

## اثربخشی بسط پرس و جو مبتنی بر خوشه‌بندی اسناد شبه‌بازخورد با الگوریتم K-NN

رضا خدایی<sup>۱</sup>، دانشجوی کارشناسی ارشد؛ محمدعلی بالافر<sup>۲</sup>، استادیار؛ سیدناصر رضوی<sup>۳</sup>، 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 پرس و جو یکی از روش‌های مؤثر در بهبود اثربخشی نتایج بازیابی اطلاعات است. روش بازخورد شبه‌مرتبط (PRF) فرض می‌کند که اسناد رتبه‌بالا از نتایج اولیه بازیابی شده مرتبط به پرس و جو 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 پرس و جو 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 پرس و جو، بازخورد شبه‌مرتبط، کلمات بسط، اسناد بازخورد.

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

Reza Khodaei<sup>1</sup>, MSc Student; Mohammad Ali Balafar<sup>2</sup>, Assistant Professor; Seyyed Naser Razavi<sup>3</sup>, 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: ۱۳۹۲/۰۸/۱۴

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

جستجو بر اساس پرس و جوی ارسالی کاربر انجام می‌شود و این امکان وجود دارد که پرس و جوی ارسالی کاربر دقیق و عاری از ابهام نباشد. معمولاً مشکلات زبان طبیعی، انتخاب کلمات نامناسب در پرس و جو، وجود کلمات مبهم با چند معنی [۱] و کوتاهی پرس و جوها [۲] مشکلات موجود در پرس و جوها هستند که معمولاً کاربران مبتدی با این مشکلات روبرو می‌شوند. تقریباً ۷ تا ۲۳ درصد از پرس و جوهای موتور جستجو کمتر از ۳ کلمه دارند [۳]. بسط پرس و جو یکی از روش‌های انطباق پرس و جو است که سعی در برطرف کردن مشکلات مذکور را دارد. در بسط پرس و جو تعدادی کلمه مرتبط به موضوع پرس و جو اضافه می‌شوند به طوری که بازیابی با پرس و جوی بسط داده شده کارایی را افزایش می‌دهد.

در بازیابی اطلاعات، روش بازخورد شبه مرتب<sup>۱</sup> (PRF) [۴] چارچوبی برای بهبود کارایی سامانه‌های بازیابی اطلاعات است. در این روش نتایج بازیابی شده اولیه، مرتبط به موضوع پرس و جو در نظر گرفته شده و تعدادی از اسناد رتبه بالا به منظور استخراج کلمات بسط انتخاب می‌شوند. در حالی که ممکن است موتور جستجو اسناد نامرتب را به عنوان اسناد مرتبط تشخیص دهد. فرض مرتبط بودن نتایج اولیه در روش بازخورد شبه مرتب ممکن است اسناد نامرتب را در فرآیند بسط دخیل کند و کلمات نامرتب به پرس و جو اضافه شود. برای جلوگیری از این مشکل روش‌هایی به منظور انتخاب بهتر اسناد بازخورد برای بسط انجام شده است [۵-۷]. به طور کلی انتخاب بهترین اسناد، هدف کلیه روش‌های بسط پرس و جوی مبتنی بر روش بازخورد شبه مرتب است تا کلمات بسط مرتبط‌تری انتخاب شود.

برای به دست آوردن اسناد بازخورد مناسب، طبقه بندی اسناد بازخورد در [۵] ارائه شده است. در این روش، از یادگیری همکار برای به دست آوردن داده‌های آموزشی و طبقه بندی اسناد نتایج اولیه استفاده شده است. اسناد با برچسب مثبت به عنوان منبع برای بسط پرس و جو استفاده می‌شوند.

خوشه بندی اسناد حاصل از نتایج اولیه بر اساس شباهت مبتنی بر کلمه در [۶] انجام گرفته است که در آن اسناد موجود در هم پوشانی خوشه‌های رتبه بالا به عنوان منبع برای بسط پرس و جو استفاده شده است. برای انتخاب کلمات بسط از مدل ربط لاورنکو<sup>۲</sup> [۸] استفاده شده است. نتایج بازیابی این روش، بهبود را در معیار متوسط میانگین دقت<sup>۳</sup>

(MAP) نشان می‌دهد. همچنین، باز رتبه بندی اسناد بازیابی شده با خوشه بندی اسناد بازیابی شده با استفاده از مدل فضای برداری برای بازیابی، نتایج موفقیت آمیزی را داشته است [۹، ۱۰]. هدف خوشه بندی قرار دادن شبیه ترین اسناد کنار هم است. شباهت کسینوسی با ویژگی‌های مبتنی بر کلمه اسناد، معیار مناسبی برای شباهت بین اسناد هست و برای بازیابی اسناد و خوشه بندی اسناد بازخورد استفاده شده است [۶، ۱۱]. شباهت کسینوسی، شباهت ضریب جاکارد<sup>۴</sup>، شباهت ضریب دایس<sup>۵</sup>، [۱۲] KL-دیورژانس<sup>۶</sup>، [۱۳] برای محاسبه شباهت بین دو سند و یا شباهت بین سند و پرس و جو به کار می‌روند.

در این مقاله، از خوشه بندی اسناد شبه بازخورد (CPRF) برای به دست آوردن منبع برای انتخاب کلمات بسط استفاده می‌شود که شباهت‌هایی به راهکار ارائه شده در مرجع [۶] دارد. در روش ارائه شده در این مقاله، برای خوشه بندی اسناد، از الگوریتم نزدیک ترین K همسایه<sup>۷</sup> (K-NN) استفاده شده است. قطعیت الگوریتم K-NN از تولید نتایج تصادفی جلوگیری می‌کند زیرا در هیچ یک از مراحل الگوریتم عملی تصادفی اتفاق نمی‌افتد. اسناد بازخورد برای بسط از اسناد رتبه بالای خوشه‌های رتبه بالا انتخاب می‌شوند در حالی که اسناد بازخورد برای بسط در [۶] از هم پوشانی خوشه‌ها با افزودن انتخاب می‌شوند. کلمات بسط با تابع رتبه بندی TF-IDF<sup>۸</sup> رتبه بندی می‌شوند. نتایج آزمایش‌های انجام گرفته روی مجموعه داده پزشکی MED [۱۴] نشان می‌دهد که خوشه بندی اسناد بازخورد به منظور انتخاب اسناد بهتر، معیار متوسط میانگین دقت (MAP) را بهبود می‌دهد.

در ادامه این مقاله کارهای پیشین آورده شده است. در قسمت ۳ روش ارائه شده در این مقاله برای بسط بیان شده و در قسمت ۴ مراحل انجام آزمایش‌ها و چگونگی انجام آن بیان شده و نتایج آزمایش‌های انجام گرفته روی مجموعه داده MED آورده شده است و در بخش ۵ نتیجه گیری مقاله بیان شده است.

## ۲- کارهای پیشین

بازخورد مرتبط<sup>۹</sup> (RF) [۱۵] روشی مؤثر برای بهبود دقت بازیابی با بازساختاری پرس و جوی 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 نامرتب برچسب می‌زند و بازیابی بعدی با اطلاعات بازخورد انجام می‌گیرد. بازخورد شبه مرتب (PRF) [۴]، شکل خودکار بازخورد مرتبط است و فرض می‌کند که تعدادی از اسناد رتبه بالا در بازیابی اولیه مرتبط به موضوع پرس و جو هستند و از این اسناد رتبه بالا به عنوان منبعی برای انتخاب کلمات بسط استفاده می‌کند. تعدادی از کلمات بسط به پرس و جو اضافه شده و بازیابی نهایی با پرس و جوی بسط داده شده انجام می‌گیرد. آزمایش‌های این روش روی مجموعه داده‌های مختلف بهبود نتایج را نشان می‌دهد.

می‌شوند. آزمایش‌های این روش روی مجموعه داده‌های 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 انجام شده 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 انتخاب شده 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 به دست آمده است. تابع رتبه‌بندی TF-IDF [Okapi BM24, IDF] و مدل زبان لاورنکو برای رتبه‌دهی کلمات بسط به کار 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 به صورت بردار با ویژگی‌های مبتنی بر کلمه نشان داده می‌شوند. کلمه‌های اسناد با تابع وزن‌دهی TF-IDF وزن‌دهی شده و به صورت بردار برای هر سند نشان داده می‌شود.

#### ۳-۱- اندازه‌گیری شباهت کسینوسی بین دو سند

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

$$\cos\_sim(d_1, d_2) = \frac{W(w, d_1) \cdot W(w, d_2)}{\|d_1\| \cdot \|d_2\|} \quad (1)$$

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  $W(w, d_2)$  وزن  $w$  در سند  $d_2$  است که با تابع وزن‌دهی TF-IDF در فرمول (۲) محاسبه شده است.

$$TF-IDF(w)_d = tf(w)_d \cdot idf(w) \quad (2)$$

$$idf(w) = \log_{10} \left( \frac{N}{df(w)} \right) \quad (3)$$

که در آن  $f(w)_d$ ، فراوانی کلمه  $w$  در سند  $d$  را نشان می‌دهد.  $N$  تعداد اسناد مجموعه داده و  $df(w)$  تعداد اسنادی از مجموعه داده که کلمه  $w$  را دارا می‌باشند.  $\|d_1\|$  اندازه سند ۱ و  $\|d_2\|$  اندازه سند ۲ هست که به صورت رابطه (۴) به دست می‌آید.

$$\|d\| = \sqrt{\sum_w W(w, d)^2} \quad (4)$$

برای انتخاب بهترین اسناد، خوشه‌بندی اسناد نتایج اولیه می‌تواند راه‌حل مناسبی باشد. اسناد رتبه‌بالاتر به صورت 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 به عنوان 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 مجموعه داده‌های حجیم مثل WT10g بهتر شدن نتایج جستجو را نسبت به مدل زبان<sup>۱۰</sup> 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 ایجاد شده 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 رتبه‌بالاتر است و سعی در نادیده گرفتن این اسناد را دارد. این نوع برگزینی اسناد برای خوشه‌ها، منجر به نتایج چندان مطلوبی روی مجموعه داده‌های NTCIR نشد.

از خوشه‌بندی با تمایل<sup>۱۱</sup> برای خوشه‌بندی اسناد نتایج اولیه در [۱۷] استفاده شده است. تعدادی از اسناد که بیشترین تفاوت را نسبت به یکدیگر داشته باشند به عنوان مرکز خوشه‌ها انتخاب می‌شوند. برای مرکزهای انتخاب شده 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 ساخته شده کلمات بسط استخراج شده و پرس‌جوهای 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 طبقه‌بند، تعدادی از اسناد رتبه بالا به عنوان طبقه مثبت و تعدادی از اسناد رتبه پایین به عنوان طبقه منفی در نظر گرفته شده‌اند [۵]. طبقه‌بند ساخته شده، سایر اسناد رتبه‌بالای اولیه را طبقه‌بندی می‌کند. اسناد برچسب‌گذاری شده با برچسب مثبت به عنوان اسناد بازخورد انتخاب می‌شوند و کلمه‌های بسط از این اسناد استخراج

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

اسناد و پرس و جو به صورت بردار نشان داده می شوند. درایه های بردار، وزن TF-IDF کلمات موجود در اسناد و پرس و جو است. شباهت اسناد با پرس و جو با شباهت کسینوسی محاسبه می شود. برای سند  $d$  و پرس و جوی  $q$ ، شباهت آن ها با شباهت کسینوسی نرمال سازی شده به صورت فرمول (۵) محاسبه می شود.

$$sim(d, q) = \frac{W(w, d) W(w, q)}{\|d\| \|q\|} \quad (5)$$

در  $W(w, d)$  و  $W(w, q)$  به ترتیب وزن TF-IDF کلمه  $w$  در پرس و جوی  $q$  و سند  $d$  است. شباهت تمام اسناد با فرمول (۵) نسبت به پرس و جو محاسبه می شود و با مرتب کردن نزولی شباهت آن ها تعدادی از اسناد به کاربر به عنوان بازیابی اولیه ارائه می شود.

### ۳-۲-۲- خوشه بندی اسناد رتبه بالا

$|R|$  تعداد از اسناد رتبه بالا به عنوان اسناد شبه مرتبط انتخاب می شود. اسناد به صورت بردار با وزن دهی TF-IDF کلمات نشان داده می شوند. از الگوریتم خوشه بندی K-NN برای قرار دادن اسناد مشابه در کنار هم استفاده می شود. استفاده از الگوریتم K-NN نتایج قطعی تولید می کند و برخلاف الگوریتم های خوشه بندی که از مرکزهای تصادفی برای خوشه بندی استفاده می کنند، نتایج یکسانی را در اجراهای مختلف تولید می کند. معیار خوشه بندی شباهت کسینوسی هست که در فرمول (۱) آورده شده است.

### ۳-۲-۳- رتبه بندی خوشه ها

خوشه ها با محاسبه مجموع شباهت کسینوسی اعضای آن ها نسبت به پرس و جو امتیازدهی می شوند و خوشه ها به صورت نزولی رتبه بندی می شوند.  $1/C$  از خوشه ها به عنوان خوشه های بازخورد برای بسط انتخاب می شوند و  $1/M$  از اسناد خوشه های انتخاب شده برای بسط بدون افزودگی برگزیده می شوند (در الگوریتم خوشه بندی K-NN هم پوشانی اتفاق می افتد و در انتخاب اسناد ممکن است افزودگی اتفاق بیافتد. در این مقاله، اسناد بدون افزودگی انتخاب می شوند در حالی که در [۶]، اسناد چیره با افزودگی انتخاب می شوند).

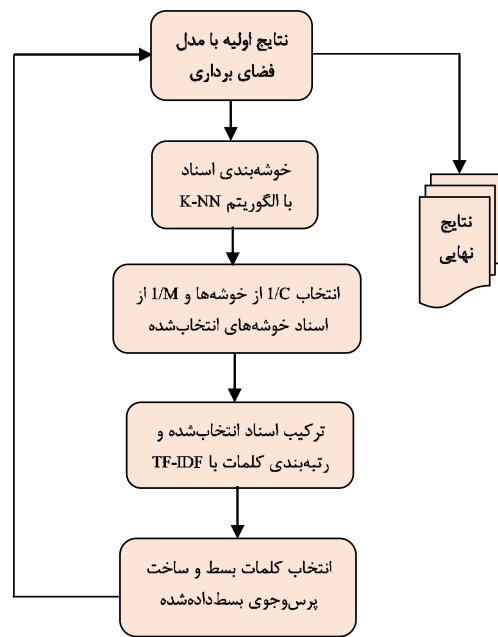
### ۳-۲-۴- ترکیب اسناد انتخاب شده

اسناد انتخاب شده با هم ترکیب می شوند به طوری که فراوانی کلمه  $w$  در آن برابر با تعدادی از اسناد انتخاب شده از خوشه ها است که کلمه  $w$  را دارا هستند. برای انتخاب کلمات بسط از تابع وزن دهی TF-IDF استفاده کرده ایم که در بازیابی اسناد کلمات با این تابع رتبه بندی شده اند و رتبه بالاترین کلمه ها برای بسط انتخاب می شوند (از مدل ربط لاورنکو برای انتخاب کلمات بسط استفاده شده است [۶] که بازیابی با مدل زبان انجام می گیرد).

### ۳-۲- خوشه بندی اسناد شبه بازخورد با شباهت کسینوسی (CPRF)

در این مقاله، بازیابی پایه با مدل فضای برداری انجام می شود. پرس و جوها و اسناد با تابع وزن دهی TF-IDF وزن دهی شده و به صورت بردار نشان داده می شوند. میزان شباهت اسناد به پرس و جو با شباهت کسینوسی نرمال سازی شده محاسبه شده و رتبه بالاترین اسناد به عنوان نتایج مرتبط برگردانده می شود. مراحل انتخاب اسناد بازخورد و بسط پرس و جو در این مقاله در پنج مرحله بیان شده است و به صورت کلی در شکل ۱ نشان داده شده است.

- بازیابی اسناد با مدل فضای برداری:  $|R|$  سند برای خوشه بندی در نظر گرفته می شود.
  - خوشه بندی اسناد رتبه بالا: اسناد به صورت بردار بر اساس تابع وزن دهی TF-IDF نشان داده می شوند. شباهت کسینوسی برای محاسبه شباهت بین اسناد استفاده شده است. شباهت کسینوسی بین ۲ سند با فرمول (۱) محاسبه می شود.
  - رتبه بندی خوشه ها: خوشه ها بر اساس شباهت اعضایشان نسبت به مرکز خوشه رتبه بندی می شوند.  $1/C$  از خوشه ها انتخاب شده و  $1/M$  از اسناد خوشه ها برای بسط انتخاب می شوند.
  - ترکیب اسناد انتخاب شده: اسناد انتخاب شده با هم ترکیب شده و کلمات حاصل از آن با تابع وزن دهی TF-IDF که در فرمول (۲) آورده شده، رتبه بندی شده و برای بسط استخراج می شوند. فراوانی کلمه ها در سند ترکیب شده برابر با تعداد اسنادی از نتایج اولیه که شامل آن کلمه هستند محاسبه می شود.
- ساخت پرس و جوی بسط: کلمات بسط انتخاب شده به پرس و جو اضافه می شوند.



شکل ۱: فلوچارت مراحل راهکار ارائه شده در مقاله

## ۳-۲-۵- ساخت پرس و جوی بسط داده شده

همانند فرمول پرس و جوی بسط در [۶]، پرس و جوی بسط داده شده به صورت ترکیبی از پرس و جوی اصلی و کلمات بسط به صورت فرمول (۶) ساخته می شود.

$$\lambda(q) + (1-\lambda)(t_1, t_2, \dots, t_e) \quad (6)$$

$q$  پرس و جوی اولیه است که به صورت مجموعه ای از کلمات مستقل از هم نشان داده می شود.  $\lambda$  وزن پرس و جوی اولیه را در پرس و جوی بسط داده شده نشان می دهد.  $t_1, t_2, \dots, t_e$  کلمات انتخاب شده برای بسط هستند که با مقدار  $(1-\lambda)$  نسبت به پرس و جوی بسط داده شده وزن دهی می شوند. پس از تشکیل این پرس و جو، بازیابی نهایی با این پرس و جو اجرا می شود.

## ۴- آزمایش ها

## ۴-۱- پیکربندی آزمایش ها

## ۴-۱-۱- مجموعه داده آزمایش

راهکار ارائه شده روی مجموعه داده پزشکی MED [۱۴] اعمال می شود. اسناد این مجموعه داده، چکیده ای از مقالات پزشکی به صورت فشرده است که خلاصه ای از جزئیات این مجموعه داده در جدول ۱ آورده شده است. اسناد مرتبط به پرس و جویها نیز در مجموعه داده مشخص شده است. همچنین نمایه زنی اسناد و پرس و جویهای مجموعه داده با موتور جستجوی متن باز: indri [۲۰] انجام شده است. پس از نمایه زنی، اسناد و پرس و جویها به صورت مجموعه ای از کلمات مستقل از هم نشان داده می شوند.

جدول ۱: خلاصه ای از جزئیات مجموعه داده

۱۰۳۳	تعداد اسناد
۱۶۲	متوسط اندازه اسناد
۹۷۶۰	تعداد نشانه ها
۳۰	تعداد پرس و جویها
۲۳	متوسط تعداد اسناد مرتبط به هر پرس و جو

## ۴-۱-۲- یادگیری و ارزیابی پارامترها

کل مجموعه داده به دو قسمت آموزشی و آزمایش تقسیم شده است. یادگیری پارامترها با ۱۰ پرس و جوی اول به عنوان داده آموزشی انجام می گیرد. تعداد اسناد شبه بازخورد انتخاب شده برای بسط می تواند مقادیر متفاوتی داشته باشد که از مجموعه مقادیر  $(|R| \in \{10, 15, 20, 25, 30, 50\})$  انتخاب می شود. تعداد کلمات بسط انتخاب شده برای اضافه شدن به پرس و جو از مجموعه مقادیر  $(e \in \{2, 5, 7, 10, 15\})$  کلمات بسط در پرس و جوی بسط داده شده با وزنی نسبت به پرس و جوی اولیه ترکیب می شود که وزن پرس و جوی اولیه از مجموعه مقادیر  $(\lambda \in \{0.2, 0.4, 0.6, 0.8\})$  انتخاب می شود. با توجه به این که کلمات بسط از اسناد انتخاب شده از خوشه ها انتخاب می شوند لذا نسبت تعداد خوشه های انتخاب شده به

تعداد همه خوشه ها در قسمت ۳ مقاله از مجموعه مقادیر  $(1/C \in \{1/2, 1/3, 1/4\})$  انتخاب می شود. تعداد اسناد انتخاب شده از خوشه های انتخاب شده نیز با نسبت های  $(1/M \in \{1/2, 1/3, 1/4\})$  مورد ارزیابی قرار گرفته است. مقادیر پارامترهای یادگیری شده در جدول ۲ آورده شده است. سعی شده است تا مقادیر مختلف برای پارامترها طوری انتخاب شود تا رفتار و کارایی راهکار ارائه شده برای مقادیر مختلف نشان داده شود. بعد از یادگیری پارامترها، از پرس و جویهای ۱۱ تا ۳۰ برای انجام آزمایش ها و مقایسه روش ارائه شده با روش های مبنا استفاده شده است.

## ۴-۲- معیار مقایسات

برای مقایسه روش های بیان شده در مقاله، کارایی را با معیار متوسط میانگین دقت (MAP) اندازه می گیریم که به صورت فرمول (۷) است.

$$MAP = \frac{1}{|Q|} \sum_{q \in Q} ap(q) \quad (7)$$

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

$$ap(q) = \frac{\sum_k^{|R|} p@k}{|R|} \quad (8)$$

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

## ۴-۲-۱- روش های مورد مقایسه مبنا

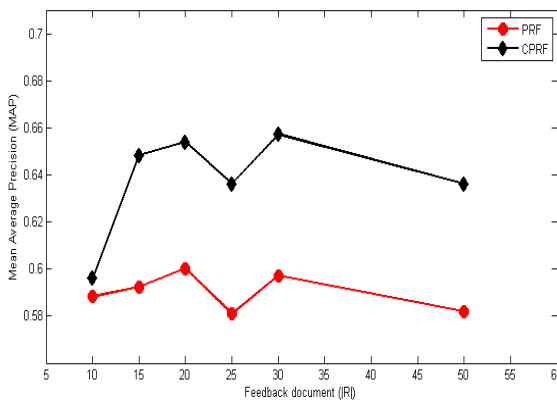
روش بسط پرس و جوی ارائه شده در این مقاله با روش های زیر به عنوان مبنا مقایسه شده است:

- مدل فضای برداری (VSM): مدل فضای برداری به عنوان روش بازیابی پایه استفاده شده است.
- روش بازخورد شبه مرتبط (PRF): در این روش  $|R|$  تعداد از رتبه بالاترین اسناد برای بسط انتخاب می شوند. اسناد با هم ترکیب می شوند و کلمات سند ترکیبی با تابع رتبه بندی TF-IDF رتبه بندی شده و رتبه بالاترین کلمات برای بسط انتخاب می شوند.
- روش بازخورد ارتباطی کامل (TrueRF<sup>11</sup>): در این روش اسناد مرتبط از بین  $|R|$  سند بازیابی شده اولیه، برای بسط انتخاب می شوند. اسناد با هم ترکیب شده و کلمات سند ترکیب شده با تابع رتبه بندی TF-IDF رتبه بندی می شوند و رتبه بالاترین کلمات برای بسط انتخاب می شوند. این روش به عنوان حد بالا در ارزیابی در نظر گرفته می شود چرا که تمام اسناد مرتبط در  $|R|$  سند بازیابی شده برای بسط انتخاب می شوند.
- روش ارائه شده در مقاله، خوشه بندی اسناد شبه بازخورد با شباهت کسینوسی (CPRF).

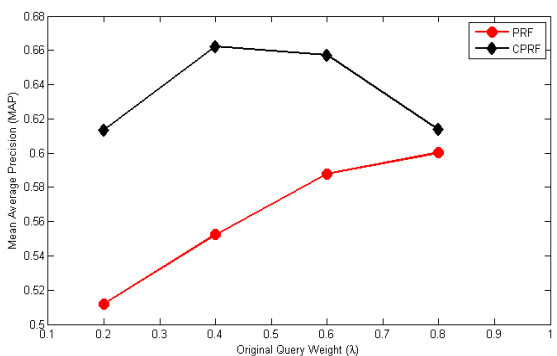
جدول ۲: یادگیری پارامترها

روش بسط پرس و جو	پارامتر وزن پرس و جوی اولیه، $\lambda$	اسناد بازخورد، $ R $	تعداد کلمات بسط، $e$	نسبت تعداد خوشه‌های بازخورد، $I/C$	نسبت اسناد انتخاب شده، $I/M$
PRF	۰/۸	۲۰	۱۰	-	-
CPRF	۰/۶	۳۰	۱۰	۰/۵	۰/۵

اختصاص وزن بیشتر به کلمات بسط، بازیابی را به بازیابی با کلمات بسط سوق می‌دهد و منجر به تشکیل پرس و جوی جدید و نادیده گرفتن پرس و جوی اولیه که منظور اصلی کاربر را نشان می‌دهد، می‌شود. با توجه به این که 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 پرس و جوی اصلی افزایش می‌یابد تا اثر کلمات بسط نامناسب را کاهش دهد.



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



شکل ۳: کارایی (MAP) نسبت به وزن پرس و جوی اولیه ( $\lambda$ ) برای PRF و CPRF

به‌طور کلی افزودن کلمات مرتبط به پرس و جو، کارایی را افزایش می‌دهد. تغییرات کارایی نسبت به تعداد کلمات بسط اضافه شده به پرس و جو در شکل ۴ نشان داده شده است. با قرار دادن مقادیر یادگیری شده 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 به دست آورده است. با توجه به شکل ۴

#### ۲-۲-۴- نتایج آزمایش‌ها

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

با توجه به جدول ۳ روش بسط پرس و جو 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 این که متوسط تعداد اسناد مرتبط به هر پرس و جو در مجموعه داده، ۲۳ سند مرتبط هست، در تعداد اسناد بازخورد بیشتر از ۲۰، کارایی (متوسط میانگین دقت) کاهش می‌یابد. کاهش کارایی به دلیل وجود اسناد خطا بین اسناد انتخاب شده برای بسط است که کلمات خطا را به پرس و جو اضافه می‌کند. سایر پارامترهای این دو روش در مقادیر یادگیری شده قرار داده شده‌اند. همان‌طور که در شکل ۲ مشخص است هر دو روش PRF و CPRF در تعداد ۲۰ سند شبه‌بازخورد بیشترین کارایی را به دست آورده‌اند.

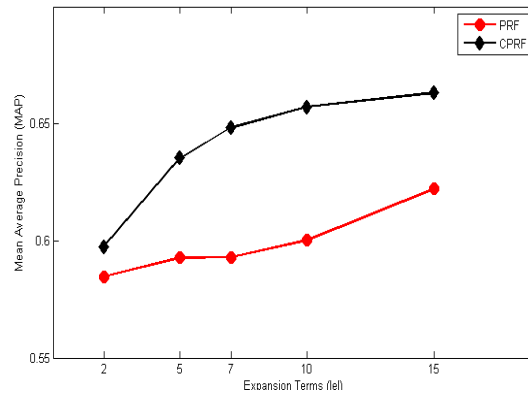
شکل ۳ تأثیر وزن پرس و جوی اصلی نسبت به کلمات بسط را نشان می‌دهد. با قرار دادن سایر پارامترها در مقادیر یادگیری شده، کارایی برای مقادیر مختلف  $\lambda$  به صورت شکل ۳ به دست می‌آید. با توجه به شکل ۲ اختصاص وزن ناچیز به پرس و جوی اصلی و یا خیلی زیاد به پرس و جوی اصلی و کلمات بسط، کارایی را کاهش می‌دهد. وزن ۰/۸ برای PRF و ۰/۴ در روش CPRF بیشترین کارایی را به دست 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 پرس و جوی اولیه تمایل خواهد داشت. درحالی که

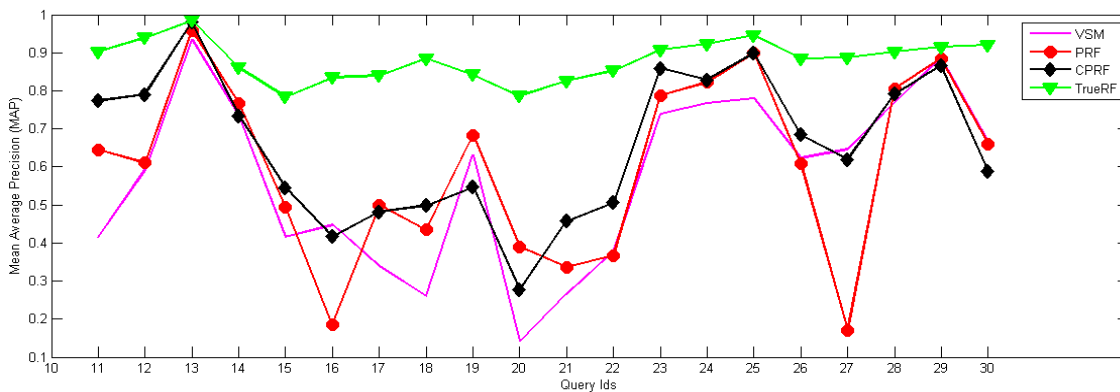
کارایی (میانگین دقت<sup>۱۲</sup>) روی تک به تک پرس و جوها ۱۱ تا ۳۰ از مجموعه داده در شکل ۵ آورده شده است. همه روشها در مقادیر یادگیری شده پارامترها آزمایش شده اند. روش TrueRF از اسناد مرتبط در ۳۰ سند رتبه بالا برای بسط استفاده می کند. روش CPRF هم از ۳۰ سند رتبه بالا برای خوشه بندی استفاده می کند و سعی در استفاده از اسناد مرتبط برای بسط را دارد. تغییرات کارایی در پرس و جوی شماره ۱۳ و ۲۳ نشان می دهد که اسناد مرتبط مشخص شده برای این پرس و جوها کیفیت لازم برای بسط پرس و جو را ندارند. در سایر پرس و جوها می توان 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 تک تک پرس و جوها نشان می دهد. همان طور که مشخص است روش CPRF نسبت به PRF تعداد بیشتری از پرس و جوها را بهبود می دهد. همچنین PRF در ۲ مورد ۵۰ تا ۷۵ درصد کارایی را کاهش داده است در حالی که CPRF برای هیچ پرس و جویی تا این مقدار کارایی را کاهش نداده است.

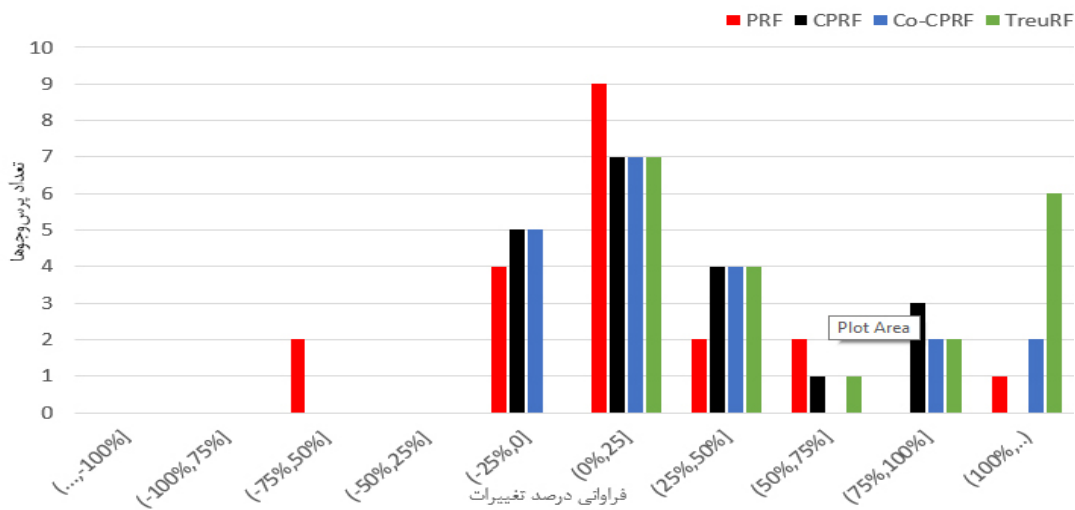
افزودن کلمات بسط تا ۵ کلمه کارایی را به طرز سریع تری افزایش می دهد و این نشان می دهد که پرس و جوها کامل نیستند و افزودن تعداد حداقلی از کلمات بسط، می تواند منظور دقیق تری از نیاز اطلاعاتی کاربر را نشان دهند.



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



شکل ۵: کارایی (MAP) برای پرس و جوهای شماره ۱۱ تا ۳۰ برای VSM, PRF, CPRF و TrueRF



شکل ۶: مقاومت پذیری CPRF نسبت به PRF و TrueRF

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

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

و روش‌های معنایی برای استخراج زمینه و مفهوم اسناد و پرس‌وجو استفاده کرد تا خوشه‌های بهتری را به دست آورد.

#### مراجع

- [1] R. Krovetz, "Homonymy and polysemy in information retrieval," *Proceedings of the 35<sup>th</sup> Annual Meeting of the Association for Computational Linguistics and 8<sup>th</sup> 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 31<sup>st</sup> 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 19<sup>th</sup> 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 24<sup>th</sup> 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.

#### 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 انتخاب شده 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 باز یابی شده، برای بسطشان انتخاب شده است. با توجه به مقادیر تعداد اسناد انتخاب شده در جدول، روش خوشه‌بندی اسناد شبه‌بازخورد، تعدادی از اسناد نامرتبط را از فرآیند بسط خارج می‌سازد.

#### ۵- نتیجه‌گیری

با توجه به نتایج آزمایش‌ها روی مجموعه داده MED، روش خوشه‌بندی اسناد شبه‌بازخورد (CPRF) نسبت به روش ارتباط شبه‌بازخورد (PRF)، کارایی بهتری را در باز یابی اسناد پزشکی دارد. کارایی روش CPRF نسبت به ارتباط شبه‌بازخورد (PRF) ۹/۵ درصد بیشتر است. روش ارائه شده (CPRF) به دلیل 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



- 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.
- [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. Glasgow, (2014/03), *Medline collection*, Available online at: [http://ir.dcs.gla.ac.uk/resources/test\\_collections/med/](http://ir.dcs.gla.ac.uk/resources/test_collections/med/).
- [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*

### زیر نویس ها

<sup>7</sup> K-Nearest Neighbor

<sup>8</sup> Term Frequency-Inverse Document Frequency

<sup>9</sup> Relevance Feedback

<sup>10</sup> Language Model

<sup>11</sup> Clustering-by-Direction

<sup>12</sup> True Relevance Feedback

<sup>13</sup> Average Precision

<sup>1</sup> Pseudo Relevance Feedback

<sup>2</sup> Lavrenko's Relevance Model

<sup>3</sup> Mean Average Precision

<sup>4</sup> Jacard's Coefficient

<sup>5</sup> Dice Coefficient

<sup>6</sup> KL-Divergence