۱۸	۱۳	۱۱	۱۱	۱۴	۱۲	۱۲	۱۲	۱۶
تعداد اسناد مرتبط انتخاب‌شده	۹/۶	۷	۱۶	۱۳	۶	۸	۱۲	۱۲	۱۰	۵	۷

۵- نتیجه‌گیری

خوشه‌بندی اسناد شبه مرتبط راه‌کار مؤثری برای انتخاب اسناد مرتبط و نادیده گرفتن اسناد خطا در فرآیند بسط است. نتایج بازیابی روش خوشه‌بندی (QS-CPRF) نسبت به روش بازخورد شبه مرتبط (PRF) و روش پایه‌ی مدل فضای برداری (VSM) بهتر شده است. استفاده از شباهت حساس به پرس‌وجو در خوشه‌بندی اثربخشی بازیابی را افزایش داد که شباهت بین دو سند را صریحاً با در نظر گرفتن پرس‌وجو اندازه می‌گیرد. آزمایش‌ها روی مجموعه داده‌ی MED بهتر شدن نتایج QS-CPRF نسبت به PRF را نشان می‌دهد که هر دو روش نسبت به روش VSM به ترتیب ۵ و ۷/۶ درصد کارایی را افزایش داده‌اند.

فاصله‌ی کارایی روش خوشه‌بندی نسبت به TrueRF نشان می‌دهد که می‌توان برای کارهای آتی بررسی‌های بیشتری روی خوشه‌های ساخته‌شده انجام داد. ادغام و یا حذف خوشه‌ها و انتخاب نسبت مناسب‌تری از خوشه‌ها و اسناد می‌تواند نتایج جستجو را بهبود دهد. استفاده از توابع رتبه-بندی بهتر برای کلمات بسط می‌تواند کارایی را تغییر دهد. برای بدست آوردن موضوع دقیق پرس‌وجو و اسناد می‌توان از روش‌های معنایی و تحلیل زبانی استفاده کرد تا اسناد با دقت بیشتری برای بسط انتخاب شوند.

۶- مراجع

[1] G. O. History. (2014). Google Annual Search Statistics. Available: http://www.statisticbrain.com/google-searches/
[2] Krovetz, R. (1997, July). Homonymy and polysemy in information retrieval. InProceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (pp. 72-79). Association for Computational Linguistics.
[3] Spink, A., & Jansen, B. J. (2004). A study of web search trends. Webology,1(2), 4.
[4] Sanderson, M. (2008, July). Ambiguous queries: test collections need more sense. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 499-506). ACM.
[5] Xu, J., & Croft, W. B. (1996, August). Query expansion using local and global document analysis. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 4-11). ACM.
[6] Huang, J. X., Miao, J., & He, B. (2013). High performance query expansion using adaptive co-training. Information Processing & Management, 49(2), 441-453.
[7] Lee, K. S., & Croft, W. B. (2013). A deterministic resampling method using overlapping document clusters for pseudo-relevance feedback. Information Processing & Management, 49(4), 792-806.
[8] Bashir, S. (2012). Improving retrievability with improved cluster-based pseudo-relevance feedback selection. Expert Systems with Applications, 39(8), 7495-7502.
[9] Lavrenko, V., & Croft, W. B. (2001, September). Relevance based language models. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 120-127). ACM.
[10] Lee, K. S., Park, Y. C., & Choi, K. S. (2001). Re-ranking model based on document clusters. Information processing & management, 37(1), 1-14.
[11] Lee, K. S., Kageura, K., & Choi, K. S. (2004). Implicit ambiguity resolution using incremental clustering in cross-language information retrieval. Information processing & management, 40(1), 145-159.
[12] Tombros, A., & van Rijsbergen, C. J. (2001, October). Query-sensitive similarity measures for the calculation of interdocument relationships. InProceedings of the tenth international conference on Information and knowledge management (pp. 17-24). ACM.
[13] Rocchio, J. J. (1971). Relevance feedback in information retrieval.
[14] Sakai, T., Manabe, T., & Koyama, M. (2005). Flexible pseudo-relevance feedback via selective sampling. ACM Transactions on Asian Language Information Processing (TALIP), 4(2), 111-135.
[15] Jardine, N., & van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. Information storage and retrieval, 7(5), 217-240.
[16] Na, S. H. (2013). Probabilistic co-relevance for query-sensitive similarity measurement in information retrieval. Information Processing & Management,49(2), 558-575.
[17] U. o. Glascow. (2014/03). Medline collection. Available: http://ir.dcs.gla.ac.uk/resources/test_collections/medl/
[18] Strohman, T., Metzler, D., Turtle, H., & Croft, W. B. (2005, May). Indri: A language model-based search engine for complex queries. In Proceedings of the International Conference on Intelligent Analysis (Vol. 2, No. 6, pp. 2-6).