Archive of SID

دو فصلنامه علمی

فناوری اطلاعات و ارتباطات ایران

سال دوازدهم، شماره‌های ۴۵ و ۴۶، پاییز و زمستان۱۳۹۹

صص: ۷۱-۹۴

ارائه مدلی برای بازیابی اطلاعات متنی با استفاده از اعداد فاصله‌ای

فرزاد قهرمانی* **هومان تحیری****

* دانشجوی دکتری بخش مهندسی و علوم کامپیوتر و فناوری اطلاعات، دانشکده مهندسی برق و کامپیوتر، دانشگاه شیراز

** استادیار بخش مهندسی و علوم کامپیوتر و فناوری اطلاعات، دانشکده برق مهندسی و کامپیوتر، دانشگاه شیراز

تاریخ دریافت: ۱۳۹۹/۰۶/۰۷ تاریخ پذیرش:۱۳۹۹/۱۰/۰۲

نوع مقاله: پژوهشی

چکیده

با گسترش و توسعه وب و افزایش محتوای آنلاین، اهمیت سیستم‌های بازیابی اطلاعات که بتوانند با دقت بالاتری به نیازهای اطلاعاتی کاربران پاسخ دهند، بیشتر از پیش مشخص است. یک بخش مهم در طراحی هر سیستم بازیابی اطلاعات، انتخاب روشی مناسب برای مدل کردن آن سیستم است که در این راستا تعیین روش وزن‌دهی به لغات جهت بیان میزان اهمیت آنها در اسناد و پرس‌وجوها، نقش به سزائی دارد. روش‌های مختلفی در خصوص چگونگی وزن‌دهی به لغات ارائه شده که غالباً یک وزن عددی را تخصیص می‌دهند اما نمی‌توان با قطعیت گفت که بهترین روش وزن‌دهی کدام است. با توجه به ابهام و عدم قطعیتی که در این زمینه وجود دارد، در این مقاله مدلی ارائه شده که به جای استفاده از یک مقدار وزنی، با استفاده از وزن‌های بدست آمده از تعدادی روش وزن‌دهی پایه که به دقت انتخاب شده‌اند، برای هر لغت بازه‌ای از وزن‌ها را به عنوان یک وزن فاصله‌ای محاسبه می‌کند. در این مدل با انجام تجمیع مناسب، میزان ارتباط هر سند با پرس‌وجوی ورودی نیز به صورت یک وزن فاصله‌ای تعیین شده و برحسب آنها می‌توان با استفاده از یکی از سه روش پیشنهادی، اسناد را رتبه‌بندی کرد. در آزمایش‌های انجام شده بر روی مجموعه داده‌های معتبر Cranfield و Medline، اثرات نرمال‌سازی طول بردار وزن‌های پایه، استفاده از مؤلفه‌های مختلف در فاکتور فرکانس لغت و فاکتور فرکانس مجموعه مورد مطالعه و بحث قرار گرفته است و مشخص شد که انتخاب مجموعه‌ای مناسب از روش‌های وزن‌دهی پایه برای اعمال روش پیشنهادی، به همراه استفاده از روش رتبه‌بندی مناسب، تأثیر به سزائی در بهبود بازدهی سیستم خواهد داشت. با انتخاب‌های مناسب، برای دو مجموعه داده مذکور به ترتیب MAP با مقادیر ۰,۴۳۳۲۳ و ۰,۵۴۵۸۰ بدست آمد. این نتایج نشان داد که روش پیشنهادی نه تنها باعث بهبود نسبت به هر یک از روش‌های وزن‌دهی پایه می‌شود، بلکه در مقایسه با چند روش وزن‌دهی پیچیده اخیر نیز بهتر عمل می‌کند.

واژگان کلیدی: بازیابی اطلاعات متنی، رتبه‌بندی اسناد، وزن‌دهی لغات، اعداد فاصله‌ای، وزن فاصله‌ای

۱. مقدمه

توسعه روزافزون وب به همراه قابلیت دسترسی به حجم زیادی از محتوای آنلاین، باعث شده که تحقیقات زیادی در زمینه‌ی موتورهای جستجو[۱] یا سیستم‌های بازیابی اطلاعات(IRS[۲]) صورت گیرد. خصوصاً هدف یک موتور جستجوی وب، بازیابی صفحاتی از وب است که به نیازهای اطلاعاتی کاربر که برحسب تعدادی کلمه

نویسنده مسئول: هومان تحیری tahayori@shirazu.ac.ir

[۱] Search engines

[۲] Information Retrieval Systems

www.SID.ir

Archive of SID

کلیدی در یک پرس‌وجو[1] بیان شده، مرتبط باشند [۱]. در واقع هدف یک سیستم IR این است که از بین اطلاعات جمع‌آوری شده، اقلام مرتبط[2] با پرس‌وجوی کاربر را استخراج کرده و اقلام ناخواسته و غیرمرتبط را فیلتر کرده و کنار بگذارد. در این راستا، سیستم IR با نمایش، ذخیره سازی و دسترسی به اسنادی[3] سر و کار دارد که می‌توانند شامل متون، تصاویر، انیمیشن، صوت، اقلام چندرسانه‌ای، صفحات وب، توئیت، وبلاگ و دیگر قطعات اطلاعاتی باشند [۲]، هرچند در این مقاله اسنادی که به صورت متنی و نوشتاری هستند، مورد توجه می‌باشند.

مساله اصلی در تمام سیستم های IR، تطبیق دادن اسناد با پرس‌وجو و پیش‌بینی این موضوع است که کاربر کدام اسناد را مرتبط با خواست خود خواهد دانست [۳]. در نمایش اسناد و پرس‌وجوها غالباً با نگاشتی از لغات به آنها کار داریم، که طبق [۲] لغات یا کلمات کلیدی مجموعه را به ترتیب به "اسناد" و "پرس‌وجوها" نگاشت می‌کنند و این غالباً با نسبت دادن وزن به این لغات صورت می‌گیرد که بیانگر میزان اهمیت آنها در سند یا پرس‌وجوی مربوطه است.

در طول عمری که از بازیابی اطلاعات می‌گذرد، روش‌های مختلفی در خصوص چگونگی وزن‌دهی به لغات در اسناد و پرس‌وجوها ارائه شده است، از جمله [۴] تا [۹]. علاوه بر IR، وزن‌دهی به لغات در حوزه‌های دیگری مانند طبقه‌بندی متون[4] [۱۰] تا [۱۳] و تحلیل احساسات[5] [۱۴]، نیز کاربرد داشته و نقش مهمی بازی می‌کند. اهمیت وزن‌دهی مناسب، و تأثیر زیاد آن در بهبود نتایج نهایی در تمامی این مطالعات مشهود بوده است [۶]، [۷]، و [۱۲].

یادگیری رتبه‌بندی[6] زمینه دیگری از تحقیقات است که یادگیری ماشین، بازیابی اطلاعات، و پردازش زبان طبیعی را با یکدیگر درمی‌آمیزد. یادگیری رتبه‌بندی در معنای عام و گسترده، اشاره به استفاده از هر روش یادگیری ماشین برای رتبه‌بندی دارد. در معنای خاص و محدود، یادگیری رتبه‌بندی به تکنیک‌های یادگیری ماشین برای ساخت مدل رتبه‌بندی، در ایجاد رتبه‌بندی و تجمیع رتبه‌بندی اشاره دارد [۱۵]. استفاده از یادگیری ماشین در IR و طبقه‌بندی متون و غیره، فقط محدود به رتبه‌بندی نبوده و در مرحله وزن‌دهی به لغات نیز کاربرد دارد. بطورکلی دو نوع اصلی وزن‌دهی به لغات وجود دارد. یکی وزن‌دهی لغت معنایی، که از معانی دسته‌ها و لغات موجود در پایگاه‌های دانشی مانند WordNet بهره می‌برد؛ دوم وزن‌دهی لغت آماری، که به چگونگی ظهور یک لغت در یک سند یا گروهی از اسناد از نقطه‌نظر آماری مرتبط است و خود به دو دسته با ناظر و بدون ناظر تقسیم می‌شود. روش‌های وزن‌دهی لغت با ناظر که از اطلاعات عضویت کلاس اسناد آموزشی در گروه‌های از قبل تعریف‌شده استفاده می‌کنند، عمدتاً از تکنیک‌های یادگیری ماشین استفاده می‌کنند [۱۱]. در این تحقیق روش های رتبه‌بندی و روش‌های وزن‌دهی آماری بدون یادگیری مدنظر می‌باشند.

در برخی از روش‌های بازیابی اطلاعات متنی، کاربر به طریقی میزان اهمیت لغات مورد جستجو را وارد می‌کند که نهایتاً به عنوان یک وزن عددی محاسبه و برای لغات منظور می‌شود؛ در برخی دیگر کاربر فقط پرس‌وجوی خود را در قالب یک متن یا چند لغت وارد کرده و سیستم IR به طریقی مشابه با وزنی که برای لغات اسناد محاسبه می‌کند، وزنی عددی نیز برای لغات پرس‌وجو محاسبه می نماید. اگر به وزن‌های تخصیص داده شده به لغات پرس‌وجوها و اسناد طبق روش‌های مختلف وزن‌دهی توجه شود، مشاهده می‌شود که این وزن ها برای یک لغت مقادیر مختلفی را در بر گرفته که ممکن است نزدیک به هم بوده یا از هم دور باشند، و سؤالی که پیش می‌آید این است که واقعاً وزن درنظرگرفته شده توسط کدام روش مناسب‌تر بوده و میزان اهمیت آن لغت را بهتر نشان می‌دهد؟ همچنین آیا کاربر از اطلاعاتی/مقادیری که به عنوان وزن و میزان اهمیت وارد کرده مطمئن بوده است؟ با توجه به نداشتن پاسخی قطعی در خصوص سؤالات مذکور، در این مقاله به جای استفاده از یک مقدار وزنی، به طریقی ساده و به صورت آفلاین، برای هر لغت بازه‌ای از وزن‌ها به عنوان یک وزن فاصله‌ای[7] در نظر گرفته می شود که بدون افزایش پیچیدگی می‌تواند منجر به افزایش دقت شود. در واقع استفاده از یک بازه یا فاصله به عنوان وزن لغات پرس‌وجو و اسناد، عدم قطعیت و عدم اطمینان موجود در انتساب وزن عددی به لغات را تعدیل ساخته و می‌تواند به محاسبه میزان ارتباط دقیق‌تر بین اسناد و پرس‌وجوها کمک کند.

روش پیشنهادی در این مقاله برای تخصیص دادن وزن فاصله‌ای به لغات، استفاده از وزن‌های محاسبه شده توسط تعدادی روش وزن‌دهی مختلف می‌باشد. بدین منظور ابتدا تعدادی از روش‌های وزن‌دهی موجود را مورد مطالعه قرار داده و با درنظر گرفتن ویژگی های بکار رفته در آنها، زیرمجموعه‌ای از آنها به عنوان روش‌های وزن‌دهی پایه[8] انتخاب می‌شوند؛ سپس بر اساس وزن‌های تعیین شده برای هر لغت توسط این روش های پایه و انجام پردازش‌های لازم، یک وزن فاصله‌ای برای آن لغت تعیین می‌شود. با توجه به لغات درون هر پرس‌وجو و با انجام تجمیع مناسب بر روی وزن‌های فاصله‌ای این لغات، میزان ارتباط هر سند با پرس‌وجوی

---

[1] Relevant
[2] Documents
[3] Text Classification
[4] Sentiment Analysis
[5] Learning to rank
[6] Interval Weight
[7] Basic Weighting Methods

www.SID.ir

Archive of SID

ورودی نیز به صورت یک وزن فاصله‌ای تعیین شده و برحسب آنها می‌توان با استفاده از یکی از سه روش پیشنهاد شده در مقاله، اسناد را رتبه‌بندی[۱] کرد. در اینجا ضمن اجرای روش پیشنهادی بر روی دو مجموعه‌داده معتبر، در آزمایش‌ها نشان داده شده که انتخاب زیرمجموعه‌ای مناسب از روش‌های وزن‌دهی پایه برای اعمال روش پیشنهادی، به همراه استفاده از روش رتبه‌بندی مناسب، تأثیر به سزائی در بهبود بازدهی سیستم خواهد داشت؛ لذا در آزمایش‌های صورت گرفته بر روی ۲۴ روش وزن‌دهی پایه، تأثیر سه دسته از معیارها در انتخاب زیرمجموعه‌های وزنی مورد مطالعه قرار گرفته و نهایتاً مناسب‌ترین ترکیب معیارها در انتخاب زیرمجموعه‌های وزنی، در کنار مناسب‌ترین روش رتبه‌بندی تعیین شده است.

در ادامه، در بخش ۲ مروری بر کارهای مرتبط و مبانی نظری خواهد شد. در بخش ۳ مقدمات لازم در خصوص اعداد فاصله‌ای، و در بخش ۴ روش وزن‌دهی پیشنهادی تشریح شده است. آزمایش‌های انجام شده و نتایج بدست‌آمده در بخش ۵ مورد بحث قرار گرفته، و نهایتاً بخش ۶ مربوط به نتیجه‌گیری است.

۲. کارهای مرتبط و مبانی نظری

در وب، رتبه‌بندی اسناد به سه دسته کلی مبتنی بر محتوا[۲]، مبتنی بر پیوند[۳]، و مبتنی بر پیوند-محتوا[۴] تقسیم می‌شود [۱۱]. در رتبه‌بندی مبتنی بر محتوا، برای بازیابی صفحات وب مرتبط با یک پرس‌وجو از محتوای صفحات به عنوان ویژگی‌هایی برای رتبه‌بندی استفاده می‌شود. رتبه‌بندی مبتنی بر پیوند یا اتصال، بر روی اطلاعات ساختاری مانند تعداد پیوندهای اشاره شده به یک صفحه وب یا تعداد لینک‌های خروجی تمرکز دارد که نشان از محبوبیت صفحات دارند. از جمله روش‌های متعارف در این زمینه می‌توان به روش‌های HITS[۵] و Page Rank اشاره کرد؛ در [۱۶] الگوریتم-های مختلف در این زمینه مرور شده و مورد مقایسه قرار گرفته‌اند. روش رتبه‌بندی مبتنی بر پیوند-محتوا، برای یافتن تعادل مناسب بین دو مورد قبل مطرح شده است که مقاله [۱۷] نمونه‌ای از این روش است که با استفاده از ویژگی‌های نسبتاً ساده اسناد وب، رتبه‌بندی مناسبی را با استفاده از معماری برنامه‌نویسی ژنتیکی ارائه می‌دهد. تحقیق پیش رو جزء دسته مبتنی بر محتوا بوده و بر روی روش‌های وزن‌دهی لغات در بازیابی اطلاعات متنی تمرکز دارد؛ به همین دلیل، در این بخش با تکیه بر روش های وزن‌دهی، به معرفی مدل‌های مرسوم IR، خصوصیات آنها و تو سعه‌هایی که از جنبه مفاهیم مختلف صورت گرفته‌است، پرداخته می‌شود.

۱،۲ مدل بولی[۶]

مدل بولی از نخستین مدل‌های بازیابی اطلاعات است که حتی می‌توان روش کار نخستین دستگاه مکانیکی [۱۸]، و استفاده‌های اولیه از کامپیوتر برای بازیابی اطلاعات [۱۹] تا [۲۱]، را جزء آن دانست. بطورکلی در مدل بولی اسناد و پرس‌وجوها به صورت مجموعه‌ای از لغات شاخص دیده می‌شوند که به آن "کیسه لغات[۷]" گفته می‌شود. در جستجوی بولی وزن لغات در هر سند یا پرس‌وجو، ۰ یا ۱ است؛ یعنی در صورتی که در سند یا پرس‌وجو هیچ تطابقی از لغت وجود نداشته باشد، درجه تطابق ۰، و در صورتی که بطور کامل مطابقت داشته باشد، درجه تطابق ۱ درنظر گرفته می‌شود. مدل بولی در بازیابی اطلاعات مبتنی بر محاسبات منطق بولی است. با یک OR بولی، یک سند برای برآورده کردن یک پرس‌وجو باید یکی از لغات پرس‌وجو را داشته باشد. در صورتیکه برای پرس‌وجوی AND بولی، سند باید شامل تمامی لغات پرس‌وجو باشد تا پرس‌وجو برآورده شده و ۱ را برگرداند. این مدل به دلیل سادگی و درک مستقیم آن، به طور گسترده‌ای در سیستم‌های اولیه پذیرفته و استفاده شد. مشکل این مدل آن است که تطابق‌ها رتبه‌بندی نمی‌شوند و مشخص نمی‌کند کدام سند بیشترین رخداد از لغت را دارد. برای OR بولی، هیچ ارجحیتی بین لغات پرس‌وجو وجود ندارد و همگی وزن یکسانی دارند. بنابراین اسناد با تطابق برابر در لیست برگشتی ممکن است حاوی لغات متفاوتی باشند [۲].

۲،۲ مدل احتمالی[۸]

نخستین بار لان[۹] در سال ۱۹۵۷ پیشنهاد نسبت دادن امتیازی به اسناد بر اساس روش احتمالاتی و اهمیت کلمات کلیدی در اسناد را داد بطوریکه این امتیاز نشان‌دهنده میزان ارتباط اسناد با پرس‌وجو باشد [۲۲]. او همچنین در سال ۱۹۵۸ پیشنهاد استفاده از فرکانس رویداد کلمه را به عنوان اندازه‌گیری مفیدی از ارزش کلمه، مطرح کرد که بعداً به عنوان وزن فرکانس لغت (TF[۱۰]) شناخته شد [۴]. مؤثر بودن این روش نسبت به جستجوی بولی در آزمایش‌های زیادی نشان داده شده است [۲۳]. با توسعه ایده‌های اولیه، در [۲۴] اصولی برای رتبه‌بندی احتمالاتی تعریف شد که مشخص می‌کرد چگونه اسناد بطور بهینه بر اساس معیارهای احتمالاتی و با توجه به

[۱] Ranked
[۲] Content-based ranking
[۳] Hyperlink-based ranking
[۴] Hyperlink-content-based ranking
[۵] Hyperlink Induced Topic Search
[۶] Boolean model
[۷] Bag of terms
[۸] Probability model
[۹] Luhn
[۱۰] Term Frequency

www.SID.ir

Archive of SID

معیارهای ارزیابی تعریف شده، رتبه‌بندی می‌شوند. از آن زمان تحقیقات زیادی انجام شده و مانند مدل‌های دیگر، نسخه‌های متعددی از این مدل نیز ارائه شده است که بخشی از آنها در [۲۵] آمده است.

نمایش‌های دیگری از مدل احتمالاتی وجود دارد که شامل مدل‌های $BM_{۱}$، $BM_{۱۱}$، $BM_{۱۵}$، $BM_{۲۵}$ هستند. این مدل‌ها شبیه هم و طرفدار بهترین تطابق[۱] هستند که $BM_{۲۵}$ معمولترین آنها بوده و به دلیل نتایج خوبی که داشته، در تحقیقات زیادی به عنوان مبنای مقایسه[۲] استفاده شده است [۷] و [۲۶] تا [۲۸]. از این روش که با عنوان Okapi-$BM_{۲۵}$ نیز نام برده می‌شود، انواع مختلفی ارائه شده است که در [۲۹] با انجام یک مطالعه تکرارپذیری[۳] بر روی هشت نوع مختلف آن نشان داده است که تفاوت معنی‌داری بین آنها وجود ندارد.

مدل‌سازی زبانی سند[۴] یک رویکرد مهم دیگر در روش‌های آماری برای بازیابی اطلاعات است [۳۰] و [۳۱]. این نوع روش‌ها ذاتاً مولد[۵] هستند، تا آنجا که هدف آنها تخمین مدلی از فرایند تولید سند است. در بازیابی اطلاعات، زمانی که مدل‌های سند تخمین زده شدند، اسناد می‌توانند بر اساس احتمال مدل سند شان در تولید پرس‌وجو، رتبه‌بندی شوند [۳۲]. مدل‌های زبانی سندی که از طریق یک فرآیند Pólya چندمتغیره، یک ویژگی خودتقویت‌کننده[۶] را ارائه می‌دهند، به طور قابل توجهی اثربخشی بازیابی را افزایش می‌دهند [۳۳]. کامینز [۳۲] این روش را بیشتر بکار گرفته و فرآیند کلی‌تری برای مدل‌سازی سند آماری ارائه داده است. او با ارائه روش‌های جدیدی در تخمین پارامترهای فرآیند Pólya، مدل‌های $GSPUD_{bs_t}$ و $GSPUD_{mc}$ را معرفی کرد که در آنها کارآیی بازیابی در مقایسه با مدل‌های قبلی افزایش یافت.

۳،۲ مدل فضای برداری[۷]

مدل معمول دیگر در IR مدل فضای برداری است. نخستین بار سوئیتزر[۸] [۵] بود که به اسناد و پرس‌وجوها به صورت بردارهایی n بعدی از لغات منحصربفرد مجموعه نگاه کرد. در اینجا به جای سیستم تطابق/عدم تطابق مدل بولی، از وزن لغات برای نشان دادن میزان اهمیت آنها در سند یا پرس‌وجوی مربوطه استفاده می شود. تابع تطابق، معیار فاصله بردار پرس‌وجو از بردار سند است، و بر اساس این میزان فاصله می‌توان یک لیست رتبه‌بندی شده در خروجی ارائه داد [۲]. سالتون در [۳۴] ایده استفاده از کسینوس زاویه بین بردارهای سند و پرس‌وجو را به عنوان معیار شباهت میان آنها مطرح کرد که یکی از معیارهای مهم در این زمینه است [۲۵]. اسپارک جونز در سال ۱۹۷۲، ایده استفاده از فرکانس سند معکوس (IDF[۹]) را مطرح کرد که به لغاتی که در تعداد اسناد کمتری مشاهده می‌شدند وزن بیشتری می‌داد [۳۵]. به دنبال آن ایده ترکیب دو وزن TF و IDF (یعنی TF-IDF[۱۰]) خیلی سریع مطرح شد که از جمله می‌توان به [۳۶] اشاره کرد که حاصلضرب آن دو را به عنوان وزن لغات در نظر گرفت. سالتون و همکاران در سال ۱۹۷۵ یک مدل فضای برداری را برای شاخص‌بندی خودکار ارائه دادند [۳۷].

با توجه به اهمیت انتخاب وزن مناسب برای لغات در بازیابی اطلاعات از اسناد متنی با روش فضای برداری، سالتون و باکلی در [۶] به اهمیت سه فاکتور در وزن‌دهی لغات اشاره کردند که از ضرب آنها در یکدیگر می‌توان وزن هر لغت را تعیین کرد، این فاکتورها عبارتند از: فرکانس لغت (TF)، فرکانس سند معکوس (IDF) (یا فرکانس مجموعه معکوس[۱۱])، و فاکتور نرمال‌سازی طول بردار[۱۲]. آنها برای هر یک از فاکتورها، مؤلفه‌های وزن‌دهی نشان‌داده شده در جدول ۱ را در نظر گرفتند تا با انتخاب یکی از مؤلفه‌ها برای هر فاکتور، و ضرب آنها در یکدیگر بتوان به یک روش وزن‌دهی لغت رسید. بدین‌ترتیب امکان انتخاب ترکیبات متنوعی برای وزن‌دهی لغات اسناد یا پرس‌وجوها وجود داشت که برحسب مؤلفه‌های بکار رفته برای فاکتورها در هر روش، آن روش با یک نام سه‌جزئی[۱۳] معرفی می‌گردد. با توجه به اینکه روش‌های وزن‌دهی درنظرگرفته شده برای لغات اسناد و لغات پرس‌وجو می‌تواند متفاوت باشد، آنها در آزمایش‌های خود، ترکیبات وزنی مختلف را توسط دو سه‌جزئی نشان دادند که این دو به ترتیب روش‌های درنظرگرفته شده برای وزن‌دهی لغات اسناد و وزن‌دهی لغات پرس‌وجو را نشان می‌دادند (مثلاً ترکیب وزنی tfc.nfx به معنی در نظر گرفتن وزن $\frac{tf.log\frac{N}{n}}{\sqrt{\sum_{vector}\left(tf.log\frac{N}{n}\right)^{۲}}}$ برای لغات اسناد و وزن $\left(۰.۵ + ۰.۵\frac{tf}{\max tf}\right).\log\frac{N}{n}$ برای لغات پرس‌وجو می‌باشد). آنها در آزمایش‌های زیادی که بر روی چند مجموعه داده انجام دادند، به مقایسه نتایج استفاده از ترکیبات مختلف پرداخته و ترکیباتی را به عنوان روش‌های برتر وزن‌دهی لغات معرفی کردند.

[۱] Best match
[۲] Baseline
[۳] Reproducibility
[۴] Document Language Modelling
[۵] Generative
[۶] Self-reinforcing
[۷] Vector space model
[۸] Switzer
[۹] Inverse Document Frequency
[۱۰] Term Frequency-Inverse Document Frequency
[۱۱] Inverse Collection Frequency
[۱۲] Vector Length Normalization
[۱۳] Triple

www.SID.ir

Archive of SID

جدول ۱. مؤلفه‌های وزن‌دهی لغت [۶]

مؤلفه فرکانس لغت		
b	۱,۰	وزن دودویی که در صورت حضور لغت در یک بردار، برابر ۱ می‌باشد (فرکانس لغت نادیده گرفته می‌شود)
t	tf	فرکانس لغت خام (تعداد دفعاتی که یک لغت در یک سند یا پرس‌وجو رخ می‌دهد)
n	$۰.۵ + ۰.۵\frac{tf}{\max tf}$	فرکانس لغت نرمال‌شده افزایش‌یافته[۱] (فاکتور tf ، توسط بیشترین مقدار tf در بردار نرمال شده، و مجددا برای قرار گرفتن بین ۰,۵ و ۱ نرمال شده است)
مؤلفه فرکانس مجموعه		
x	۱,۰	بدون تغییر در وزن؛ استفاده از مؤلفه فرکانس لغت اصلی (b، t، یا n)
f	$\log\frac{N}{n}$	ضرب مؤلفه tf اصلی، در یک مؤلفه فرکانس مجموعه معکوس (N تعداد کل اسناد در مجموعه، و n تعداد اسنادی است که یک لغت در آنها مشاهده شده است)
p	$\log\frac{N-n}{n}$	ضرب مؤلفه tf، در یک مؤلفه فرکانس مجموعه معکوس احتمالی
مؤلفه نرمال‌سازی طول بردار		
x	۱,۰	بدون تغییر؛ استفاده از فاکتورهای بدست آمده از فرکانس لغت و فرکانس مجموعه به تنهایی (بدون نرمال‌سازی)
c	$1/\sqrt{\sum_{vector} w_i^2}$	استفاده از نرمال‌سازی کسینوسی؛ تقسیم هر وزن لغت w بر فاکتوری که نشان‌دهنده طول بردار اقلیدسی است

ابراهیم و لا ندا-سیلوا [۸] برای بهبود وزن‌دهی لغات، از روش وزن‌دهی فرکانس لغت همراه با میانگین رویداد لغت (TF-ATO[۲]) استفاده کردند که در آن وزن هر لغت در هر سند، از تقسیم TF آن لغت در سند بر میانگین رویداد لغات موجود در آن سند بدست می‌آید. آنها همچنین از یک روش متمایزکننده (DA[۳]) بر پایه بردار مرکز ثقل اسناد، به منظور حذف وزن‌های کم ارزش از اسناد، استفاده کردند و در آزمایش‌های خود تأثیر مثبت هر یک از این دو ایده جدید وزن‌دهی را در دو حالت با حذف کلمات توقف[۴] و بدون حذف کلمات توقف نشان دادند.

تحقیقات زیاد دیگری نیز در راستای بهبود روش TF-IDF انجام شده است که با درنظر گرفتن فاکتورها و مؤلفه‌هایی دیگر سعی بر این مهم داشته‌اند، از جمله این تحقیقات می‌توان به [۱۰]، [۱۳]، و [۳۸] تا [۴۲] اشاره کرد.

۴،۲ مدل‌های بازیابی فازی

مدل‌های بازیابی فازی یا مدلهای بولی توسعه‌یافته که بر پایه مجموعه‌های فازی تعریف شده‌اند، به منظور رسیدگی به ابهام و عدم دقت کاربران (که در پرس‌وجوهای آنها انعکاس می‌یابد) و ارتباط جزئی بین اسناد و پرس‌وجوها پیشنهاد شدند [۳]. این مدل‌های توسعه‌یافته، وزن‌های لغت و توابع عضویت یک مدل مجموعه فازی را ترکیب کرده و تطابق‌هایی که به دقت یک مدل بولی هستند را برمی‌گردانند. از جمله این مدل‌ها می‌توان به MMM[۵]، Paice، و P-Norm اشاره کرد [۴۳]. توسعه‌های فازی سیستم‌های IR و تعمیم‌های فازی مدل بازیابی بولی در [۴۴] و [۴۵] نیز مرور شده است.

۳. مقدمات

در این بخش برخی از مفاهیم اولیه که در مقاله مورد استفاده قرار گرفته است تعریف خواهد شد.

[۱] Augmented
[۲] Term Frequency with Average Term Occurrence
[۳] Discriminative Approach
[۴] Stop-Words
[۵] Min, Max, and Mixed

www.SID.ir

Archive of SID

۱،۳ اعداد فاصله‌ای[۱]

عدد فاصله‌ای A زیرمجموعه‌ای از اعداد حقیقی است که یک محدوده پیوسته و متصل به هم را در بر می‌گیرد.

$$\boldsymbol{A} = \{\boldsymbol{x} \in \boldsymbol{R},\ \boldsymbol{l_a} \le \boldsymbol{x} \le \boldsymbol{r_a}\} = [\boldsymbol{l_a}, \boldsymbol{r_a}] \qquad (۱)$$

$\boldsymbol{l_a}$ و $\boldsymbol{r_a}$ به ترتیب حدود چپ (پایین) و راست (بالا) این عدد هستند و $\boldsymbol{l_a} \le \boldsymbol{r_a}$ ؛ همانطور که مشاهده می شود تمامی مقادیر بین این دو کرانه متعلق به A می باشند.

بطورکلی اگر A و B دو عدد فاصله‌ای با مقادیر $A = [l_a, r_a]$ و $B = [l_b, r_b]$ باشند، برخی از مفاهیم لازم به صورت زیر تعریف می‌شوند [۴۶]:

- پهنا یا وسعت یک عدد فاصله‌ای:

$$width(A) = r_a - l_a \qquad (۲)$$

- نقطه میانی یا مرکز یک عدد فاصله‌ای:

$$midpoint(A) = \frac{l_a + r_a}{۲} \qquad (۳)$$

- جمع دو عدد فاصله‌ای:

$$A + B = [l_a + l_b\ , r_a + r_b] \qquad (۴)$$

- تفریق دو عدد فاصله‌ای:

$$A - B = [l_a - r_b\ , r_a - l_b] \qquad (۵)$$

- ضرب دو عدد فاصله‌ای:

$$A \times B = [min\ (l_a \times l_b\ , l_a \times r_b\ , r_a \times l_b\ , r_a \times r_b)\ , max\ (l_a \times l_b\ , l_a \times r_b\ , r_a \times l_b\ , r_a \times r_b)] \qquad (۶)$$

- تقسیم دو عدد فاصله‌ای:

$$\frac{A}{B} = A \times \frac{۱}{B} \quad , ۰ \notin B \text{ و } \frac{۱}{B} = \left[\frac{۱}{r_b}, \frac{۱}{l_b}\right] \qquad (۷)$$

توجه شود که بجز پهنا و نقطه میانی یک عدد فاصله‌ای که حاصل آنها یک عدد حقیقی است، در بقیه موارد حاصل هر یک از عملیات نیز یک عدد فاصله‌ای خواهد بود.

۲،۳ میانگین وزنی فاصله‌ای (IWA[۲])

این نوع میانگین‌گیری شبیه به میانگین‌گیری وزنی معمولی است که در آن برای داده‌ها و وزن آنها، به جای اعداد حقیقی از اعداد فاصله‌ای استفاده می‌شود و نتیجه میانگین وزنی نهایی آن نیز نه یک عدد، بلکه یک بازه عددی (عدد فاصله‌ای) خواهد بود. فرض کنید مجموعه‌های X و W هر یک شامل N عدد فاصله‌ای باشند که به ترتیب داده‌ها و وزن‌های متناظر هر یک را در بر می‌گیرند [۴۷]:

$$x_i = \left[l_{x_i}, r_{x_i}\right],\ x_i \in X \quad i = ۱,۲, \dots, N \qquad (۸)$$

$$w_i = \left[l_{w_i}, r_{w_i}\right], w_i \in W \quad i = ۱,۲, \dots, N$$

در اینصورت میانگین وزنی فاصله‌ای x_i های مختلف، عدد فاصله‌ای Y بوده و به صورت زیر محاسبه می‌شود:

$$Y = \frac{\sum_{i=۱}^{N} x_i \times w_i}{\sum_{i=۱}^{N} w_i} = \left[l_y, r_y\right] \qquad (۹)$$

در رابطه (۹) نیاز به انجام عملیات ضرب، جمع، و تقسیم بین اعداد فاصله‌ای می‌باشد که جهت انجام آنها از روابط (۳) تا (۷) استفاده می‌شود.

۴. روش پیشنهادی وزن‌دهی

در این مقاله که بر پایه مدل فضای برداری عمل می‌شود، روش پیشنهادی کلی این است که بر اساس وزن‌های مختلفی که از چند روش مختلف وزن‌دهی برای یک لغت بدست آمده‌اند، یک وزن فاصله‌ای (بازه‌ای) برای آن لغت محاسبه شود تا بدین‌ترتیب هر یک از وزن‌های اولیه بتوانند به نحوی استفاده شده و اثرگذار باشند. در اینجا ماژول تطابق نیز در محاسبات خود بر اساس اصول اعداد فاصله‌ای عمل کرده و نهایتاً میزان ارتباط اسناد مختلف با هر پرس‌وجو را در قالب یک سری اعداد فاصله‌ای ارائه می‌دهد که بر مبنای آنها اسناد رتبه‌بندی می‌شوند. مراحل روش پیشنهادی و جزئیات بیشتر در ادامه شرح داده شده است.

۱،۴ پیش‌پردازش داده‌ها

جهت پالایش اولیه مستندات، از معمول‌ترین عملیات در مرحله پیش‌پردازش استفادمی گردد: حذف کلمات توقف[۳](کلماتی که بی‌معنی بوده و تاثیری در برآورد میزان ارتباط اسناد و پرس‌وجوها ندارند)، انجام ریشه‌یابی[۴](یکسان گرفتن کلماتی که ریشه مشترکی دارند)، و نادیده گرفتن بزرگ و کوچک بودن حروف[۵].

۲،۴ انتخاب روش‌های وزن‌دهی پایه[۶]

همانطور که اشاره شد روش‌های وزن‌دهی مختلفی در تحقیقات پیشین معرفی شده‌اند. در این مرحله بایستی تعدادی از این روش‌ها را به عنوان روش‌های وزن‌دهی پایه انتخاب کرد تا از آنها جهت ساخت وزن‌های فاصله‌ای لغات فرهنگ بدست‌آمده، استفاده نمود. با

[۱] Interval Numbers
[۲] Interval Weighted Average
[۳] Stop words
[۴] Stemming
[۵] Case sensitivity
[۶] Basic weighting methods

www.SID.ir

Archive of SID

توجه به اینکه روش‌های وزن‌دهی مختلف مطرح شده در [۶] که در بخش قبل شـــرح داده شـــدند از تنوع زیادی برخوردار بوده و پایه خیلی از روشـــهای وزن‌دهی در تحقیقات بعدی بوده‌اند، در این تحقیق نیز از آنها به عنوان روش‌های وزن‌دهی پایه اســـتفاده شـــده است. البته در اینجا به مؤلفه‌های فاکتور فرکانس لغت در جدول ۱، مؤلفه دیگری تحت عنوان مؤلفه فرکانس لغت نرمال‌شـده (معمولی، نه افزایش‌یافته) با فرمول $\frac{tf}{\max tf}$ (و با نماد z) اضافه شده تا توسط آن بتوان مقدار فاکتور TF لغات هر ســـند را با توجه به ماکزیمم TF لغات آن ســـند بین ۰ و ۱ محدود کرد. در این صـــورت امکان ایجاد ۲۴ ترکیب متنوع به عنوان روش‌های وزن‌دهی پایه برای لغات اسناد یا پرس‌وجوها بوجود خواهد آمد که در جدول ۳ قابل مشاهده هستند.

واضح است که انتخاب روش‌های وزن‌دهی اولیه مناسب، از اهمیت بالایی در بهبود نتایج نهایی برخوردار بوده و ممکن اســـت در نظر گرفتن تمامی این ۲۴ روش وزن‌دهی، جهت انجام بقیه مراحل روی آنها لازم نباشد و انتخاب زیرمجموعه‌ای مناسب از آنها کفایت کند. در بخش ۵ آزمایشـــهای مختلفی در این خصـــوص انجام شـــده و انتخاب زیرمجموعه‌های مختلفی مورد بحث و بررســـی قرار گرفته است.

۳،۴ محاسبه وزن‌های پایه

در این مرحله با اســـتفاده از روش‌های مختلف وزن‌دهی پایه که در مرحله قبل انتخاب شـــده‌اند، برای هر لغت تعدادی وزن پایه و اولیه تعیین می شود. بدین ترتیب اگر پس از انجام پیش‌پردازش‌ها، تعداد n_t لغت منحصربفرد در فرهنگ لغات وجود داشته باشد، می‌توان هر ســـند d را در روش وزن‌دهی m به صـــورت برداری از وزن لغات به فرم زیر نمایش داد:

$$D_{d,m} = \langle wd_{d,m,1}, wd_{d,m,2}, \dots, wd_{d,m,n_t} \rangle \quad (10)$$
$$d \in \{1,2,\dots,n_d\} \quad m \in \{1,2,\dots,n_m\}$$

که در آن n_d، تعداد اســـناد در مجموعه، n_m، تعداد روش‌های مختلف درنظر گرفته شـــده جهت بدســـت آوردن وزن‌های پایه، و $wd_{d,m,t}$، وزن پایه مربوط به لغت t در ســـند d طبق روش وزن‌دهی پایه m می‌باشـــد. همچنین به طریق مشـــابه می‌توان هر پرس‌وجو را در روش وزن‌دهی m به صورت برداری از وزن لغات به فرم زیر نمایش داد:

$$Q_m = \langle wq_{m,1}, wq_{m,2}, \dots, wq_{m,n_t} \rangle \quad (11)$$
$$m \in \{1,2,\dots,n_m\}$$

که در آن $wq_{m,t}$، وزن مربوط به لغت t در آن پرس‌وجو، طبق روش وزن‌دهی پایه m می‌باشد.

۴،۴ نرمال‌سازی وزن‌های پایه

وزن‌های مختلفی که در مرحله قبل برای هر لغت بدســـت آمد، ممکن است بسته به روش‌های وزن‌دهی انتخابی، مقادیر مختلفی را در بر بگیرند. مثلاً ممکن است وزن‌های بدست آمده برای لغات طی یک روش، تماماً در بازه [۰،۱]، یا بازه [۰،۵،۱] قرار گرفته، و یا مقادیری بیش از ۱، یا حتی منفی را نیز در بر گیرند. بنابراین در این مرحله باید تمامی وزن‌های پایه مربوط به لغات مختلف سند d را که درروش وزن‌دهی m بدست آمده‌اند، در بازه [۰،۱] نرمال‌سازی کرد تا بدین‌ترتیب اثر پراکندگی وزن‌های روش‌های مختلف برای هر لغت تعدیل گردد. نرمال‌سازی می‌تواند بر اساس رابطه زیر انجام شود [۴۸]:

$$nwd_{d,m,t} = \frac{wd_{d,m,t} - wd_{d,m}^{min}}{wd_{d,m}^{max} - wd_{d,m}^{min}} \quad t \in \{1,2,\dots,n_t\} \quad (12)$$

جائیکه $wd_{d,m}^{min}$ و $wd_{d,m}^{max}$ به ترتیب کمترین و بیشترین مقادیر وزنی پایه مجموعه لغات در سند d، طی روش وزن‌دهی m، و $nwd_{d,m,t}$ وزن پایه نرمال‌شده مربوط به لغت t در سند d طبق روش وزن‌دهی پایه m می‌باشند. به طریق مشابه بایستی وزن‌های پایه بدست آمده برای لغات پرس‌وجو را نیز نرمال‌سازی کرد تا $nwq_{m,t}$های مختلف برای بردار پرس‌وجوی Q_m بدست آید. توجه شود که انجام نرمال‌سازی وزن‌ها، به معنی ۱ شدن طول بردار وزن‌ها نمی‌باشد.

۵،۴ محاسبه وزن فاصله‌ای

در این مرحله بایستی به طریقی مناسب با استفاده از وزن‌های پایه نرمال‌شـــده مختلفی که از روش‌های مختلف وزن‌دهی برای هر لغت در هر ســـند یا پرس‌وجو بدســـت آمده، یک وزن فاصـــله‌ای برای آن لغت بدســـت آورد. در این تحقیق این وزن‌های فاصـــله‌ای بر اســـاس انحراف استاندارد وزن‌های مختلف در پیرامون میانگین آنها تشکیل می‌شود و برای لغات اسناد مختلف طبق رابطه زیر بدست می‌آیند:

$$iwd_{d,t} = [lwd_{d,t}, rwd_{d,t}] \quad (13)$$
$$= [avg_{d,t} - std_{d,t}, avg_{d,t} + std_{d,t}]$$

جائیکه $lwd_{d,t}$ و $rwd_{d,t}$ به ترتیب حد چپ و راست بازه درنظر گرفته شـــده برای وزن فاصـــله‌ای، و $avg_{d,t}$ و $std_{d,t}$ به ترتیب میانگین و انحراف اســـتاندارد وزن‌های پایه نرمال‌شـــده مختلفی هستند که برای لغت t در سند d بدست آمده بودند و به صورت زیر محاسبه می‌شوند:

www.SID.ir

Archive of SID

$$avg_{d,t} = \frac{\sum_{m=1}^{n'_m} nwd_{d,m,t}}{n'_m} \quad , n'_m \le n_m \qquad (۱۴)$$

$$std_{d,t} = \sqrt{\frac{\sum_{m=1}^{n'_m} (nwd_{d,m,t} - avg_{d,t})^2}{n'_m}} \qquad (۱۵)$$

همانطور که قبلاً اشاره شد ممکن است در برخی موارد از تمامی روش‌های وزن‌دهی پایه درنظر گرفته شده استفاده نشود و تنها زیرمجموعه‌ای از آنها درنظر گرفته شود؛ در همین خصوص، n'_m تعداد روش‌های وزن‌دهی پایه مورد استفاده می‌باشد. وزن‌های فاصله‌ای لغات پرس‌وجو (iwq_tها) نیز به طریق مشابه محاسبه می‌شود.

۶،۴ رتبه‌بندی اسناد

پس از اینکه در مرحله قبل برای هر یک از لغات اسناد و پرس‌وجوها بازه‌ای از وزن‌ها به عنوان وزن فاصله‌ای تعیین شد، حال سوال این است که جهت رتبه‌بندی اسناد، میزان ارتباط بین پرس‌وجوها و اسناد چگونه باید مشخص شود؟ می بایست توجه شود که کاربر نیازهای اطلاعاتی خود را در قالب پرس‌وجوی ورودی از سیستم می‌خواهد و وزن لغات در آن پرس‌وجو میزان اهمیت هر لغت را تعیین می‌کند؛ لذا هرچه سندی حاوی لغات خواسته شده در پرس‌وجو بوده و خصوصاً لغات مهمتر پرس‌وجو، در آن سند وزن بیشتری داشته باشند، می‌توان گفت که آن سند به پرس‌وجوی خواسته شده مرتبط‌تر است. برای نشان دادن این میزان ارتباط می‌توان از رابطه (۱۶) استفاده کرد. در عمل رابطه (۱۶) مبین محاسبه میانگین وزنی بر روی لغات سند است که هدف آن برآورد یک وزن کلی برای سند در ارتباط با پرس‌وجوی ورودی است. به منظور تأثیر دادن اهمیت لغات پرس‌وجو، وزن این لغات نیز به عنوان ضریب وزنی در محاسبه دخالت داده شده است. در اینجا چون وزن‌های محاسبه شده برای لغات به صورت فاصله‌ای هستند، از روش محاسبه میانگین وزنی فاصله‌ای استفاده می گردد که فرمول کلی آن در رابطه (۹) آورده شد؛ کافی است برای محاسبه میزان ارتباط بین سند D_d و پرس‌وجوی Q، در رابطه (۹) وزن‌های فاصله‌ای لغات سند ($iwd_{d,t}$ها) را به عنوان داده‌های اصلی، و وزن‌های فاصله‌ای لغات پرس‌وجو (iwq_tها) را به عنوان وزن‌های داده‌های اصلی درنظر گرفت. بدین‌ترتیب داریم:

$$Relevancy(D_d, Q) = IWA(D_d, Q) = \frac{\sum_{t=1}^{n_t} (iwd_{d,t} \times iwq_t)}{\sum_{t=1}^{n_t} iwq_t} = [l_d, r_d] \qquad (۱۶)$$
$$d: ۱, ۲, \dots, n_d$$

l_d و r_d به ترتیب حد چپ (پایین) و راست (بالا) عدد فاصله‌ای نتیجه شده می‌باشند. با اینکه ظاهرا سری عملیات صورت و مخرج رابطه (۱۶) باید بر روی کل لغات (n_t) انجام شوند اما در عمل تنها کافی است که به تعداد لغات پرس‌وجو (بعد از مرحله پیش‌پردازش) و بر روی آنها انجام شوند، زیراکه وزن فاصله‌ای لغات غیرموجود در پرس‌وجو صفر بوده و در محاسبات بی‌تأثیر هستند. این موضوع سرعت محاسبات را بالا می‌برد.

در رابطه (۱۶) در واقع مقدار وضعیت بازیابی (RSV) سند D_d به صورت یک عدد فاصله‌ای محاسبه می شود؛ پس از محاسبه RSV اسناد مختلف، برای رتبه‌بندی آنها می‌توان به یکی از سه روش زیر اقدام کرد:

- رتبه‌بندی اسناد بر اساس مقدار چپ (Lrank[۱]): در اینجا بایستی اسناد را بر اساس مقدار چپ (حد پایین) ارتباط‌های بدست آمده با پرس‌وجو (l_dها) به صورت نزولی مرتب کرد تا رتبه آنها بدست آید.
- رتبه‌بندی اسناد بر اساس مقدار راست (Rrank[۲]): در اینجا بایستی اسناد را بر اساس مقدار راست (حد بالای) ارتباط‌های بدست آمده با پرس‌وجو (r_dها) به صورت نزولی مرتب کرد تا رتبه آنها بدست آید.
- رتبه‌بندی اسناد بر اساس مقدار نقطه میانی (Mrank[۳]): در اینجا برای بدست آوردن رتبه اسناد، ابتدا نقطه میانی (مرکز) اعداد فاصله‌ای بدست آمده به عنوان ارتباط اسناد با پرس‌وجو را طبق رابطه (۲) بدست آورده و سپس این مقادیر را به صورت نزولی مرتب کرده تا رتبه اسناد بدست آید.

۵. نتایج و بحث

به دلیل محدودیت سخت‌افزاری در کار بر روی مجموعه داده‌های بزرگ، آزمایش‌ها بر روی دو مجموعه داده محک[۴] Cranfield و Medline انجام شد که موضوع آنها به ترتیب در زمینه علوم مربوط به گاز و هوا[۵] و پزشکی می‌باشد [۲۶]. اطلاعات آماری این مجموعه‌ها در جدول ۲ مشاهده می‌شود [۶].

[۱] Left rank
[۲] Right rank
[۳] Midpoint rank
[۴] Benchmark
[۵] Aerodynamics

www.SID.ir

Archive of SID

جدول ۲. اطلاعات آماری مجموعه داده‌ها (شامل میانگین طول بردار و انحراف استاندارد طول بردار) [۶]

مجموعه داده	تعداد بردارها (اسناد یا پرس‌وجوها)	میانگین طول بردارها (تعداد لغات)	انحراف استاندارد طول بردارها	میانگین فرکانس لغات در بردارها
Cranfield				
documents	۱۳۹۸	۵۳,۱۳	۲۲,۵۳	۱,۵۸
queries	۲۲۵	۹,۱۷	۳,۱۹	۱,۰۴
Medline				
documents	۱۰۳۳	۵۱.۶۰	۲۲.۷۸	۱.۵۴
queries	۳۰	۱۰.۱۰	۶.۰۳	۱.۱۲

پس از انجام پیش‌پردازش‌های اشاره شده در بخش ۴-۱ بر روی هر یک از مجموعه داده‌ها، آزمایش‌های مختلفی در خصوص چگونگی انتخاب زیرمجموعه‌های وزنی پایه برای اسناد و پرس‌وجوها و انجام رتبه‌بندی بر اساس روش‌های پیشنهادی صورت گرفت. کارآیی روش وزن‌دهی پیشنهادی برا ساس داوری‌های ا سناد مرتبط به هر پرس‌وجو (که به صورت دودوئی انجام شده برای هر مجموعه داده موجود است)، و برحسب معیار میانگین متوسط دقت (MAP[۱]) و دقت در نقطه n (P@n) ارزیابی می‌شود.

- معیار میانگین متوسط دقت (MAP): این معیار در قالب یک عدد، مؤثربودن سیستم[۲] را نشان داده و به طور گسترده‌ای مورد استفاده قرار می‌گیرد [۴۹]. متوسط دقت (AP) ترکیبی از دقت و بازخوانی[۳] است که، برای یک پرس‌وجو، بر روی لیست رتبه‌بندی‌شده اسناد، اعمال شده و به صورت زیر محاسبه می‌شود [۷]:

$$AP = \frac{1}{R}\sum_{r=1}^{n_d} P(d_r).Rel(d_r) \quad (۱۷)$$

$$R = \sum_{r=1}^{n_d} Rel(d_r) \quad (۱۸)$$

جائیکه d_r سند در رتبه r، $P(d_r)$ دقت در رتبه r، R تعداد اسناد مرتبط به پرس‌وجو در مجموعه، و $Rel(d_r)$ داوری دودویی (۰ و ۱) انجام شده برای سند رتبه r از نظر داشتن ارتباط با پرس‌وجو است. پس از اینکه AP برای تمامی پرس‌وجوهای یک مجموعه داده بدست آمد، میانگین آنها به عنوان MAP در نظر گرفته خواهد شد، یعنی:

$$MAP = \frac{\sum_{q=1}^{n_q} AP(q)}{n_q} \quad (۱۹)$$

جائیکه n_q تعداد پرس‌وجوها، و AP(q) میانگین دقت مربوط به پرس‌وجوی q است.

- معیار دقت در نقطه n (P@n): این معیار، نسبت تعداد اسناد مرتبط را به کل اسناد در برترین n سند ارائه شده، مشخص می‌کند؛ یعنی [۱۶]:

$$P@n = \frac{\#\ of\ relevant\ documents\ in\ top\ n\ results}{n} \quad (۲۰)$$

در اینجا معیارهای P@۵، P@۱۰، و P@۲۰ مورد ارزیابی قرار می‌گیرند.

در ادامه آزمایش‌های انجام‌شده و تحلیل‌های لازم شرح داده خواهد شد.

۱،۵ روش‌های وزن‌دهی پایه

همانطور که در بخش ۴-۲ اشاره شد، در این تحقیق ۲۴ روش وزن‌دهی پایه درنظر گرفته شد که نهایتاً بایستی زیرمجموعه‌هایی مناسب از آنها جهت اعمال روش‌های پیشنهادی بر روی آنها، انتخاب شوند. در اینجا ابتدا هر یک از این روش‌ها به تنهایی به عنوان روش وزن‌دهی لغات اسناد و پرس‌وجوها انتخاب، و وزن‌های بدست آمده از هر روش طبق رابطه (۱۲) در بازه [۰،۱] نرمال‌سازی شدند. سپس با استفاده از فرمول مر سوم ضرب بردارهای سند و پرس‌وجو در یکدیگر، شباهت و ارتباط اسناد با پرس‌وجوها محاسبه و رتبه‌بندی شدند. نتایج MAP هر دو مجموعه داده برای تمامی ۲۴ روش وزن‌دهی، در جداول ۳ و ۴ مشاهده می‌شود.

اگر به نتایج جدول ۳ دقت شود، نکاتی دیده می‌شود که توجه به آنها می‌تواند به انتخاب زیرمجموعه‌هایی مناسب جهت وزن‌های پایه

[۱] Mean Average Precision
[۲] System effectiveness
[۳] Recall

www.SID.ir

Archive of SID

جدول ۳. روش‌های وزن‌دهی پایه، به همراه مقدار MAP هر روش بر روی مجموعه داده‌های Cranfield و Medline

	روش‌های ۱-۱۲ (بدون نرمال‌سازی طول بردار)												روش‌های ۱۳-۲۴ (همراه با نرمال‌سازی طول بردار)											
شماره روش	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵	۱۶	۱۷	۱۸	۱۹	۲۰	۲۱	۲۲	۲۳	۲۴
نام سه‌جزئی روش	bxx	bfx	bpx	txx	tfx	tpx	nxx	nfx	npx	zxx	zfx	zpx	bxc	bfc	bpc	txc	tfc	tpc	nxc	nfc	npc	zxc	zfc	zpc
Cranfield	۰,۲۷۱۱۹	۰,۳۲۱۲۴	۰,۳۱۴۴۱	۰,۳۴۰۰۴	۰,۳۸۷۱	۰,۳۷۴۱۳	۰,۳۶۵۹۶	۰,۳۷۷۹۹	۰,۳۶۶۰۷	۰,۳۴۰۰۴	۰,۳۸۷۱	۰,۳۷۴۱۳	۰,۳۱۷۷۵	۰,۳۳۸۲۵	۰,۳۳۰۱۶	۰,۳۷۲۲۳	۰,۴۰۳۴۱	۰,۳۹۴۸۵	۰,۳۷۵۲۹	۰,۳۸۸۵۸	۰,۳۷۹۵۶	۰,۳۷۲۳۱	۰,۴۰۳۴۱	۰,۳۹۴۸۵
Medline	۰,۴۲۶۳۳	۰,۴۷۲۹۵	۰,۴۷۲۹۷	۰,۴۴۳۱۹	۰,۵۰۱۸	۰,۵۰۰۸۴	۰,۴۸۱۵۷	۰,۵۱۵۹۸	۰,۵۱۵۲۱	۰,۴۴۳۱۹	۰,۵۰۱۸	۰,۵۰۰۸۸	۰,۴۵۸۳۶	۰,۴۸۲۷۱	۰,۴۸۲۱۴	۰,۴۶۷۳۹	۰,۵۱۵۲۵	۰,۵۱۴۸۶	۰,۴۹۶۱۹	۰,۵۲۳۶۵	۰,۵۲۲۳	۰,۴۶۷۳۹	۰,۵۱۵۲۵	۰,۵۱۴۸۶

جدول ۴. بیشترین، کمترین، و میانگین مقادیر MAP بدست آمده از ۲۴ روش وزن‌دهی پایه

	Cranfield	Medline
بیشترین	۰,۴۰۳۴۱	۰,۵۲۳۶۵
کمترین	۰,۲۷۱۱۹	۰,۴۲۶۳۳
میانگین	۰,۳۶۲۰۹	۰,۴۸۹۰۴

کمک کند؛ از جمله:

- استفاده از روش‌های ۱۳-۲۴، به دلیل نرمال بودن طول بردار وزن ها، در تمامی موارد نسبت به استفاده از روش‌های ۱-۱۲ متناظر بهتر است.
- روش‌های وزن‌دهی ۱ و ۱۳ (bxx و bxc)، به دلیل دادن مقادیر وزنی بولی (۰ و ۱) به لغات، نسبت به سایر روش‌ها نتایج ضعیف‌تری دارند.
- استفاده از مؤلفه پیشنهادی z به جای مؤلفه n در فاکتور اول (یعنی نرمال سازی معمولی TF در وزنهای ۱۰-۱۲ و ۲۲-۲۴ به جای نرمال‌سازی افزایش‌یافته آن در وزنهای ۷-۹ و ۱۹-۲۱)، در مجموعه داده Cranfield باعث نتایج بهتری شده است.
- استفاده از مؤلفه p به جای مؤلفه f در فاکتور دوم (یعنی فرکانس مجموعه معکوس احتمالی به جای فرکانس مجموعه معکوس)، در برخی موارد باعث نتایج ضعیف‌تری می‌شود. علت این موضوع به خاطر این است که در محاسبه مؤلفه p (یعنی $\log\frac{N-n}{n}$)، در مواردی که برای یک لغت داریم: $n > \frac{N}{2}$ ، حاصل لگاریتم و در نتیجه وزن لغت منفی می شود؛ یعنی وزن چنین لغاتی حتی از لغاتی که اصلاً در سند وجود ندارند و وزن صفر گرفته‌اند هم کمتر می‌شود. با این حساب حتی پس از نرمال‌سازی وزن‌های پایه طبق رابطه (۱۲) نیز این وضعیت باقی مانده و وزن این لغات از لغاتی که اصلاً در سند وجود نداشته‌اند کمتر می‌شود؛ بنابراین طبیعی است که در چنین مواردی، روش‌های مذکور ضعیف‌تر عمل کنند.

۲،۵ اعمال روش پیشنهادی بر روی زیرمجموعه‌هایی از وزن‌های پایه

با توجه به اینکه ایده اصلی این تحقیق، تعیین وزن‌های فاصله‌ای برای لغات با اعمال روش پیشنهادی بر روی زیرمجموعه‌ای از وزن‌های پایه است که با روش‌های مختلف وزن‌دهی بدست آمده‌اند، بر اساس مشاهدات جدول ۳ و نکات برگرفته از آن که در بخش ۵-۱ اشاره شد، در انتخاب زیرمجموعه‌های مختلف از ۲۴ روش وزن‌دهی پایه، می‌توان موارد زیر را درنظر گرفت که در جدول ۵ نیز با جزئیات بیشتری آورده شده‌اند:

أ) درنظرگرفتن زیرمجموعه هایی فقط از روش های ۱-۱۲ (A_1)، فقط از روش های ۱۳-۲۴ (A_2)، یا از روش های ۱-۲۴ (A_3)

ب) انتخاب روش های فاقد مؤلفه z (B_1)، روش های فاقد مؤلفه n (B_2)، یا روش‌های شامل هر دو مؤلفه z و n (B_3)

ج) انتخاب زیرمجموعه‌هایی شامل روش‌های ۱ یا ۱۳ (C_1)، فاقد روش‌های ۱ و ۱۳ (C_2)، و علاوه بر کنار گذاشتن روش های ۱ و ۱۳ فاقد مؤلفه p در فاکتور فرکانس مجموعه (C_3)

در واقع برای هر یک از سه دسته از عواملی که ممکن است در انتخاب زیرمجموعه‌های وزنی تأثیرگذار باشند، سه گزینه برای

www.SID.ir

Archive of SID

انتخاب در نظر گرفته و هر کدام با نمادهایی معرفی شـــده‌اند. با در نظر گرفتن این موارد و انتخاب گزینه‌های مختلف، بطورکلی امکان

جدول ۵. نمادهای مورد استفاده جهت معرفی زیرمجموعه‌های وزنی پایه انتخابی

نمادهای مربوط به نرمال‌سازی طول	
A_1	درنظرگرفتن زیرمجموعه‌هایی فقط شامل روش‌های وزن‌دهی پایه بدون نرمال‌سازی طول (روش‌های ۱: bxx تا ۱۲: zpx)
A_2	درنظرگرفتن زیرمجموعه‌هایی فقط شامل روش‌های وزن‌دهی پایه همراه با نرمال‌سازی طول (روش‌های۱۳:bxc تا ۲۴: zpc)
A_3	درنظرگرفتن زیرمجموعه‌هایی بدون محدودیت در این خصوص، یعنی زیرمجموعه‌هایی شامل هم روش‌های وزن‌دهی پایه بدون نرمال‌سازی طول و هم روش‌های وزن‌دهی پایه همراه با نرمال-سازی طول (روش‌های ۱: bxx تا ۲۴: zpc)
نمادهای مربوط به مؤلفه فرکانس لغت انتخابی	
B_1	درنظرگرفتن زیرمجموعه‌هایی فقط شامل روش‌های وزن‌دهی پایه که در آنها از مؤلفه z استفاده نشده (عدم استفاده از روش-های ۱۰: zxx تا ۱۲: zpx و ۲۲: zxc تا ۲۴: zpc)
B_2	درنظرگرفتن زیرمجموعه‌هایی فقط شامل روش‌های وزن‌دهی پایه که در آنها از مؤلفه n استفاده نشده (عدم استفاده از روش-های ۷: nxx تا ۹: npx و ۱۹: nxc تا ۲۱: npc)
B_3	درنظرگرفتن زیرمجموعه‌هایی از روش‌های وزن‌دهی پایه بدون محدودیت در مؤلفه فرکانس لغت مورد استفاده (یعنی در زیرمجموعه‌های وزنی هم از روش‌های شامل مؤلفه z می‌توان استفاده کرد و هم از روش‌های شامل مؤلفه n)
نمادهای مربوط به استفاده از روش بولی و روش‌های وزن‌دهی شامل مؤلفه *p* فرکانس مجموعه	
C_1	درنظرگرفتن زیرمجموعه‌هایی از روش‌های وزن‌دهی پایه بدون محدودیت در این خصوص (یعنی در زیرمجموعه‌های وزنی هم از روش‌های وزن‌دهی بولی استفاده می‌شود و هم از روش‌های وزن‌دهی شامل مؤلفه p)
C_2	درنظرگرفتن زیرمجموعه‌هایی فاقد روش‌های وزن‌دهی پایه بولی (عدم استفاده از روش‌های ۱: bxx و ۱۳: bxc)
C_3	درنظرگرفتن زیرمجموعه‌هایی فاقد روش‌های وزن‌دهی پایه بولی (روش‌های ۱: bxx و ۱۳: bxc) و همچنین فاقد روش-های وزن‌دهی شامل مؤلفه p فرکانس مجموعه (یعنی روش-های ۳: bpx، ۶: tpx، ... ، ۲۴: zpc)

انتخاب ۲۷ زیرمجموعه مختلف از ۲۴ روش پایه وجود دارد که در آزمایش‌ها مورد بررســی و ارزیابی قرار گرفته‌اند. در آزمایش‌ها برای هر یک از این زیرمجموعه ها، پس از اینکه بر اســاس هر یک از روش های وزن‌دهی در زیرمجموعه، وزن های پایه‌ی لغات مختلف اســناد و پرس‌وجوها طبق روابط (۱۰) و (۱۱) تعیین و طبق رابطه (۱۲) در بازه [۰،۱] نرمال سازی شد، بر اساس روش پیشنهادی در بخش ۴-۵ (رابطه (۱۳)) وزن فاصـله‌ای لغات مشـخص گردید و در ادامه طبق بخش ۴-۶ نیز بر اســاس میانگین وزنی فاصــله‌ای (IWA)، میزان ارتباط اســناد با پرس‌وجوی ورودی تعیین شــد و نهایتا بر اساس یکی از روش‌های سه‌گانه، رتبه‌بندی اسناد صورت گرفته و طبق روابط (۱۷ تا ۲۰) با بدســت آوردن مقدار MAP و P@n، نتیجه هر روش مورد ارزیابی قرار گرفت.

جدول ۶، نتایج MAP مربوط به این آزمایش‌ها را بر روی تمامی ۲۷ زیرمجموعه ممکن از وزن های پایه، در مجموعه داده های Cranfield و Medline، و برحسب روش‌های رتبه‌بندی مختلف (Lrank، Rrank، و Mrank) نشــان می‌دهد. در این جدول هر ردیف مربوط به یک زیرمجموعه از روش‌های وزن‌دهی پایه است که در سه ستون سمت چپ جدول، شرایط در نظر گرفته شده جهت انتخاب زیرمجموعه‌های وزنی پایه (طبق جدول ۵) برای آن ردیف معرفی شده است. بر اساس این شرایط، شماره روش‌های وزنی پایه انتخابی مورد اســتفاده در زیرمجموعه مربوط به هر ردیف از جدول دقیقاً مشخص می‌شود که در ستون چهارم جدول آورده شده است. در آزمایش‌های اولیه، زیرمجموعه‌های وزنی یکسـانی برای اسـناد و پرس‌وجوها استفاده شد؛ هرچند در آزمایش‌ها مشاهده شد که کنار گذاشــتن روش‌های وزن‌دهی شــامل مؤلفه p، از زیرمجموعه‌های وزنی، ممکن اســت برای اســناد مفید باشــد اما برای پرس‌وجوها کارآیی سیستم را پایین می‌آورد و بهتر است این روش‌های وزن‌دهی در زیرمجموعه‌های وزنی پرس‌وجوها حتماً مورد استفاده قرار گیرند؛ به همین دلیل در جدول ۶ بجز در ردیف‌هایی که در ســتون چهارم صــراحتاً برای پرس‌وجوها زیرمجموعه‌ی دیگری ذکر شــده، در بقیه موارد از زیرمجموعه‌های وزنی یکســانی برای اســناد و پرس‌وجوها استفاده شده است. به این ترتیب به عنوان مثال، سطر اول جدول ۶ بیان‌کننده این اســت که اگر در تشــکیل زیرمجموعه وزنی، از وزن‌های پایه ۱ تا ۱۲ (A1) که نرمال سازی طول بردار بر روی آنها انجام نشده استفاده گردد، بطوریکه فاقد روش‌های وزن‌دهی برگرفته از مؤلفه p بوده (B1) و شــامل روش‌های وزن‌دهی بولی ۱ یا ۱۳ باشند (C1)، در اینصورت زیرمجموعه انتخابی مورد آزمایش شامل روش های وزن‌دهی پایه ۱ تا ۹ (یا به عبارتی روش های bxx تا npx از جدول ۳) خواهد بود که در شش ستون بعد نتایج MAP برای دو مجموعه داده و با اســتفاده از هر یک از ســه روش مختلف رتبه‌بندی آورده شــده اســت. برای هر زیرمجموعه در هر مجموعه داده، بهترین روش رتبه‌بندی با زمینه طوسی مشخص شده است. همچنین مقادیر ضـخیم شـده در هر سـتون، معرف زیرمجموعه‌ای اسـت که بهترین نتیجه را برای آن روش رتبه‌بندی، در آن مجموعه

www.SID.ir

Archive of SID

داده، داشته است. نتایج این جدول از جنبه‌های مختلفی قابل بحث می‌باشد که در ادامه به آنها اشاره خواهد شد.

جدول ۶. مقدار MAP زیرمجموعه‌های وزنی مختلف، در روش وزن‌دهی فاصله‌ای، برحسب روش‌های رتبه‌بندی مختلف، بر روی مجموعه داده‌های Cranfield و Medline (مقادیر با زمینه طوسی، بیشترین مقدار بین سه روش رتبه‌بندی در هر مجموعه داده، و مقادیر ضخیم شده، بیشترین مقدار در هر ستون هستند)

				Cranfield			Medline		
			روش‌های وزن‌دهی پایه مورد استفاده در زیرمجموعه‌ها	Lrank	Rrank	Mrank	Lrank	Rrank	Mrank
روش‌های A_1: ۱-۱۲ (بدون نرمال‌سازی طول)	B_1:Lacking z component	C_1: including ۱ or ۱۳	۱-۹	۰,۳۹۰۱۱	۰,۳۷۴۶۱	۰,۳۹۸۲۶	۰,۵۲۷۴۶	۰,۴۹۸۵۳	۰,۵۰۹۹۶
		C_2: excluding ۱ & ۱۳	۲-۹	۰,۳۹۰۶	۰,۴۰۳۲۳	۰,۴۱۸۲۴	۰,۵۳۲۱۳	۰,۵۲۲۲۶	۰,۵۲۹۵۴
		C_3: excluding ۱, ۱۳ & p	۲, ۴-۵, ۷-۸ (weighting subset for queries: ۲-۹)	۰,۳۹۹۹۳	۰,۴۰۲۰۵	۰,۴۱۷۰۸	۰,۵۳۴۸۹	۰,۵۱۷۱۲	۰,۵۲۸۶۴
	B_2:Lacking n component	C_1: includ. ۱ or ۱۳	۱-۶, ۱۰-۱۲	۰,۳۷۲۰۳	۰,۳۸۵۰۶	۰,۴۰۷۷۵	۰,۴۸۲۶۴	۰,۵۰۴۷۰	۰,۵۲۳۴۲
		C_2: exclud. ۱& ۱۳	۲-۶, ۱۰-۱۲	۰,۳۸۹۸۷	۰,۳۹۶۹۱	۰,۴۰۶۲۱	۰,۵۰۴۳۸	۰,۵۲۴۸۹	۰,۵۲۸۶۷
		C_3: excluding ۱, ۱۳ & p	۲, ۴-۵, ۱۰-۱۱ (queries: ۲-۶, ۱۰-۱۲)	۰,۳۹۳۴۱	۰,۳۹۹۳۷	۰,۴۰۸۱۷	۰,۴۹۹۹۷	۰,۵۱۹۲۹	۰,۵۲۱۵۲
	B_3:including either z or n component	C_1: includ. ۱ or ۱۳	۱-۱۲	۰,۳۸۶۶۴	۰,۳۸۶۰۷	۰,۴۰۸۴۲	۰,۵۱۳۰۲	۰,۵۰۳۲۱	۰,۵۱۹۷۲
		C_2: exclud. ۱& ۱۳	۲-۱۲	۰,۳۹۱۲۴	۰,۴۰۲۹۷	۰,۴۱۳۰۴	۰,۵۱۹۱۷	۰,۵۲۳۱۵	۰,۵۳۲۱۹
		C_3: excluding ۱, ۱۳ & p	۲, ۴-۵, ۷-۸, ۱۰-۱۱ (queries: ۲-۱۲)	۰,۳۹۹۰۳	۰,۴۰۱۳۳	۰,۴۱۵۴۳	۰,۵۱۶۳۹	۰,۵۱۹۱۱	۰,۵۲۸۸۳
روش‌های A_2: ۱۳-۲۴ (همراه با نرمال‌سازی طول)	B_1:Lacking z component	C_1: includ. ۱ or ۱۳	۱۳-۲۱	۰,۳۷۵۸۵	۰,۴۲۷۹	۰,۴۳۳۲۳	۰,۵۳۲۹۸	۰,۵۲۵۶۷	۰,۵۳۷۹۸
		C_2: exclud. ۱& ۱۳	۱۴-۲۱	۰,۳۸۱۶۴	۰,۴۲۵۴۴	۰,۴۳۲۳۵	۰,۵۳۴۴۱	۰,۵۲۳۶۹	۰,۵۴۱۳۴
		C_3: excluding ۱, ۱۳ & p	۱۴, ۱۶-۱۷, ۱۹-۲۰ (queries: ۱۴-۲۱)	۰,۳۹۷۹۹	۰,۴۲۶۹۸	۰,۴۳۲۴۲	۰,۵۴۳۱۳	۰,۵۲۱۰۰	۰,۵۳۹۰۵
	B_2:Lacking n component	C_1: includ. ۱ or ۱۳	۱۳-۱۸, ۲۲-۲۴	۰,۴۱۴۳۲	۰,۴۱۹۲۷	۰,۴۲۹۶	۰,۵۴۳۹۴	۰,۵۱۴۴۰	۰,۵۲۸۸۶
		C_2: exclud. ۱& ۱۳	۱۴-۱۸, ۲۲-۲۴	۰,۴۱۱۱۱	۰,۴۱۶۲۳	۰,۴۲۴۷۶	۰,۵۴۲۰۰	۰,۵۱۲۵۰	۰,۵۲۵۹۴
		C_3: excluding ۱, ۱۳ & p	۱۴, ۱۶-۱۷, ۲۲-۲۳ (queries:۱۴-۱۸, ۲۲-۲۴)	۰,۴۱۹۳۷	۰,۴۱۱۹۹	۰,۴۲۲۳۶	۰,۵۴۳۷۰	۰,۵۰۶۶۸	۰,۵۲۱۱۵
	B_3:including either z or n component	C_1: includ. ۱ or ۱۳	۱۳-۲۴	۰,۴۰۳۵۵	۰,۴۲۴۳۵	۰,۴۳۱۸۹	۰,۵۴۵۸۰	۰,۵۱۹۲۷	۰,۵۳۵۹۷
		C_2: exclud. ۱& ۱۳	۱۴-۲۴	۰,۴۰۵۸۷	۰,۴۲۳۱۲	۰,۴۲۹۵۱	۰,۵۴۳۳۰	۰,۵۱۹۱۶	۰,۵۳۶۱۹
		C_3: excluding ۱, ۱۳ & p	۱۴،۱۶-۱۷،۱۹-۲۰, ۲۲-۲۳ (queries: ۱۴-۲۴)	۰,۴۱۵۳۸	۰,۴۲۱۶۹	۰,۴۳۱۳۲	۰,۵۴۵۴۴	۰,۵۱۴۷۱	۰,۵۳۱۸۶
روش‌های A_3: ۱-۲۴ (شامل هر دو دسته)	B_1:Lacking z component	C_1: includ. ۱ or ۱۳	۱-۹, ۱۳-۲۱	۰,۲۷۶۲۷	۰,۴۰۴۸۷	۰,۴۱۵۷۲	۰,۴۴۱۷۰	۰,۵۱۳۴۹	۰,۵۲۱۰۷
		C_2: exclud. ۱& ۱۳	۲-۹, ۱۴-۲۱	۰,۳۴۵۳	۰,۴۲۱۹	۰,۴۲۵۰۲	۰,۵۱۱۶۲	۰,۵۲۸۶۹	۰,۵۳۱۱۳
		C_3: excluding ۱, ۱۳ & p	۲, ۴-۵, ۷-۸, ۱۴, ۱۶-۱۷, ۱۹-۲۰ (queries: ۲-۹, ۱۴-۲۱)	۰,۳۶۳۰۳	۰,۴۲۰۶۶	۰,۴۲۵۸۲	۰,۵۱۷۵۹	۰,۵۳۰۶۳	۰,۵۳۳۹۴
		C_1: includ. ۱ or ۱۳	۱-۶, ۱۰-۱۲, ۱۳-۱۸, ۲۲-	۰,۲۷۱۱۸	۰,۴۱۲۳۳	۰,۴۱۸۸۶	۰,۳۷۹۷۱	۰,۵۲۲۶۲	۰,۵۳۱۳۰
		C_2: exclud. ۱& ۱۳	۲-۶, ۱۰-۱۲, ۱۴-۱۸, ۲۲-	۰,۳۶۹۷	۰,۴۱۲۱۳	۰,۴۱۷۲۱	۰,۴۹۳۴۹	۰,۵۲۶۴۶	۰,۵۲۹۶۱

www.SID.ir

Archive of SID

	B_2:Lacking n component	C_4: excluding ۱, ۱۳ & p	۲, ۴-۵, ۱۰-۱۱, ۱۴, ۱۶-۱۷, ۲۲-۲۳ (queries:۲-۶, ۱۰-۱۲, ۱۴-۱۸, ۲۲-۲۴)	۰,۳۷۹۳۳	۰,۴۱۵۷۷	۰,۴۱۷۸۹	۰,۴۹۱۹۸	۰,۵۲۳۰۰	۰,۵۲۴۳۵
	B_3:including either z or n component	C_1: includ. ۱ or ۱۳	۱-۲۴	۰,۲۹۲۰۴	۰,۴۱۴۵۴	۰,۴۲۲۲۷	۰,۴۳۰۰۶	۰,۵۲۲۶۹	۰,۵۳۰۰۴
		C_2: exclud. ۱& ۱۳	۲-۱۲, ۱۴-۲۴	۰,۳۵۷۱۳	۰,۴۱۷۷۲	۰,۴۲۱۲۹	۰,۴۹۷۹۰	۰,۵۲۹۸۲	۰,۵۳۴۰۰
		C_3: excluding ۱, ۱۳ & p	۲, ۴-۵, ۷-۸, ۱۰-۱۱, ۱۴, ۱۶-۱۷, ۱۹-۲۰, ۲۲-۲۳ (queries: ۲-۱۲, ۱۴-۲۴)	۰,۳۶۸۳۱	۰,۴۲۱۱	۰,۴۲۳۱۸	۰,۴۹۵۵۲	۰,۵۲۹۱۵	۰,۵۳۰۲۹

۵_۲_۱ ارزیابی کلی نتایج

با نگاه به جدول ۶، مشاهده می‌شود که در مجموعه داده Creanfield، برای کلیه زیرمجموعه‌ها بهترین نتایج زمانی بدست آمده است که رتبه‌بندی بر اساس "مرکز" اعداد فاصله‌ای مربوط به شباهت‌ها (Mrank) صورت گرفته است. در مجموعه داده Medline نیز غالباً در استفاده از زیرمجموعه‌هایی از A_1 و A_3 رتبه‌بندی Mrank، و در استفاده از زیرمجموعه‌هایی از A_2 (که اتفاقا شامل بهترین نتیجه هم هست)، رتبه‌بندی Lrank بهتر بوده است.

شکل ۱ که بر اساس جدول ۶ بدست آمده، مقادیر بیشترین، کمترین، و میانگین MAP تمامی ۲۷ زیرمجموعه وزنی مختلف از روش‌های وزن‌دهی پایه را به ازاء هر یک از سه روش رتبه‌بندی پیشنهادی (Lrank، Rrank، و Mrank)، در هر یک از مجموعه داده‌های Cranfield و Medline، نشان می‌دهد. در اینجا نیز از روی مقادیر میانگین نتایج در سه روش رتبه‌بندی مختلف مشاهده می‌شود که در هر دو مجموعه داده، بطور متوسط نتایج روش رتبه‌بندی Rrank بهتر از Lrank، و روش رتبه‌بندی Mrank بهتر از آن دو بوده است.

با مقایسه فواصل بین مقادیر بیشترین و کمترین، در روش‌های رتبه‌بندی مختلف مشاهده می‌شود که در هر دو مجموعه داده، پراکندگی نتایج بدست آمده از زیرمجموعه‌های وزنی مختلف، در روش رتبه‌بندی Rrank کمتر از Lrank، و در روش رتبه‌بندی Mrank کمتر از آن دو بوده است. این موضوع با محاسبه انحراف استاندارد نتایج نیز تایید شد، جائیکه انحراف استاندارد نتایج بدست آمده از زیرمجموعه‌های وزنی مختلف، برای روش‌های رتبه‌بندی Lrank، Rrank، و Mrank، در مجموعه داده Cranfield، به ترتیب ۰,۰۳۹، ۰,۰۱۴، و ۰,۰۰۹، و در مجموعه داده Medline، به ترتیب ۰,۰۴، ۰,۰۰۸، و ۰,۰۰۷ بدست آمد. پراکندگی کمتر بدین معنی است که وابستگی نتایج به زیرمجموعه‌های انتخاب‌شده از وزن‌های پایه، کمتر بوده و صرف‌نظر از زیرمجموعه انتخابی، بیشتر می‌توان به نتایج بدست آمده اطمینان کرد.

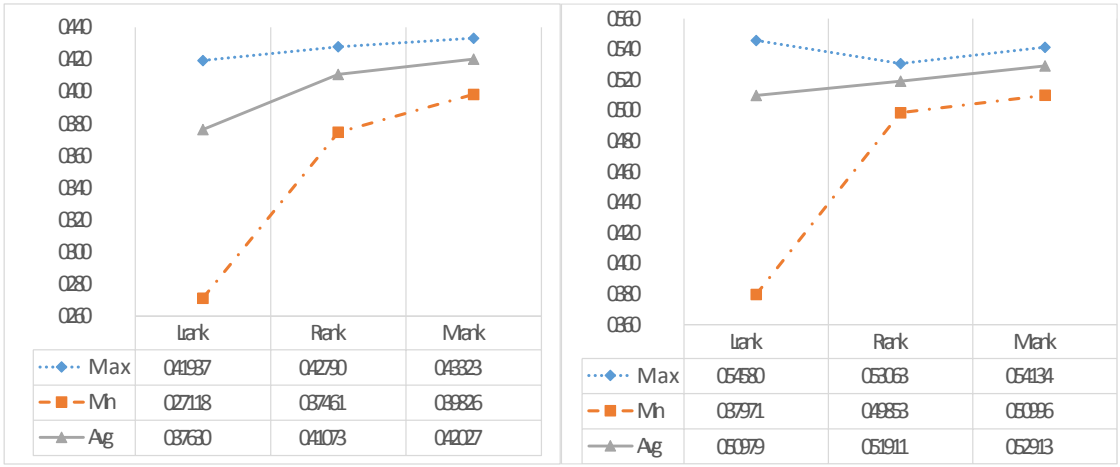
با محاسبه میانگین نتایج روش‌های مختلف در هریک از مجموعه داده‌ها مشاهده می‌شود که میانگین نتایج روش وزن‌دهی فاصله‌ای، در مجموعه داده‌های Cranfield و Medline، به ترتیب ۰,۴۰۲۴۳ و ۰,۵۱۹۳۴ می‌باشد که نه تنها از میانگین نتایج روش‌های وزن‌دهی پایه متناظر هر مجموعه داده در جدول ۵ بهتر هستند بلکه نزدیک به بیشترین مقدار این مجموعه داده‌ها در جدول ۵ هستند که همگی نشان‌دهنده تأثیر مثبت درنظر گرفتن وزن‌های فاصله‌ای است.

۵_۲_۲ اثر نرمال‌سازی طول وزن‌های پایه

جهت بررسی اینکه نرمال‌سازی طول وزن‌های پایه مورد استفاده در زیرمجموعه‌های وزنی، چه تأثیری در نتایج دارد، میانگین نتایج MAP روش‌های وزن‌دهی در هر یک از نواحی A_1، A_2، و A_3، برای هر یک از سه روش رتبه‌بندی Lrank، Rrank، و Mrank، در هر یک از مجموعه داده‌های Cranfield و Medline، در جدول ۶ بدست آمد که در شکل ۲ در قالب سه خط مختلف قابل مشاهده هستند.

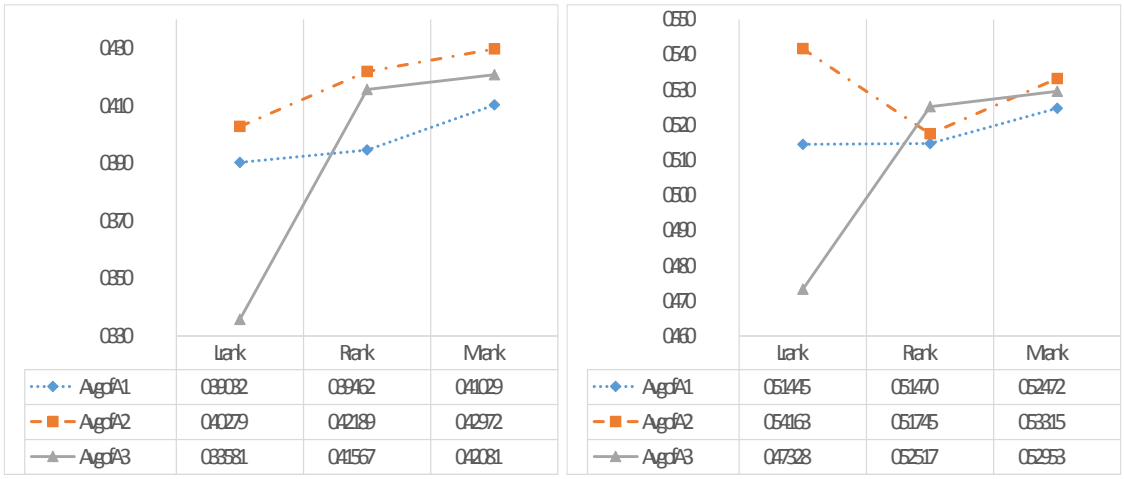
www.SID.ir

Archive of SID



(a) Cranfield (b) Medline

شکل ۱. مقایسه بیشترین، کمترین، و میانگین MAP تمامی ۲۷ زیرمجموعه وزنی، در روش‌های رتبه‌بندی مختلف جدول ۶، در مجموعه داده‌های Cranfield و Medline



(a) Cranfield (b) Medline

شکل ۲. تاثیر نرمال‌سازی طول: میانگین MAP زیرمجموعه‌های وزنی نواحی A_1، A_2، و A_3، در روش‌های رتبه‌بندی مختلف جدول ۶

همانطور که مشاهده می‌شود، در کل، نتایج بدست آمده از زیرمجموعه‌های وزنی ناحیه A_2 مناسب‌تر از A_3، و این نیز غالباً بهتر از A_1 بوده است که نشان‌دهنده تأثیر مثبت نرمال سازی طول بردار وزن‌های پایه مورد استفاده است. البته در روش رتبه‌بندی Rrank در مجموعه داده Medline، نتایج A_3 بهتر از A_2 شده، اما همانطور که مشاهده می شود و در بخش ۵-۲-۱ نیز اشاره شد، در این مجموعه داده نتایج روش‌های رتبه‌بندی Lrank و Mrank مناسبتر بوده و بیشتر مدنظر هستند. همچنین مشاهده می‌شود که در استفاده از زیرمجموعه‌های وزنی ناحیه A_2، در مجموعه داده Cranfield استفاده از روش رتبه‌بندی Mrank، و در مجموعه داده Medline استفاده از روش رتبه‌بندی Lrank، مناسب‌تر بوده است.

۵_۲_۳ اثر استفاده از مؤلفه‌های نرمال‌سازی TF مختلف، در وزن‌های پایه

شکل ۳ از میانگین‌گیری بر روی نتایج MAP زیرمجموعه های وزنی در هر یک از نواحی B_1، B_2، و B_3، برای هر یک از سه روش رتبه‌بندی Lrank، Rrank، و Mrank، در مجموعه داده های Cranfield و Medline، در جدول ۶ بدست آمده تا تأثیر استفاده از هر یک مشخص شود. همانطور که مشاهده می‌شود، بطورکلی در هر دو مجموعه داده، استفاده از زیرمجموعه‌های وزنی B_1 و B_3 منجر به نتایج بهتری نسبت به B_2 شده است؛ یعنی بهتر است در زیرمجموعه‌های انتخابی حتماً از روش‌های وزن‌دهی پایه شامل مؤلفه n، استفاده شود. در شکل ۳ مشاهده می‌شود که در هر دو مجموعه داده، در روش رتبه‌بندی Mrank نتایج بهتری بدست آمده است اما اگر نتایج بدست آمده در بخش ۵-۲-۲ نیز مدنظر

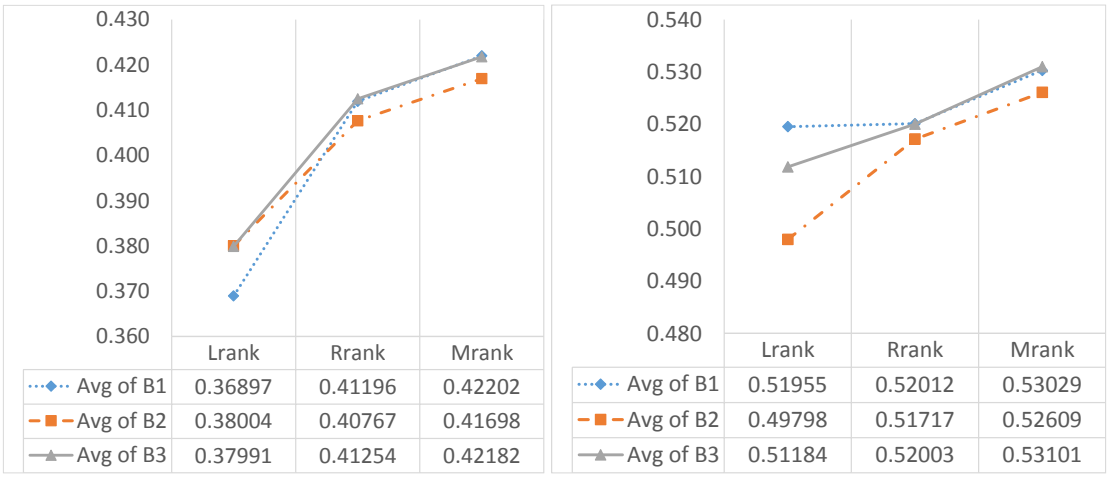
www.SID.ir

Archive of SID

قرار داده و تأثیر این مؤلفه‌ها فقط در ناحیه برتر A_2 مورد بررسی قرار گیرد، مشاهده می‌شود که در A_2، در مجموعه داده Cranfield، زیرمجموعه‌های وزنی B_1 در روش رتبه‌بندی Mrank با میانگین ۰٫۴۳۲۶۷، بهترین نتایج را داده‌اند. در مجموعه داده Medline، نیز با اینکه نتایج این ناحیه خوب هستند اما بهترین نتایج به ازاء زیرمجموعه‌های وزنی B_3 در روش رتبه‌بندی Lrank با میانگین ۰٫۵۴۴۸۵، بدست آمده است.

۵_۲_۴ اثر استفاده از روش‌های بولی، و روش‌های وزن‌دهی شامل مؤلفه *P*

شکل ۴ از میانگین‌گیری بر روی نتایج MAP زیرمجموعه‌های وزنی در هر یک از نواحی C_1، C_2، و C_3، در هر یک از سه روش رتبه‌بندی Lrank، Rrank، و Mrank، برای هر دو مجموعه داده در جدول ۶ بدست آمده است. همانطور که مشاهده می‌شود، استفاده از روش‌های وزن‌دهی بولی در زیرمجموعه وزن‌های پایه (C_1)، در هر دو مجموعه داده باعث کاهش در نتایج MAP شده است. بعد از کنار گذاشتن روش‌های وزن‌دهی بولی، بطور متوسط، در مجموعه داده های Cranfield عدم استفاده از وزن های وزن‌دهی شامل مؤلفه p فرکانس مجموعه در زیرمجموعه‌های وزنی مناسب‌تر است (یعنی C_3 بهتر از C_2)، در حالیکه در مجموعه داده Medline، بجز در روش Lrank در روش‌های Rrank و Mrank برعکس است.



(a) Cranfield (b) Medline

شکل ۳. تاثیر مؤلفه‌های نرمال‌سازی TF مختلف: میانگین MAP زیرمجموعه‌های وزنی نواحی B_1، B_2، و B_3، در روش‌های رتبه‌بندی مختلف جدول ۶

در شکل ۴ نیز شاید درکل، در هر دو مجموعه داده، روش رتبه‌بندی Mrank نتایج بهتری داده است اما اگر با توجه به نتایج بخش ۵-۲-۲ فقط ناحیه برتر A_2 مورد بررسی قرار گیرد، مشاهده می شود که در A_2، در مجموعه داده Cranfield، زیرمجموعه های وزنی C_1 در روش رتبه‌بندی Mrank با میانگین ۰٫۴۳۱۵۷، و در مجموعه داده Medline، زیرمجموعه‌های وزنی C_3 در روش رتبه‌بندی Lrank با میانگین ۰٫۵۴۴۰۹ بهترین نتایج را داده‌اند.

۵_۲_۵ بررسی بهترین روش‌ها

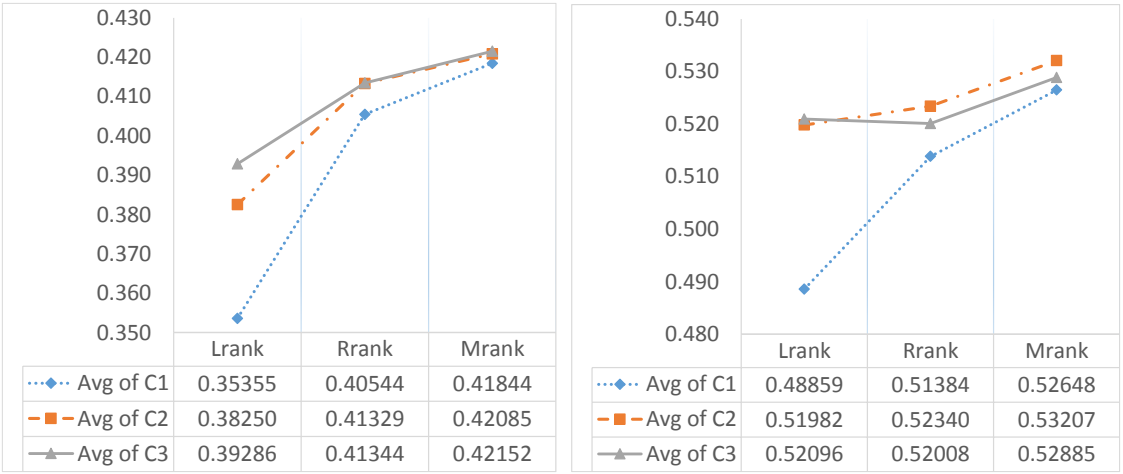
در جدول ۶، نتایج مربوط به ۸۱ آزمایش مختلف بر روی دو مجموعه داده آورده شد و در بخش‌های قبل اثر روش‌ها، پارامترها، و مؤلفه‌های مختلف در این نتایج مورد بررسی و بحث قرار گرفت.

اگر به پنج روش برتر در هر یک از مجموعه داده‌ها که در جدول ۶ زیر آنها خط کشیده شده توجه شود نیز به همان نتایجی که در بحث‌های قبلی اشاره شد، رسیده و مشاهده می‌گردد که:

- در مجموعه Cranfield بهترین نتایج مربوط به استفاده از زیرمجموعه هایی از روش های وزن‌دهی ۱۳-۲۴ بوده (A_2) که در آنها حتماً از وزن‌های پایه شامل مؤلفه n استفاده شده (B_1 یا B_3) و جهت رتبه‌بندی نهایی از Mrank استفاده شده است. در این شرایط، استفاده از زیرمجموعه‌های وزنی C_1 و سپس C_3 مناسب‌تر بوده‌اند، یعنی استفاده از وزن‌های بولی و وزن‌های شامل مؤلفه p در زیرمجموعه‌های وزنی (C_1)، یا عدم استفاده از هیچیک از آنها (C_3)

www.SID.ir

Archive of SID



(a) Cranfield (b) Medline

شکل ۴. تاثیر استفاده از روش‌های بولی و مؤلفه p: میانگین MAP زیرمجموعه‌های وزنی نواحی C_1،C_2، وC_3، در روش‌های رتبه‌بندی مختلف جدول ۶

- در مجموعه Medline بهترین نتایج مربوط به استفاده از زیرمجموعه هایی از روش های وزن‌دهی ۱۳-۲۴ بوده (A_2) که در آنها حتماً از وزن‌های پایه شامل مؤلفه z استفاده شده (B_2 یا B_3) و جهت رتبه‌بندی نهایی از Lrank استفاده شده است. در این شرایط نیز مانند Cranfield، استفاده از زیرمجموعه های وزنی C_1 و سپس C_3 مناسب‌تر بوده‌اند.

م شاهده می‌گردد که در ر سیدن به بهترین نتایج، خیلی از شرایط بین دو مجموعه داده یکسان است اما به هر حال در مواردی که با هم اختلاف دار ند، چه مؤلفه ها و روش هایی را درنظر گر فت که صرف‌نظر از خصوصیات مجموعه داده، نتایج بهتری بدهد؟ برای ر سیدن به پا سخ این سوال، ابتدا رتبه‌ی هر یک از ۸۱ روش مذکور در مجموعه داده مربوطه بدست آمد و سپس میانگین رتبه این روش‌ها در دو مجموعه داده محاسبه و به صورت صعودی مرتب شدند تا برترین روش‌ها بدست آیند. در جدول ۷ میانگین رتبه‌ی ده روش برتر در دو مجموعه داده مشاهده می‌شود، ضمن اینکه مؤلفه‌ها و روش‌های بکار رفته در هر روش به همراه رتبه و MAP آنها در هر یک از مجموعه داده‌ها آورده شده است. طبق جدول ۷ مشاهده می‌شود که:

- در ۸۰ درصد از ده روش برتر، از وزن های پا یه‌ای با نرمال‌سازی طول در زیرمجموعه‌ها (A_2) استفاده شده و
- در ۲۰ درصد، از هر دو سری وزن‌های پایه‌ی با نرمال‌سازی طول و بدون نرمال سازی طول (A_3) استفاده شده است، اما زیرمجموعه‌هایی که در آنها فقط وزن‌های پایه‌ی بدون نرمال‌سازی طول (A_1) بکار رفته باشد، مشاهده نمی‌شود.
- در ۵۰ درصد از آنها، از وزن‌های پایه فاقد مؤلفه z در فاکتور TF (زیرمجموعه‌هایی از B_1)، در ۲۰ درصد از روش ها از وزن های فا قد مؤلفه n (زیرمجموعه هایی از B_2)، و در ۳۰ درصد دیگر از هر دو مؤلفه (زیرمجموعه-هایی از B_3) استفاده شده است.
- ۳۰ درصد از این روش‌ها، از وزن‌های پایه بولی و وزن‌های شامل مؤلفه p در فرکانس مجموعه (زیرمجموعه‌هایی از C_1) استفاده کرده‌اند، در ۳۰ درصد، بدون استفاده از وزن‌های بولی، از وزن‌های پایه شامل مؤلفه p در فرکانس مجموعه (زیرمجموعه‌هایی از C_2) ا ستفاده شده و در ۴۰ درصد باقیمانده علاوه بر عدم استفاده از وزن‌های بولی، از روش‌های فاقد مؤلفه p (زیرمجموعه‌هایی از C_3) ا ستفاده شده است.

www.SID.ir

Archive of SID

جدول ۷. روش‌های برتر وزن‌دهی و بازیابی، به همراه رتبه و مقدار MAP هر یک در هر مجموعه داده و متوسط رتبه آنها در دو مجموعه

	Methods				Rank / MAP		Mean rank for ۲ datasets
	Length Normalization	Using n & z components	Using p components	Ranking Method	Cranfield	Medline	
۱	$A_۲$	$B_۱$	$C_۱$	Mrank	۱ / ۰.۴۳۳۲۳	۱۰ / ۰.۸۳۷۹۱	۵,۵
۱	$A_۲$	$B_۱$	$C_۲$	Mrank	۳ / ۰.۴۳۲۳۸	۸ / ۰.۸۴۱۳۴	۵,۵
۱	$A_۲$	$B_۱$	$C_۳$	Mrank	۲ / ۰.۴۳۲۴۲	۹ / ۰.۸۳۹۰۸	۵,۵
۴	$A_۲$	$B_۳$	$C_۱$	Mrank	۴ / ۰.۴۳۱۸۹	۱۲ / ۰.۸۳۸۹۷	۸
۵	$A_۲$	$B_۳$	$C_۲$	Mrank	۷ / ۰.۴۲۹۸۱	۱۱ / ۰.۸۳۶۱۹	۹
۶	$A_۲$	$B_۳$	$C_۳$	Mrank	۵/ ۰.۴۳۱۳۲	۲۰/ ۰.۸۳۱۸۶	۱۲,۵
۷	$A_۳$	$B_۱$	$C_۳$	Mrank	۱۰ / ۰.۴۲۸۸۲	۱۶ / ۰.۸۳۳۹۴	۱۳
۸	$A_۲$	$B_۲$	$C_۳$	Lrank	۲۴ / ۰.۴۱۹۳۷	۴ / ۰.۸۴۳۷۰	۱۴
۹	$A_۳$	$B_۱$	$C_۲$	Mrank	۱۲ / ۰.۴۲۸۰۲	۲۲ / ۰.۸۳۱۱۳	۱۷
۱۰	$A_۲$	$B_۲$	$C_۱$	Mrank	۶ / ۰.۴۲۹۶	۳۰ / ۰.۸۲۸۸۶	۱۸

- در ۹۰ درصد از موارد از روش رتبه‌بندی Mrank، و در ۱۰درصد از روش Lrank استفاده شده است، اما در هیچیک از ده روش برتر از روش رتبه‌بندی Rrank استفاده نشده است.

با توجه به جدول ۷ و بررسی‌های فوق، می‌توان گفت که بطورکلی در روش پیشنهادی با ۲۴ روش وزن‌دهی پایه در نظر گرفته شده، استفاده از زیرمجموعه‌هایی از وزن‌های پایه که نرمال‌سازی طول روی آنها انجام شده ($A_۲$)، فاقد مؤلفه z در فاکتور TF هستند ($B_۱$)، در کنار استفاده از روش رتبه‌بندی Mrank مناسب‌تر است. با نگاه به شرایط سه سطر اول جدول ۷، مشاهده می‌شود که استفاده یا عدم استفاده از وزن‌های بولی و وزن‌های پایه شامل مؤلفه p در فرکانس مجموعه (C_iهای مختلف)، تأثیری در میانگین رتبه‌های آنها نداشته است؛ با این وجود با توجه به شکل ۴، در روش رتبه‌بندی Mrank، شاید عدم استفاده از وزن‌های بولی ($C_۲$) نتایج بهتری را داشته باشد.

۵_۲_۶ بررسی معیار P@N

باید توجه داشت که اکثر کاربران موتورهای جستجو تنها نتایج ابتدای لیست اسناد برگشتی را بررسی کرده و برای ۸۰ درصد آنها تنها سه نتیجه برگشتی ابتدایی جالب می‌باشد [۵۰]. برای بررسی نتایج از این نقطه نظر، در جدول ۸ مقادیر معیارهای P@۵، P@۱۰، و P@۲۰، برای وزن‌های پایه ۱۳ تا ۲۴ (که زیرمجموعه‌های انتخابی از آنها در بخش قبل مناسب‌تر تشخیص داده شدند)، و در جدول ۹ مقادیر این معیارها برای وزن‌های فاصله‌ای بدست آمده از زیرمجموعه‌های مختلف وزن‌های ۱۳ تا ۲۴، به ازاء روش‌های رتبه‌بندی مختلف، در دو مجموعه داده Cranfield و Medline آورده شده است.

www.SID.ir

Archive of SID

جدول ۸. مقادیر دقت در برترین ۵، ۱۰، و ۲۰ سند ارائه شده، به ازاء روش‌های وزن‌دهی پایه ۱۳ تا ۲۴، در هر مجموعه داده

شماره و نام سه جزئی روش			Cranfield			Medline		
			P@۵	P@۱۰	P@۲۰	P@۵	P@۱۰	P@۲۰
روش‌های ۱۳-۲۴ (همراه با نرمال‌سازی طول)	۱۳	bxc	۰,۳۳۹۵۶	۰,۲۴۴۴۴	۰,۱۶۳۷۸	۰,۷۲۶۶۷	۰,۵۸۰۰۰	۰,۴۶۳۳۳
	۱۴	bfc	۰,۳۴۹۳۳	۰,۲۵۶۰۰	۰,۱۶۹۱۱	۰,۶۳۳۳۳	۰,۵۸۳۳۳	۰,۵۱۱۶۷
	۱۵	bpc	۰,۳۴۳۱۱	۰,۲۴۹۷۸	۰,۱۶۶۶۷	۰,۶۴۰۰۰	۰,۵۸۶۶۷	۰,۵۰۶۶۷
	۱۶	txc	۰,۳۹۲۰۰	۰,۲۷۷۳۳	۰,۱۸۳۷۸	۰,۶۷۳۳۳	۰,۵۷۰۰۰	۰,۴۶۸۳۳
	۱۷	tfc	۰,۴۲۷۵۶	۰,۳۰۳۱۱	۰,۲۰۱۱۱	۰,۶۶۶۶۷	۰,۶۰۳۳۳	۰,۵۳۳۳۳
	۱۸	tpc	۰,۴۱۲۴۴	۰,۲۹۷۷۸	۰,۱۹۶۲۲	۰,۶۶۰۰۰	۰,۶۰۶۶۷	۰,۵۳۳۳۳
	۱۹	nxc	۰,۴۰۹۷۸	۰,۲۸۲۲۲	۰,۱۸۲۶۷	۰,۷۲۶۶۷	۰,۶۱۶۶۷	۰,۵۰۰۰۰
	۲۰	nfc	۰,۴۱۰۶۷	۰,۲۸۷۵۶	۰,۱۸۷۱۱	۰,۶۷۳۳۳	۰,۶۳۶۶۷	۰,۵۳۵۰۰
	۲۱	npc	۰,۴۰۲۶۷	۰,۲۸۵۳۳	۰,۱۸۴۰۰	۰,۶۷۳۳۳	۰,۶۳۳۳۳	۰,۵۳۸۳۳
	۲۲	zxc	۰,۳۹۲۰۰	۰,۲۷۷۳۳	۰,۱۸۳۷۸	۰,۶۷۳۳۳	۰,۵۷۰۰۰	۰,۴۶۸۳۳
	۲۳	zfc	۰,۴۲۷۵۶	۰,۳۰۳۱۱	۰,۲۰۱۱۱	۰,۶۶۶۶۷	۰,۶۰۳۳۳	۰,۵۳۳۳۳
	۲۴	zpc	۰,۴۱۲۴۴	۰,۲۹۷۷۸	۰,۱۹۶۲۲	۰,۶۶۰۰۰	۰,۶۰۶۶۷	۰,۵۳۳۳۳
بیشترین مقدار			۰,۴۲۷۵۶	۰,۳۰۳۱۱	۰,۲۰۱۱۱	۰,۷۲۶۶۷	۰,۶۳۶۶۷	۰,۵۳۸۳۳
میانگین مقادیر			۰,۳۹۳۲۶	۰,۲۸۰۱۵	۰,۱۸۴۶۳	۰,۶۷۲۷۸	۰,۵۹۹۷۲	۰,۵۱۰۴۲

همانگونه که در جداول ۸ و ۹ مشاهده می‌شود، در وزن‌های فاصله‌ای نیز مانند وزن‌های پایه، مقادیر دقت در nهای کوچکتر که بیشتر مورد توجه کاربران است، بالاتر می‌باشد. همچنین در مقایسه جداول ۶ و ۹، مشاهده می‌گردد که در معیار P@n نیز مانند معیار MAP در مجموعه داده Cranfield نتایج رتبه‌بندی بر اساس Mrank، و در مجموعه داده Medline نتایج رتبه‌بندی بر اساس Lrank مناسب‌تر بوده است. با مقایسه جداول ۸ و ۹ می‌توان متوجه شد که میانگین مقادیر بدست آمده از زیرمجموعه‌های مختلف در روش وزن‌دهی فاصله‌ای، به ازاء هر یک از روش‌های رتبه‌بندی، از میانگین روش‌های وزن‌دهی پایه‌ی مربوطه بهتر شده است. در روش وزن‌دهی فاصله‌ای، در مجموعه داده Cranfield، مقادیر میانگین Mrank (و در مجموعه داده Medline، مقادیر میانگین Lrank) حتی از بیشترین مقادیر معادل در جدول ۸ هم بهتر شده‌اند. با مقایسه نتایج MAP جدول ۶ با جداول ۴ و ۵ نیز مشاهده می‌گردد که نه تنها میانگین نتایج روش‌های پیشنهادی مختلف در این تحقیق از میانگین نتایج استفاده از هر یک از وزن‌های پایه به تنهایی، با اختلاف بهتر است، بلکه نتایج MAP روش پیشنهادی از اکثر نتایج روش‌های وزن‌دهی پایه در جدول ۴ نیز بهتر شده است. بنابراین بر اساس هر دو معیار، اثربخشی ایجاد وزن‌های فاصله‌ای بر اساس وزن‌های پایه، نسبت به استفاده از هر یک از وزن‌های پایه به تنهایی،کاملاً مشهود می‌باشد. همچنین باید توجه داشت که محاسبه وزن‌های پایه اولیه مورد نظر و محاسبه وزن‌های فاصله‌ای برای لغات مختلف اسناد، به صورت آفلاین صورت گرفته و در زمان ورود یک پرس‌وجوی جدید، تنها محاسبه میزان ارتباط آن با اسناد مختلف (آن هم فقط بر اساس لغات درون پرس‌وجو) انجام می‌گیرد؛ بنابراین روش پیشنهادی ضمن افزایش دقت، از روشی ساده استفاده کرده و افزایش پیچیدگی و زمان اجرا را به همراه ندارد.

www.SID.ir

Archive of SID

جدول ۹. مقادیر دقت در برترین ۵، ۱۰، و ۲۰ سند ارائه شده، برای زیرمجموعه‌های مختلف از روش‌های وزن‌دهی پایه ۱۳ تا ۲۴، در روش وزن‌دهی فاصله‌ای، برحسب روش‌های رتبه‌بندی مختلف، در هر مجموعه داده (مقادیر با زمینه طوسی، بیشترین مقدار بین سه روش رتبه‌بندی، و مقادیر ضخیم شده، بیشترین مقدار در هر ستون هستند)

	روش‌های وزن‌دهی پایه مورد استفاده در زیرمجموعه‌ها	روش رتبه‌بندی	Cranfield			Medline		
			P@۵	P@۱۰	P@۲۰	P@۵	P@۱۰	P@۲۰
روش‌های ۱۳-۲۴ (همراه با نرمال‌سازی طول) A_{2}:	۱۳-۲۱ ($B_1.C_1$)	Lrank	۰,۴۰۳۵۶	۰,۲۸۴۰۰	۰,۱۸۶۲۲	۰,۷۲۶۶۷	۰,۶۴۰۰۰	۰,۵۳۸۳۳
		Rrank	۰,۴۵۲۴۴	۰,۳۱۲۰۰	۰,۲۰۵۳۳	۰,۷۱۳۳۳	۰,۶۴۶۶۷	۰,۵۴۰۰۰
		Mrank	۰,۴۵۶۰۰	۰,۳۱۴۶۷	۰,۲۰۶۶۷	۰,۷۲۶۶۷	۰,۶۵۰۰۰	۰,۵۴۰۰۰
	۱۴-۲۱ ($B_1.C_2$)	Lrank	۰,۴۱۰۶۷	۰,۲۸۴۸۹	۰,۱۸۶۰۰	۰,۷۱۳۳۳	۰,۶۵۰۰۰	۰,۵۴۱۶۷
		Rrank	۰,۴۴۷۱۱	۰,۳۱۰۶۷	۰,۲۰۶۶۷	۰,۶۹۳۳۳	۰,۶۴۳۳۳	۰,۵۴۳۳۳
		Mrank	۰,۴۶۰۴۴	۰,۳۱۲۰۰	۰,۲۰۷۵۶	۰,۷۲۰۰۰	۰,۶۶۰۰۰	۰,۵۳۶۶۷
	۱۴, ۱۶-۱۷, ۱۹-۲۰ (queries: ۱۴-۲۱) ($B_1.C_3$)	Lrank	۰,۴۱۴۲۲	۰,۲۹۳۳۳	۰,۱۹۵۵۶	۰,۷۲۶۶۷	۰,۶۵۳۳۳	۰,۵۴۶۶۷
		Rrank	۰,۴۴۸۸۹	۰,۳۱۱۱۱	۰,۲۰۵۳۳	۰,۷۰۶۶۷	۰,۶۴۰۰۰	۰,۵۳۱۶۷
		Mrank	۰,۴۵۷۷۸	۰,۳۱۵۱۱	۰,۲۰۸۴۴	۰,۷۴۰۰۰	۰,۶۴۳۳۳	۰,۵۴۰۰۰
	۱۳-۱۸, ۲۲-۲۴ ($B_2.C_1$)	Lrank	۰,۴۴۵۳۳	۰,۳۰۴۸۹	۰,۱۹۸۶۷	۰,۷۲۶۶۷	۰,۶۵۶۶۷	۰,۵۴۸۳۳
		Rrank	۰,۴۴۱۷۸	۰,۳۱۱۵۶	۰,۲۰۴۴۴	۰,۷۰۶۶۷	۰,۶۲۳۳۳	۰,۵۳۰۰۰
		Mrank	۰,۴۵۷۷۸	۰,۳۱۲۴۴	۰,۲۰۹۳۳	۰,۷۱۳۳۳	۰,۶۳۳۳۳	۰,۵۳۸۳۳
	۱۴-۱۸, ۲۲-۲۴ ($B_2.C_2$)	Lrank	۰,۴۴۰۸۹	۰,۳۰۳۱۱	۰,۱۹۶۸۹	۰,۷۳۳۳۳	۰,۶۵۶۶۷	۰,۵۵۶۶۷
		Rrank	۰,۴۳۸۲۲	۰,۳۰۹۷۸	۰,۲۰۴۰۰	۰,۷۰۰۰۰	۰,۶۳۰۰۰	۰,۵۲۶۶۷
		Mrank	۰,۴۵۰۶۷	۰,۳۱۰۶۷	۰,۲۰۶۴۴	۰,۷۰۰۰۰	۰,۶۳۶۶۷	۰,۵۳۵۰۰
	۱۴, ۱۶-۱۷, ۲۲-۲۳ (queries: ۱۴-۱۸, ۲۲-۲۴) ($B_2.C_3$)	Lrank	۰,۴۴۲۶۷	۰,۳۰۹۳۳	۰,۲۰۴۲۲	۰,۷۴۶۶۷	۰,۶۵۳۳۳	۰,۵۵۶۶۷
		Rrank	۰,۴۳۷۳۳	۰,۳۰۳۱۱	۰,۲۰۳۱۱	۰,۷۱۳۳۳	۰,۶۱۶۶۷	۰,۵۲۱۶۷
		Mrank	۰,۴۴۷۱۱	۰,۳۱۰۶۷	۰,۲۰۶۲۲	۰,۷۱۳۳۳	۰,۶۴۳۳۳	۰,۵۳۰۰۰
	۱۳-۲۴ ($B_3.C_1$)	Lrank	۰,۴۲۵۷۸	۰,۲۹۸۲۲	۰,۱۹۳۵۶	۰,۷۴۶۶۷	۰,۶۵۶۶۷	۰,۵۵۳۳۳
		Rrank	۰,۴۴۶۲۲	۰,۳۱۴۲۲	۰,۲۰۳۷۸	۰,۷۰۶۶۷	۰,۶۲۶۶۷	۰,۵۳۱۶۷
		Mrank	۰,۴۵۶۰۰	۰,۳۱۶۰۰	۰,۲۰۹۳۳	۰,۷۲۶۶۷	۰,۶۴۶۶۷	۰,۵۳۶۶۷
	۱۴-۲۴ ($B_3.C_2$)	Lrank	۰,۴۳۴۶۷	۰,۲۹۶۰۰	۰,۱۹۵۵۶	۰,۷۲۶۶۷	۰,۶۵۶۶۷	۰,۵۵۶۶۷
		Rrank	۰,۴۴۷۱۱	۰,۳۱۳۷۸	۰,۲۰۴۸۹	۰,۷۰۰۰۰	۰,۶۱۶۶۷	۰,۵۳۵۰۰
		Mrank	۰,۴۵۶۸۹	۰,۳۱۲۴۴	۰,۲۰۸۲۲	۰,۷۱۳۳۳	۰,۶۵۰۰۰	۰,۵۳۸۳۳
	۱۴, ۱۶-۱۷, ۱۹-۲۰, ۲۲-۲۳ (queries: ۱۴-۲۴) ($B_3.C_3$)	Lrank	۰,۴۳۷۳۳	۰,۳۰۰۸۹	۰,۲۰۳۳۳	۰,۷۳۳۳۳	۰,۶۶۰۰۰	۰,۵۵۱۶۷
		Rrank	۰,۴۴۱۷۸	۰,۳۰۸۸۹	۰,۲۰۵۳۳	۰,۷۰۰۰۰	۰,۶۳۳۳۳	۰,۵۳۳۳۳
		Mrank	۰,۴۵۴۲۲	۰,۳۱۰۲۲	۰,۲۰۹۱۱	۰,۷۲۶۶۷	۰,۶۴۳۳۳	۰,۵۳۳۳۳
میانگین مقادیر Lrank			۰,۴۲۸۳۵	۰,۲۹۷۱۹	۰,۱۹۵۵۶	۰,۷۳۱۱۱	۰,۶۵۳۷۰	۰,۵۵۰۰۰
میانگین مقادیر Rrank			۰,۴۴۴۵۴	۰,۳۱۰۵۷	۰,۲۰۴۷۷	۰,۷۰۴۴۴	۰,۶۳۰۷۴	۰,۵۳۲۵۹
میانگین مقادیر Mrank			۰,۴۵۵۲۱	۰,۳۱۲۶۹	۰,۲۰۷۹۳	۰,۷۲۰۰۰	۰,۶۴۵۱۹	۰,۵۳۶۴۸

www.SID.ir

Archive of SID

۳،۵ مقایسه با دیگر روش‌ها

جهت مقایسه با دیگر روش‌های وزن‌دهی، مقالات زیادی مورد برر سی قرار گرفتند؛ در برخی از مقالات معیار MAP مدنظر نبوده و در برخی نیز نتایج بر اساس انجام آزمایش بر روی گزیده‌ای از پرس‌وجوها (نه تمامی آنها) ارائه شده ا ست و امکان مقایسه وجود نداشت. در جدول ۱۰، بهترین نتایج روش پیشنهادی در بین ترکیبات مختلف پارامترهای بررسی شده برای تعیین وزن‌های فاصله‌ای لغات، که در واقع مربوط به بهترین ترکیب در هر مجموعه داده هستند، به همراه نتایج ترکیب بهینه بر روی هر دو مجموعه داده ($A_۲$, $B_۱$, $C_۲$, Mrank) که در بخش ۵-۲-۵ مشخص شد، آورده شده است. این نتایج با بهترین نتایج روش بهترین تطابق BM۲۵ [۷] و [۲۶]، روش وزن‌دهی فرکانس لغت همراه با میانگین رویداد لغت (TF-ATO) در دو حالت عدم استفاده و استفاده از روش متمایزکننده (DA) به منظور حذف وزن های کم ارزش از اسناد [۸]، و مدل‌های زبانی سند $GSPUD_{bs_t}$ و $GSPUD_{mc}$ در فرآیند Pólya [۳۲]، مقایسه شده است. به دلیل موجود نبودن مقادیر P@n در این مقالات، امکان مقایسه از منظر این معیار میسر نبود.

جدول ۱۰. مقایسه مقدار MAP روش پیشنهادی با دیگر روش‌ها

مجموعه داده / روش/مدل		Cranfield	Medline
Wistchel [۷]	BM۲۵	۰,۴۱۶	۰,۵۰۹
Cummins [۲۶]	BM۲۵	۰,۴۲۲۳	۰,۵۳۴۳
Ibrahim and Landa-Silva [۸]	TF-ATO without DA	۰,۳۵۴۷	-
	TF-ATO with DA	۰,۳۹۹۸	-
Cummins [۳۲]	$GSPUD_{bs_t}$	۰,۴۲۷	۰,۵۲۳
	$GSPUD_{mc}$	۰,۴۳۲	۰,۵۳۳
روش پیشنهادی	بهترین نتایج بین روش‌های مختلف	۰,۴۳۳۲۳	۰,۵۴۵۸۰
	ترکیب بهینه پارامترها ($A_۲$,$B_۱$,$C_۲$,Mrank)	۰,۴۳۲۳۵	۰,۵۴۱۳۴

همانگونه که مشاهده می‌شود، در هر دو مجموعه داده Cranfield و Medline، نه فقط بهترین نتایج بدست آمده در روش پیشنهادی، بلکه نتایج ترکیب بهینه بدست آمده نیز از نتایج تمامی دیگر روش‌ها بهتر شده است که این نشان‌دهنده تأثیر مثبت استفاده از وزن‌دهی فاصله‌ای در روش پیشنهادی است.

۶. نتیجه‌گیری

در این تحقیق نشان داده شد که می‌توان به جای درنظر گرفتن یک مقدار وزنی برای لغات اسناد و پرس‌وجوها، از تمامی وزن‌هایی که توسط تعدادی روش وزن‌دهی مختلف تعیین شده به عنوان وزن‌های پایه بهره برد و به روشی ساده و به صورت آفلاین، یک فاصله (بازه) وزنی برای لغات بدست آورد که بدون افزایش پیچیدگی و زمان اجرا، باعث دقت می‌شود. در این مقاله ۲۴ روش وزن‌دهی پایه درنظر گرفته شد که از نظر مؤلفه‌های بکار رفته در فاکتور فرکانس لغت، فاکتور مجموعه، و فاکتور نرمال‌سازی طول بردار وزنی، تفاوت‌هایی با هم داشتند؛ بر این اساس، آزمایش‌های مختلفی صورت گرفت تا در کنار بررسی سه روش رتبه‌بندی مختلف، تأثیر بکارگیری مؤلفه های مختلف فاکتورهای فوق را در وزن های پایه بررسی کرده، تا بتوان بهترین زیرمجموعه‌های وزنی از ۲۴ روش وزن‌دهی پایه را جهت ایجاد وزن‌های فاصله‌ای تعیین کرد. در این بررسی‌ها نتایج زیر بدست آمد:

- در مجموعه داده Cranfield، بهترین نتایج مربوط به استفاده از زیرمجموعه‌هایی از روش‌های وزن‌دهی پایه است که نرمال‌سازی طول بردار روی آنها انجام شده و در آنها حتماً از وزن‌های پایه شامل مؤلفه n استفاده شده، و جهت رتبه‌بندی نهایی از روش Mrank استفاده شده است. در این شرایط، استفاده از وزن‌های بولی و وزن‌های شامل مؤلفه p در زیرمجموعه‌های وزنی، و سپس عدم استفاده از هیچیک از آنها در زیرمجموعه های وزنی، مناسب‌تر بوده‌اند.
- در مجموعه Medline نیز بهترین نتایج مربوط به استفاده از زیرمجموعه‌هایی از روش‌های وزن‌دهی پایه ا ست که نرمال سازی طول بردار روی آنها انجام شده، اما در آنها باید حتماً از وزن‌های پایه شامل مؤلفه z ا ستفاده شده و جهت رتبه‌بندی نهایی از Lrank استفاده شده باشد. در این شرایط نیز مانند Cranfield، استفاده از وزن‌های بو لی و وزن‌های شامل مؤ لفه p در زیرمجموعه‌های وزنی، و سپس عدم استفاده از هیچیک از آنها در زیرمجموعه‌های وزنی، مناسب‌تر بوده‌اند.

در آزمایش‌های انجام شده مشخص شد که استفاده از وزن‌های فاصله‌ای می‌تواند باعث بهبود اثربخشی بازیابی اطلاعات متنی

www.SID.ir

Archive of SID

نسبت به هریک از وزن‌های پایه شود. با بررسی که بر روی ده ترکیب برتر در بین ترکیب های مورد آزمایش صورت گرفت، مشخص شد که صرف‌نظر از خصوصیات مجموعه داده‌ها، جهت ایجاد زیرمجموعه‌های وزنی و بدست آوردن بهترین نتایج، استفاده از شرایط زیر مناسب‌تر است:

- استفاده از زیرمجموعه هایی از وزن های پایه که نرمال‌سازی طول روی بردار وزنی آنها صورت گرفته است.
- استفاده از زیرمجموعه‌هایی از وزن‌های پایه که فاقد مؤلفه z در فاکتور TF هستند.
- استفاده از روش رتبه‌بندی Mrank جهت رتبه‌بندی میزان ارتباط اسناد مختلف با یک پرس‌وجو.
- در شرایط فوق، هرچند استفاده یا عدم استفاده از وزن‌های بولی و وزن‌های پایه شامل مؤلفه p فرکانس مجموعه، تأثیر زیادی در نتایج ندارد، اما شاید عدم استفاده از وزن‌های بولی مناسب‌تر باشد.

در تحقیقات آتی می‌توان بر روی دیگر روش های وزن‌دهی پایه و دیگر مجموعه داده‌ها، آزمایش‌ها را انجام داده و در کنار آن بر روی دیگر روش ها یا فاکتورها در تعیین زیرمجموعه های وزنی، تعیین وزن‌های فاصله‌ای و غیره کار کرد.

مراجع

[۱] S. Marrara, G. Pasi and M. Viviani, "Aggregation operators in information retrieval," *Fuzzy Sets and Systems*, vol. ۳۲٤, pp. ۳-۱۹, ۲۰۱۷.

[۲] D. H. Kraft and E. Colvin, Fuzzy Information Retrieval, North Carolina: Morgan and Claypool, ۲۰۱۷.

[۳] D. H. Kraft, E. Colvin, G. Bordogna and G. Pasi, "Fuzzy information retrieval systems: A historical perspective," in *Fifty Years of Fuzzy Logic and its Applications*, Springer, Cham, ۲۰۱٥, pp. ۲٦۷-۲۹٦.

[٤] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of research and development*, vol. ۲, no. ۲, pp. ۱٥۹-۱٦٥, ۱۹٥۸.

[٥] P. Switzer, "Vector Images in Document Retrieval," in *Statistical Association Methods for Mechanized Documentation: Symposium Proceedings*, Washington, ۱۹٦٤.

[٦] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. ۲٤, no. ٥, pp. ٥۱۳-٥۲۳, ۱۹۸۸.

[۷] R. Cummins, *The evolution and analysis of term-weighting schemes in information retrieval*, Galway: National University of Ireland, ۲۰۰۸.

[۸] O. A. S. Ibrahim and D. Landa-Silva, "Term frequency with average term occurrences for textual information retrieval," *Soft Computing*, vol. ۲۰, no. ۸, pp. ۳۰٤٥-۳۰٦۱, ۲۰۱٦.

[۹] K. Goslin and M. Hofmann, "A Wikipedia powered state-based approach to automatic search query enhancement," *Information Processing & Management*, vol. ٥٤, no. ٤, pp. ۷۲٦-۷۳۹, ۲۰۱۸.

[۱۰] K. Chen, Z. Zhang, J. Long and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," *Expert Systems with Applications*, vol. ٦٦, pp. ۲٤٥-۲٦۰, ۲۰۱٦.

[۱۱] S. Plansangket, *New weighting schemes for document ranking and ranked query suggestion*, PhD diss., University of Essex, ۲۰۱۷.

[۱۲] D. Kandé, R. M. Marone, S. Ndiaye and F. Camara, "A Novel Term Weighting Scheme Model," in *Proceedings of the ٤th International Conference on Frontiers of Educational Technologies(ICFET ۱۸)*, Moscow, ۲۰۱۸.

[۱۳] T. Dogan and A. K. Uysal, "Improved inverse gravity moment term weighting for

www.SID.ir

Archive of SID

research and development, vol. ۱, no. ٤, pp. ۳۰۹-۳۱۷, ۱۹۵۷.

[۲۳] K. S. Jones, Information retrieval experiment, Newton, MA: Butterworth-Heinemann, ۱۹۸۱.

[۲٤] S. E. Robertson, "The probability ranking principle in IR," *Journal of documentation*, vol. ۳۳, no. ٤, pp. ۲۹٤-۳۰٤, ۱۹۷۷.

[۲۵] M. Sanderson and W. B. Croft, "The history of information retrieval research," *Proceedings of the IEEE*, vol. ۱۰۰, no. Special Centennial Issue, pp. ۱٤٤٤-۱٤۵۱, ۲۰۱۲.

[۲٦] H. F. Witschel, "Global term weights in distributed environments," *Information Processing & Management*, vol. ٤٤, no. ۳, pp. ۱۰٤۹-۱۰٦۱, ۲۰۰۸.

[۲۷] Y. Gupta, A. Saini and A. K. Saxena, "A new fuzzy logic based ranking function for efficient information retrieval system," *Expert Systems with Applications*, vol. ٤۲, no. ۳, pp. ۱۲۲۳-۱۲۳٤, ۲۰۱۵.

[۲۸] A. I. Kadhim, "Term Weighting for Feature Extraction on Twitter: A Comparison Between BM۲۵ and TF-IDF," in *International Conference on Advanced Science and Engineering (ICOASE)*, ۲۰۱۹.

[۲۹] C. Kamphuis, A. P. de Vries, L. Boytsov and J. Lin, "Which BM۲۵ Do You Mean? A Large-Scale Reproducibility Study of Scoring Variants," in *European Conference on Information Retrieval*, Cham, ۲۰۲۰.

[۳۰] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *In Proceedings of the ۲۱st annual international ACM SIGIR conference on Research and development in information retrieval*, ۱۹۹۸.

[۳۱] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM* text classification," *Expert Systems with Applications*, vol. ۱۳۰ , pp. ٤۵-۵۹, ۲۰۱۹.

[۱٤] S. Balbi, M. Misuraca and G. Scepi, "Combining different evaluation systems on social media for measuring user satisfaction," *Information Processing & Management*, vol. ۵٤, no. ٤, pp. ٦۷٤-٦۸۵, ۲۰۱۸.

[۱۵] H. Li, "Learning to rank for information retrieval and natural language processing," *Synthesis Lectures on Human Language Technologies*, vol. ٤, no. ۱, pp. ۱-۱۱۳, ۲۰۱۱.

[۱٦] S. Gugnani, T. Bihany and R. K. Roul, "A complete survey on web document ranking," *International Journal of Computer Applications (۹۷۵ ۸۸۸۷)*, vol. ICACEA, no. ۲, pp. ۱-۷, ۲۰۱٤.

[۱۷] A. H. Keyhanipour, M. Piroozmand and K. Badie, "A GP-adaptive web ranking discovery framework based on combinative content and context features," *Journal of Informetrics*, vol. ۳, no. ۱, pp. ۷۸-۸۹, ۲۰۰۹.

[۱۸] E. Goldberg, "Statistical machine". U.S. Patent ۱۸۳ ۸۳۸ ۹۲۹-۱۹۳۱, ۱۹۳۱.

[۱۹] J. E. Holmstrom, "Section III. Opening plenary session," in *The Royal Society Scientific Information Conference*, London, U.K., ۱۹٤۸.

[۲۰] H. F. Mitchell Jr, "The use of the univ AC FAC-tronic system in the library reference field," *American Documentation*, vol. ٤, no. ۱, pp. ۱٦-۱۷, ۱۹۵۳.

[۲۱] M. Taube, C. D. Gull and I. S. Wachtel, "Unit terms in coordinate indexing," *American Documentation*, vol. ۳, no. ٤, pp. ۲۱۳-۲۱۸, ۱۹۵۲.

[۲۲] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM Journal of*

www.SID.ir

Archive of SID

distribution," *Information Processing & Management,* vol. ٥٤, no. ٣, pp. ٣٧٠-٣٧٩, ٢٠١٨.

[٤١] R. Lakshmi and S. Baskar, "Novel Term Weighting Schemes for Document Representation based on Ranking of Terms and Fuzzy Logic with Semantic Relationship of Terms," *Expert Systems with Applications,* vol. ١٣٧, pp. ٤٩٣-٥٠٣, ٢٠١٩.

[٤٢] F. Carvalho and G. P. Guedes, *TF-IDFC-RF: A Novel Supervised Term Weighting Scheme,* arXiv preprint arXiv:٢٠٠٣,٠٧١٩٣, ٢٠٢٠.

[٤٣] W. B. Frakes and R. Baeza-Yates, Eds., Information retrieval: Data structures & algorithms, vol. ٣٣١, Englewood Cliffs, NJ: prentice Hall, ١٩٩٢.

[٤٤] G. Bordogna and G. Pasi, "Controlling retrieval through a user-adaptive representation of documents," *International Journal of Approximate Reasoning,* vol. ١٢, no. ٣-٤, pp. ٣١٧-٣٣٩, ١٩٩٥.

[٤٥] D. H. Kraft, G. Bordogna and G. Pasi, "An extended fuzzy linguistic approach to generalize Boolean information retrieval," *Information Sciences-Applications,* vol. ٢, no. ٣, pp. ١١٩-١٣٤, ١٩٩٥.

[٤٦] Y. Y. Yao, "Interval-set algebra for qualitative knowledge representation," in *Proceedings of ICCI'٩٣: ٥th International Conference on Computing and Information*, ١٩٩٣.

[٤٧] J. M. Mendel and D. Wu, Perceptual computing: Aiding people in making subjective judgments, vol. ١٣, John Wiley & Sons, ٢٠١٠.

[٤٨] J. Han, J. Pei and M. Kamber, Data mining: concepts and techniques, Elsevier, ٢٠١١.

[٤٩] A. Turpin and F. Scholer, "User performance versus precision measures for *Transactions on Information Systems (TOIS),* vol. ٢٢, no. ٢, pp. ١٧٩-٢١٤, ٢٠٠٤.

[٣٢] R. Cummins, *Modelling Word Burstiness in Natural Language: A Generalised Polya Process for Document Language Models in Information Retrieval,* arXiv preprint arXiv:١٧٠٨,٠٦٠١١, ٢٠١٧.

[٣٣] R. Cummins, J. H. Paik and Y. Lv, "A Pólya urn document language model for improved information retrieval," *ACM Transactions on Information Systems (TOIS),* vol. ٣٣, no. ٤, p. ٢١, ٢٠١٥.

[٣٤] G. Salton, Automatic Information Organization and Retrieval, New York: McGraw-Hill, ١٩٦٨.

[٣٥] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation,* vol. ٢٨, no. ١, pp. ١١-٢١., ١٩٧٢.

[٣٦] G. Salton and C.-S. Yang, "On the specification of term values in automatic indexing," *Journal of documentation,* vol. ٢٩, no. ٤, pp. ٣٥١-٣٧٢, ١٩٧٣.

[٣٧] G. Salton, A. Wong and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM,* vol. ١٨, no. ١١, pp. ٦١٣-٦٢٠, ١٩٧٥.

[٣٨] F. S. Al-Anzi, D. AbuZeina and S. Hasan, "Utilizing standard deviation in text classification weighting schemes," *Int J Innov Comput Inf Control,* vol. ١٣, no. ٤, pp. ١٣٨٥-١٣٩٨, ٢٠١٧.

[٣٩] J. Beel, S. Langer and B. Gipp, "Tf-iduf: A novel term-weighting scheme for user modeling based on users' personal document collections," in *iConference ٢٠١٧*, Wuhan, China, ٢٠١٧.

[٤٠] L. Bernauer, E. J. Han and S. Y. Sohn, "Term discrimination for text search tasks derived from negative binomial

www.SID.ir

Archive of SID

simple search tasks," in *In Proceedings of the ۲۹th annual international ACM SIGIR conference on Research and development in information retrieval*, ۲۰۰٦.

[۵۰] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval: the concepts and technology behind search, ۲ ed., Harlow: England: Pearson Education Ltd., ۲۰۱۱.

www.SID.ir

Archive of SID

www.SID.ir