دوره دوازدهم، شماره اول، بهار و تابستان ۱۳۹۸
مجله فناوری اطلاعات در طراحی مهندسی
Information Technology in Engineering Design
http://ited.sinaweb.net

بهبود نتایج موتور جستجوی مبتنی بر مدل بولی و برداری با استفاده از رویکرد معنایی

نرجس راز قندی * (۱) **حسن شاکری** (۲)

(۱) گروه مهندسی کامپیوتر، واحد سبزوار، دانشگاه آزاد اسلامی، سبزوار، ایران.*

(۲) گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران.

(تاریخ دریافت : ۱۳۹۷/٤/٤ تاریخ پذیرش : ۱۳۹۸/۱۰/۹)

چکیده

برای تسهیل جستجو در وب، موتورهای جستجویی طراحی‌شده‌اند که کاربر را دررسیدن به پاسخ موردنظر یاری می‌دهند. مدل بولی و مدل برداری دو مدل بهینه در موتور جستجو هستند که در تولید نتایج مرتبط با نیازمندی کاربر تأثیر بسزایی دارند. در این مقاله برای بهبود نتایج حاصل از موتور جستجو، در ابتدا برای کشف دقیق‌تر نیاز کاربر، روابط معنایی بین کلمات پرس‌وجوی کاربر در نظر گرفته می‌شود. سپس موتور جستجو، اسناد مرتبط با پرس‌وجوی کاربر را با استفاده از ترکیب مدل بولی و برداری می‌یابد. در مرحله بعد اسناد یافت شده بر اساس میزان ارتباطشان با پرسش با فن TF-IDF رتبه‌بندی می‌شوند و بالاخره لیست بلندی از اسناد مرتبط که بیشترین امتیازها را کسب نموده‌اند به‌عنوان پاسخ به کاربر بازگردانده می‌شوند. نتایج حاصل از این روش پیشنهادی حاکی از آن است که روش پیشنهادی ازنظر دقت و کیفیت نتایج جستجو نسبت به تحقیقات قبلی ارائه‌شده در این زمینه، بهبود قابل‌ملاحظه‌ای را نشان می‌دهد.

واژه‌های کلیدی: موتور جستجو، مدل بولی، مدل برداری، روابط معنایی

* عهده‌دار مکاتبات:
نشانی: گروه مهندسی کامپیوتر، واحد سبزوار، دانشگاه آزاد اسلامی، سبزوار، ایران.
تلفن: ۰۹۳٦٦٤۳٤۲۳۰ پست الکترونیکی: *narges_razghandi@yahoo.com*

۱. مقدمه

امروزه موتورهای جستجو از ابزارهای ضروری برای پیدا کردن اطلاعات روی صفحات وب هستند [۱]. در موتور جستجو کاربر کلیدواژه خود را وارد می‌کند، سپس برنامه جستجو با جستجو در بانک اطلاعاتی خود سایت‌های مرتبط با موضوع کاربر را نمایش خواهند داد [۲].

یکی از موتورهای جستجویی که توسط محققان برای مطالعات و آزمایش‌های پژوهشی مورداستفاده قرار می‌گیرد، موتور جستجوی علم و فنّاوری است. این موتور جستجو ابتدا اطلاعات منتشرشده روی وب سایت‌های دانشگاهی و صفحات شخصی محققان را جمع‌آوری می‌کند، سپس این اطلاعات را پردازش و خلاصه کرده و به کاربران ارائه می‌دهد. موتور جستجو در ابتدا باید منابع اطلاعاتی و مستندات وب را از طریق نرم‌افزاری به نام کاوشگر وب جمع‌آوری کند. کاوشگر در مدل کلی صفحات مربوط به سایت‌ها را درخواست می‌کند و در صورت مجاز بودن، صفحات را برای شاخص‌گذاری به واحد مدیریت ذخیره‌سازی می‌دهد تا در مخزن قرار گیرند [۳]. تمام اطلاعات جمع‌آوری‌شده توسط کاوشگر وب بعد از طی مراحل ذخیره‌سازی در مخزن در اختیار شاخص‌گذار قرار می‌گیرد [٤] .

در این بخش، اطلاعات ارسالی مورد تجزیه‌وتحلیل قرار می‌گیرد. ازآنجاکه حضور همه کلمات متن در شاخص‌گذاری سربار زیادی برای سیستم دارد، به نظر می‌رسد یکی از مشکل‌ترین فعالیت‌ها درروند شاخص‌گذاری انتخاب کلیدواژه‌هایی است که نشان‌دهنده محتویات سند می‌باشد .برای استخراج کلمات کلیدی یک سری پیش‌پردازش‌هایی باید روی متن انجام بگیرد. یکی از این پیش‌پردازش‌ها تعیین کلمات است. پس از تعیین کلمات، کلمات عمومی را حذف کرده و بقیه متن را ریشه‌یابی می‌کنیم و سپس کلمات را وزن دهی کرده و تبدیل به بردار می‌کنیم و با اعمال آستانه، لیست کلمات کلیدی استخراج می‌شود .شاخص‌گذاری یکی از قسمت‌های اساسی و مهم در موتورهای جستجو است به‌طوری‌که شاخص‌گذاری مناسب می‌تواند تأثیر قابل‌توجهی در بالا بردن کارایی موتور جستجو داشته باشد [۵].

یکی از نکات اصلی که برای کاربر اهمیت زیادی دارد نحوه رتبه‌بندی نتایج به‌دست‌آمده توسط موتور جستجوگر است. تفاوت در کارایی موتورهای جستجو ناشی از الگوریتم‌ها و مدل‌های مختلفی است که در این قسمت از موتور جستجو پیاده‌سازی شده‌اند. یکی دیگر از نکات این مدل‌ها، رفتار متفاوت آن‌ها در زبان‌های مختلف و مجموعه اسناد مختلف است. به این معنی که مدل‌های بازیابی اطلاعاتی که در موتورهای جستجو به‌منظور یافتن مشابه‌ترین سند به پرسش کاربر از میان اسناد موجود استفاده می‌شود، برخی از مهم‌ترین آن‌ها مدل فضای برداری و مدل بولی هستند. ما هم مشابه موتور جستجوی علم و فنّاوری که از دو مدل فضای برداری و مدل بولی استفاده می‌کند و آن‌ها را با یکدیگر ترکیب می‌کند، از ترکیب این دو مدل استفاده می‌کنیم. این مدل‌ها با توجه به مجموعه داده‌های مورداستفاده و زبان مقصد کارایی متفاوتی دارند [٦]. در مدل بولی، نیاز اطلاعاتی کاربر به‌صورت عبارتی منطقی با عملگرها and ،or و not بیان می‌شود و هر سندی که این عبارت در مورد آن صحیح باشد بازیابی می‌شود [۷و۸]. در مدل برداری، هر سند را به‌صورت برداری از کلمات در نظر می‌گیریم [۹].

یکی از پرکاربردترین روابط در حوزه بازیابی اطلاعات، پارامتر TF-IDFمی‌باشد که از حاصل‌ضرب فراوانی کلمه در فراوانی معکوس سند به دست می‌آید [۱۰]. دلیل مقبولیت این روش نسبت به سایر روش‌ها را

می‌توان با توجه به سهولت در استفاده از این روش محاسبات کم و نتایج قابل‌قبول دانست [۱۱].

در این مقاله راهکار جدیدی برای بهبود دقت نتایج جستجو در موتورهای جستجو ارائه می‌کنیم که از ایده TF-IDF مبتنی بر رویکرد معنایی استفاده می‌کند. دلیل اینکه رویکرد معنایی را به TF-IDF اضافه می‌کنیم، این است که TF-IDF محدودیتی دارد و آن این است که ارتباط و یکی بودن کلمات مترادف و به‌طورکلی کلماتی را که با یکدیگر ارتباط معنایی دارند را در نظر نمی‌گیرد. ما سعی می‌کنیم با مشخص کردن سه نوع ارتباط معنایی بین کلمات از قبیل رابطه مترادف بودن،HAS-A و IS-A با استفاده از این رویکرد بر این مشکل غلبه کرده و دقت موتور جستجو را افزایش دهیم. مزیت این روش نسبت به دیگر روش‌های استفاده‌شده در پژوهش‌های گوناگون این است که نه‌تنها به محاسبه‌ی وزن کلمه مورد جستجو کاربر پرداخته، بلکه وزن و ارزش کلماتی که تشابه معنایی با کلمات مورد جستجوی کاربر دارند هم در نظر گرفته می‌شود.

ساختار ادامه این مقاله به شرح زیر است: در بخش ۲ پیشینه تحقیق شامل فن‌های ارائه‌شده برای بهبود نتایج جستجو، موردبررسی قرار می‌گیرد. در بخش ۳ راهکار پیشنهادی برای افزایش دقت موتور جستجو ارائه می‌شود.. بدین منظور برای بهبود نتایج موتور جستجوی مبتنی بر مدل بولی و برداری آن را با استفاده از رویکرد معنایی بسط و توسعه می‌دهیم. در بخش ٤ روش پیشنهادی مورد ارزیابی قرار می‌گیرد و نتایج حاصل از این ارزیابی‌ها با تحقیقات قبلی ارائه‌شده در این حوزه مقایسه می‌شود و درنهایت در بخش‌های ۵ و ٦ به ترتیب نتیجه‌گیری مقاله و ایده‌های پیشنهادی برای کارهای آینده مطرح می‌شود.

۲. کارهای پیشین انجام‌شده

در زمینه بهبود دقت و صحت نتایج جستجو توسط موتورهای جستجو کارهای تحقیقاتی متعددی صورت گرفته است. در این بخش، برخی از مهم‌ترین پژوهش‌های ارائه‌شده در این زمینه را موردبررسی قرار می‌دهیم.

طبق مرجع [۱] یکی از چالش‌های موجود در موتور جستجو کشف توضیح دقیقی از نیاز کاربر است، زیرا معمولاً کاربران پرس‌وجوهای کوتاه و مبهمی ارسال می‌کنند و به همین دلیل در بسیاری از موارد برای پرسش کاربر پاسخی غیر مرتبط ارائه می‌شود. در این حال همان‌طور که در مرجع [۱]اشاره‌شده است محققان از فن‌های بازخورد ربط استفاده می‌کنند. در بازخورد ربط، فهرستی از پرسش‌های مرتبط شده با پرسش اولیه کاربر نمایش داده می‌شود که می‌تواند برای کاربر مفید واقع شود. برای مثال اگر کاربر در کلیه موتورهای جستجو عبارت "modonna" را جستجو کند پرس‌وجوهای مرتبطی از قبیل " modonnamp3","modonnamusic" ,"modonnalyrics" به او ارائه خواهد شد. این بازخورد برای اصلاح پرس‌وجوی اولیه و دریافت بهتر نیازهای کاربر به کار می‌رود. در اینجا کاربر می‌تواند به‌صراحت پرس‌وجوی خود را اصلاح کند، ابهامات احتمالی را از بین ببرد و پرس‌وجوی خود را به یک موضوع دیگر که مربوط به پرس‌وجوی اولیه است هدایت کند. در این فن از قوانین انجمنی استفاده می‌شود که موجب می‌شود نتایج خوبی به کاربر نشان داده شود و دقت موتور جستجو بیشتر شود. ایده این روش این است که الگوی جستجوی قبلی را پیدا می‌کند و با جستجوی فعلی مطابقت می‌دهد تا از این اطلاعات برای نشان دادن پیشنهادها مرتبط استفاده کند.

یکی دیگر از فن‌هایی که محققان در مرجع [۱۲]برای رفع چالش فوق ارائه داده‌اند، چارچوبی است که توسط خوشه‌بندی پرس‌وجو به بهبود موتور جستجو کمک می‌کند. در این تحقیق جانسن و همکاران به این نتیجه رسیدند که بیشتر پرس‌وجوها کوتاه هستند.(یعنی در هر پرس‌وجو به‌صورت میانگین حدوداً دو اصطلاح استفاده‌شده است). این نتیجه به‌دست‌آمده ممکن است ناشی از این باشد که کاربران اطلاعات کمی در مورد موضوع مدنظر خوددارند حتی در بدترین حالت ممکن است آن‌ها ندانند که به دنبال جستجوی چه چیزی هستند. بااین‌حال، کاربران نیاز به آشنایی با اصطلاحات خاصی در حیطه موردنظر خوددارند تا بتوانند پرس‌وجوی مؤثرتری را از موتور جستجو درخواست نمایند. از این فن‌ها برای کشف اطلاعاتی در رابطه با پرسش‌های مرتبط استفاده می‌شود اما موتورهای جستجو برای انجام رتبه‌بندی باکیفیت عالی باید از فن‌های پیشرفته‌تری استفاده کنند. به‌عبارتی‌دیگر آن‌ها باید قادر به پردازش ده‌ها هزار پرس‌وجو روی ده‌ها هزار صفحه باشند، بنابراین پرس‌وجو یک مسئله مهم تلقی می‌شود.

کار تحقیقاتی ارائه‌شده در مرجع [۱۳] برای به دست آوردن اسناد مرتبط با پرسش کاربر تنها فن TF-IDF را موردبررسی قرارداده است. در بخش‌های مختلف مقاله خود، در رابطه با این پارامتر به‌عنوان قسمتی از راهکار پیشنهادی صحبت خواهیم کرد و نتیجه ارزیابی حاصل از این کار تحقیقاتی صورت گرفته را با رویکرد خود موردقیاس قرار خواهیم داد.

۳. روش پیشنهادی

در این روش ارائه‌شده با انجام مرحله پیش‌پردازش اصطلاحات مهم واصلی شناسایی می‌شوند، سپس توسط چند رابطه معنایی کلمات مترادف و هم‌معنی با این اصطلاحات را کشف می‌کنیم. بعد از طی کردن این مراحل با استفاده از مدل بولی اسنادی را که شامل این اصطلاحات می‌باشند بازیابی کرده و درنهایت با درنظرگرفتن مدل برداری به رتبه‌بندی این اسناد و ارائه نتایج مورد انتظار کاربر می‌پردازیم.

اولین فاز، مرحله پیش‌پردازش است. قبل از فرآیند اصلی جستجو باید عملیات پیش‌پردازشی شامل نشانه‌گذاری، توقف کردن و ریشه‌یابی روی اسناد را انجام دهیم که هرکدام را به شکل خلاصه تعریف می‌کنیم. نشانه‌گذاری روشی است برای جدا کردن کلمات، نمادها و عناصر معنی‌دار دیگر که نشانه نامیده می‌شوند. متوقف شدن فرآیندی است که شامل حذف کلمات پرتکرار می‌باشد. ریشه‌یابی فرآیندی برای به دست آوردن ریشه کلمات است. البته پیش‌پردازش موارد فرعی دیگری را نیز در برمی‌گیرد. نمونه‌های دیگری که برای استخراج کلمات کلیدی بهتر در یک متن استفاده می‌شوند عبارت‌اند از :

نرمال‌سازی

نرمال‌سازی از دو جنبه مطرح می‌شود:

۱. ازلحاظ ظاهری به این معنا که کلمات مشابه که ازنظر ظاهر متفاوت‌اند را به ظاهری یکسان تبدیل نماییم. منظور از تفاوت‌های ظاهری، کوچک و بزرگ بودن حروف در زبان انگلیسی است. لذا ابتدا برای نمایش‌های مختلف یک کلمه، کلاس واحدی در نظر گرفته می‌شود. سپس هر زمان که این کلاس در نقاط مختلف متن تکرار شد، یکی از نمایش‌ها که از قبل انتخاب‌شده است جایگزین کلاس می‌شود. به‌عنوان‌مثال، کلمات FriEnd وfriend و Friend همه مشابه یکدیگر ولی در ظاهر متفاوت‌اند. هرکدام از آن‌ها در قسمت از متن که مشاهده شد، با کلمه‌ی friendجایگزین خواهد شد.

۲. ازلحاظ مفهومی نرمال‌سازی مفهومی شامل عملیات ریشه‌یابی است که در بالا مفهوم ریشه‌یابی را توضیح دادیم.

حذف برخی از عناصر ازنظر ظاهری

این مرحله شامل حذف تمام علائم نقطه‌گذاری در متن است. علائم نقطه‌گذاری مانند نقطه، ویرگول و علامت سؤال اطلاعات کمی را در اختیار ما قرار می‌دهند و در فرآیند بازیابی، نویز محسوب می‌شوند.

به‌این‌ترتیب با حذف آن‌ها، اطلاعات زیادی از دست نخواهیم داد. در عوض آنچه از متن باقی می‌ماند فقط کلمات هستند.

در فاز دوم، روابط معنایی بین کلمات شناسایی و استخراج می‌شود. منظور از روابط معنایی مطالعه و درک ارتباط واژه‌ها با یکدیگر و منطق حاکم بر این ارتباط است. در حقیقت، روابط معنایی می‌خواهد با درک نیت کاربر از طریق معنای کلمات، دقت نتایج جستجو را افزایش دهد و نتایج بهتری را پیش روی کاربر بگذارد. برای این منظور یافتن روابط بین کلمات از اهمیت ویژه‌ای برخوردار است. برای پیدا کردن ارتباط بین کلمات از شبکه واژگان استفاده می‌کنیم. شبکه واژگآنیکی از مباحثی است که در چند سال اخیر به‌شدت موردتوجه قرارگرفته است. وردنت نام عمومی است که بر شبکه‌های واژگانی مختلفی برای بسیاری زبان‌های جهان اطلاق می‌شود. این شبکه‌ها عموماً در نقش واژه ستان شناسی و یا واژگان معنایی محاسباتی در خدمت دستگاه‌های هوشمند دانش‌پایه و معناگرا قرار دارند.

شبکه واژگان به دلیل دارا بودن یک پایگاه داده غنی از لغات و روابط بین آن‌ها به‌طور مؤثری استفاده می‌شود. این بانک اطلاعاتی اسم‌ها، فعل‌ها، صفت‌ها و قیدها را به مجموعه‌ای از لغات مترادف دسته‌بندی می‌نماید که هر دسته یک مفهوم مجزا را بیان می‌کند. مجموعه مترادف‌ها با استفاده از روابط معنایی و ارتباطات لغوی به یکدیگر پیوند داده‌شده‌اند. شبکه به‌دست‌آمده که شبکه‌ای است از لغات و مفاهیم مرتبط ازلحاظ معنایی، می‌تواند توسط مرورگرها پیمایش شود.

در راهکار پیشنهادی ما مشخصاً سه نوع ارتباط معنایی بین کلمات تعیین می‌شود:

رابطه مترادف بودن: این رابطه اصلی‌ترین رابطه معنایی مورداستفاده است که در واکشی و مرتب‌سازی نتایج جستجو مورداستفاده قرار می‌گیرد. به‌عنوان‌مثال دو کلمه «اتومبیل» و «خودرو» به‌عنوان مترادف یکدیگر در نظر گرفته می‌شوند و جستجو به دنبال هر یک، باید اسناد شامل دیگری را هم ارائه کند.

رابطه IS-A: در علم بازنمایی و شیءگرایی، رابطه IS-A عبارت است از یک رابطه‌ای بین دو مفهوم یا کلاس که در آن‌یک مفهوم زیر کلاس مفهوم دیگر است. به‌عنوان نمونه کلمه «خودرو» رابطه IS-A با «وسیله نقلیه» دارد چون خودرو زیرکلاسی از کلاس وسیله نقلیه است.

رابطه HAS-A: یک رابطه ترکیبی است که یک شیء که جزء نامیده می‌شود، به‌عنوآن‌یک بخش یا عضو به شیء دیگری که شیء مرکب نامیده می‌شود، تعلق دارد. به‌عنوان‌مثال «فرمان» جزئی از اجزای «اتومبیل» است بنابراین بین این دو مفهوم رابطه HAS-A وجود دارد.

در سومین فاز از روش پیشنهادی بر اساس رویکرد بولی، اسنادی که کلمات مورد جستجوی کاربر در آن‌ها موجود است پیدا و استخراج می‌شود. این فرآیند باهدف کاهش فضای جستجو و یافتن کلیه اسناد حاوی کلمات موردنظر صورت می‌گیرد و بدون وزن دهی یا رتبه‌بندی خاصی همه اسنادی را که کلمات کاربر را در بردارند، استخراج می‌کند تا در فاز بعدی پردازش بیشتر جهت انتخاب نهایی و رتبه‌بندی آن‌ها صورت گیرد. بالاخره در فاز چهارم با استفاده از رویکرد برداری نتایج نهایی جستجو به‌صورت مرتب‌شده استخراج می‌شود. برای این

منظور معیار TF-IDF اسناد حاصل از فاز سوم هم برای کلمات موردنظر کاربر و هم برای کلیه کلماتی که با کلمات مذکور یکی از سه رابطه مترادف بودن، IS-A و یا HAS-A داشته باشند، محاسبه می‌شود و بر اساس آن اسناد مرتب و در خروجی ارائه می‌گردد. وزن کلمات بر اساس رابطه (۱) به دست می‌آید[۱۵و۱۶].

$$TF.IDF=(TermFrequency*InverseDocument\ Frequenc) \quad (۱)$$

معیار TFبه معنای فراوانی تکرار کلمه موردنظر در یک سند است [۱۷]. بدیهی است که هر چه فرکانس تکرار یک کلمه در یک سند بیشتر باشد، آن سند امتیاز بیشتری برای حضور در نتایج جستجو خواهد داشت.

IDF یا فرکانس معکوس سند به معنای فراوانی تکرار کلمه در کل اسناد است. مثلاً کلماتی مانند «و»، «در»، «از» و به‌طورکلی حروف ربط و حروف اضافه بااینکه مکرراً در یک سند تکرار شده‌اند، اما چون در تمام اسناد دیگر هم تکرار شده‌اند، ارزشی ندارند. وزن یک کلمه tk در یک سند dj با درنظرگرفتن یک مؤلفه محلی مربوط به سند و یک مؤلفه سراسری مربوط به کلیه اسناد تعیین می‌گردد[۱۸و۱۹]. فاکتور محلی یا TF متناظر است با فراوانی تکرار کلمه tk در سند dj یعنی تکرار دفعاتی که کلمه مذکور در سند موردنظر وجود دارد. فاکتور سراسری با استفاده از فرکانس معکوس سند یا IDF تخمین زده می‌شود. این معیار به‌صورت رابطه(۲)تعریف می‌شود:

$$IDF(tk) = \log(N/nk) \quad (2)$$

که در رابطه فوق، N تعداد کل اسناد شاخص‌گذاری شده و nk تعداد اسنادی است که حداقل یک‌بار حاوی کلمه tk هستند [۲۰و۲۱].

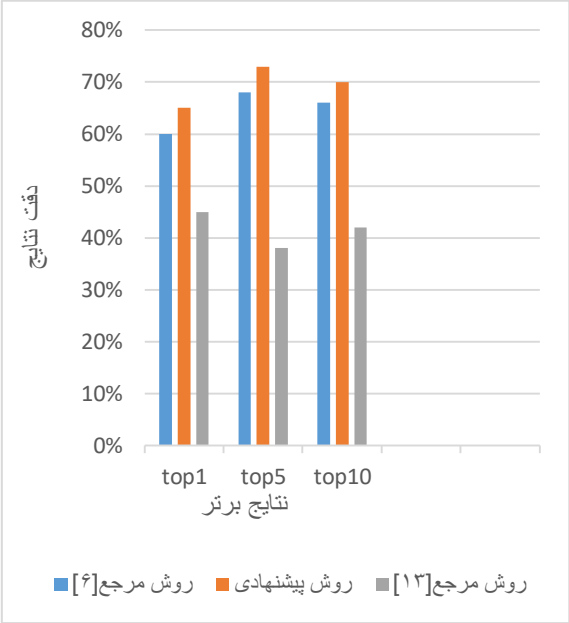
به‌طورکلی در این فضا هر کلمه‌ای که وزن بیشتری داشته باشد ارزش بیشتری دارد و موتور جستجو را در رساندن کاربر به سند مرتبط یاری می‌دهد. درنهایت موتور جستجو با استفاده از این فن اسناد مرتبط و رتبه‌بندی شده‌ای را در اختیار کاربر قرار می‌دهد. ما برای ارزیابی روش پیشنهادی خود از۶۰ سند مربوط به حوزه‌های هوش مصنوعی، مهندسی نرم‌افزار، سرویس وب، کارگزار- هوشمند، منطق و چندین مبحث مرتبط به این عناوین استفاده کرده‌ایم که برای دریافت نتایج مرتبط‌تر، برای هر عنوان مربوطه ده سند در نظر گرفته‌ایم بعد از درنظرگرفتن این داده‌ها فازهای روش پیشنهادی مطرح‌شده را بر روی آن‌ها اعمال کرده و نتایج حاصل‌شده از این رویکرد را با نتایج حاصل از مرجع [۶و۱۳] مقایسه می‌کنیم.

٤. ارزیابی روش پیشنهادی

در این بخش تمام فازهای روش پیشنهادی بر روی اسناد جمع‌آوری‌شده اعمال می‌شود و ثابت می‌شود که روش پیشنهادی نتایج برتری نسبت به روش ارائه‌شده در مراجع دیگر تولید می‌کند. درنهایت با آنالیز زمان پاسخ و بار اضافی حاصل از افزایش تعداد کلمات مورد جستجو رویکرد پیشنهادی را مورد ارزیابی قرار می‌دهیم.

٤-۱ مقایسه نتایج حاصل از روش پیشنهادی با نتایج حاصل از روش مراجع دیگر

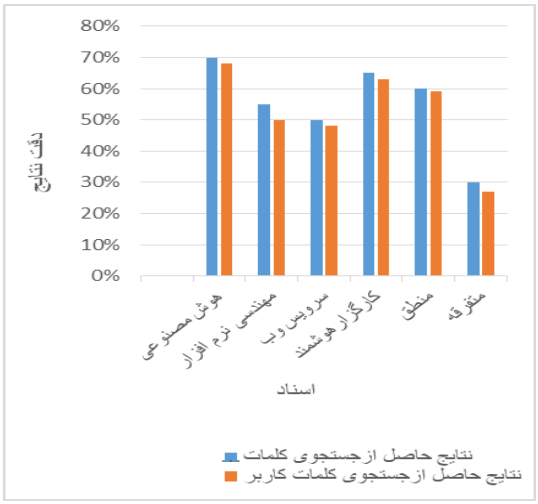
نتایج حاصل‌شده در این بخش، از اعمال فازهای مختلف روش پیشنهادی بر روی اسناد جمع‌آوری‌شده به‌دست‌آمده است که در شکل ٤-۱ نتایج حاصل از راهکار پیشنهادی ما ازنظر دقت در مقایسه با راهکار ارائه‌شده در مراجع [۶و۱۳] نشان داده‌شده است.



شکل ٤-١ مقایسه دقت نتایج جستجو در راهکار پیشنهادی در مقایسه با راهکار ارائه‌شده در مراجع [٦و١٣]

همانطورکه در شکل ٤-١ مشاهده می‌شود روش پیشنهادی ما از روش ارائه شده در مراجع [٦و١٣] مناسب‌تر می‌باشد زیرا این روش با درنظرگرفتن روابط معنایی بین کلمات، دقت موتور جستجو را افزایش داده و باعث می‌شود اسناد مرتبط‌تری با پرسش کاربر پیدا شود وبرترین نتایج تحت عنوان top1و top5 و top10 در اختیار کاربر قرار بگیرد

٤-٢ بار اضافی به وجود آمده از بررسی روابط معنایی



شکل٤-٢ مقایسه بار در صورت اعمال یا عدم اعمال روابط معنایی درروش پیشنهادی

همان‌طور که قابل پیش‌بینی بود، با اعمال رویکرد معنایی بار موتور جستجو کمی افزایش‌یافته است، که این افزایش بار افزایش زمان پاسخ را هم به همراه دارد، اما این افزایش فقط در حدود ده درصد است و در برابر بهبود دقت حاصل از این رویکرد، قابل‌چشم‌پوشی است.

٥. نتیجه‌گیری

در این پژوهش روشی برای بهبود نتایج موتور جستجو ارائه شد. در مدل پیشنهادی از رویکرد روابط معنایی استفاده شد که باعث می‌شود که کاربر به سندی که مدنظر دارد نزدیک‌تر شود و آن را بازیابی کند. در این رویکرد کلماتی که با پرس‌وجوی کاربر شباهت معنایی داشتند، استخراج شدند و در جایگاه کلمات ارزشمند قرار گرفتند. درنهایت موتور جستجو را در هدایت کاربر به سمت هدف موردنظر یاری دادند. با ارزیابی و آزمایشی که بر روی روش پیشنهادی اعمال کردیم و مقایسه‌ای که با روش‌های ارائه‌شده در مراجع [٦و١٣] انجام دادیم به این نتیجه رسیدیم که روابط معنایی درنظرگرفته شده درروش پیشنهادی، سبب بالاتر رفتن دقت موتور جستجو می‌شود. قابل‌ذکر است که با توجه به ارزیابی به‌عمل‌آمده رویکرد معنایی افزایش بار اضافی کلمات و زمان پاسخ را به همراه خواهد داشت. البته این افزایش زمان پاسخ به حدی نیست که ما را از تأثیر مثبت رویکرد معنایی در ارائه اسناد مرتبط با پرس‌وجوی کاربر دورنگه دارد.

٦. کارهای آتی

یکی از ایده‌های پیشنهادی برای ادامه این پژوهش استفاده از روابط معنایی گسترده‌تر برای بهبود دقت

نتایج جستجو می‌باشد. همچنین این امکان وجود دارد که از فن‌های یادگیری ماشین و روش‌های بهینه‌سازی فرا ابتکاری و الهام گرفته از طبیعت استفاده شود، که این فن‌ها باعث بهبود دقت انتخاب اسناد نهایی می‌شود. یک رویکرد دیگر ثبت رفتار کاربر در انتخاب و کلیک روی نتایج جستجوی کاربر در هر پرس‌وجو می‌باشد. نگهداری این اطلاعات در یک بانک اطلاعاتی به‌منظور شخصی‌سازی جستجو برای کاربران مختلف صورت می‌گیرد.

۷. مراجع

[1] Fonseca, Bruno M., Paulo Braz Golgher, Edleno Silva de Moura, and Nivio Ziviani. "Using association rules to discover search engines related queries." In Proceedings of the IEEE/LEOS 3rd International Conference on Numerical Simulation of Semiconductor Optoelectronic Devices (IEEE Cat. No. 03EX726), pp. 66-71. IEEE,(2003).

[2] Lee, Jihyun, et al. "Effective ranking and search techniques for Web resources considering semantic relationships." Information Processing & Management 50.1 (2014): 132-155.

[3] Wachsmuth, Henning, et al. "Building an argument search engine for the web." Proceedings of the 4th Workshop on Argument Mining. (2017).

[4] Martin, Carlstedt. "Using NLP and context for improved search result in specialized search engines." (2017).

[5] Chauhan, Ekta, and Amit Asthana. "Review of Indexing Techniques in Information Retrieval." International Journal of Engineering Science 13940 (2017).

[6] Armentano, Marcelo G., Daniela Godoy, Marcelo Campo, and Analia Amandi. "NLP-based faceted search: Experience in the development of a science and technology search engine." Expert Systems with Applications 41, no. 6 (2014): 2886-2896.

[7] Jayalakshmi, T., and C. Chethana. "A semantic search engine for indexing and retrieval of relevant text documents." Int. J 4, no. 5 (2016): 1-5.

[8] Hristovski, Dimitar, et al. "Biomedical question answering using semantic relations." BMC bioinformatics 16.1 (2015): 6.

[9] Meng, Zeyu, Dong Yu, and Endong Xun. "Chinese microblog entity linking system combining wikipedia and search engine retrieval results." Natural Language Processing and Chinese Computing. Springer, Berlin, Heidelberg, (2014). 449-456.
[10] Nesi, Paolo, Gianni Pantaleo, and Gianmarco Sanesi. "A Distributed Framework for NLP-Based Keyword and Keyphrase Extraction From Web Pages and Documents." DMS. (2015).
[11] Kumari, Madhu, Akshat Jain, and Ankit Bhatia. "Synonyms Based Term Weighting Scheme: An Extension to TF. IDF." Procedia Computer Science 89 (2016): 555-561.
[12] Baeza-Yates, Ricardo, Carlos Hurtado, and Marcelo Mendoza. "Improving search engines by query clustering." Journal of the Association for Information Science and Technology 58, no. 12 (2007): 1793-1804.
[13] Ramos, Juan. "Using tf-idf to determine word relevance in document queries." In Proceedings of the first instructional conference on machine learning, vol. 242, pp. 133-142. (2003).
[14] Zhang, Jiangong, Xiaohui Long, and Torsten Suel. "Performance of compressed inverted list caching in search engines." In Proceedings of the 17th international conference on World Wide Web, pp. 387-396. ACM, (2008).
[15] Boudin, Florian, Hugo Mougard, and Damien Cram. "How Document Pre-processing affects Keyphrase Extraction Performance." arXiv preprint arXiv:1610.07809 (2016).
[16] Baquerizo, R., P. Leyva, J. Febles, Hubert Viltres, and Vivian Estrada Sentí Sala. "Algorithm for calculating relevance of documents in information retrieval systems." (2017).
[17] Khan, Sharifullah, and Jibran Mustafa. "Effective semantic search using thematic similarity." Journal of King Saud University-Computer and Information Sciences 26.2 (2014): 161-169.
[18] Elhadad, Mohamed K., KhaledM Badran, and Gouda I. Salama. "A novel approach for ontology-based dimensionality reduction for web text document classification." In Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference on, pp. 373-378. IEEE, (2017).
[19] Ram, Naik Ramesh, and C. Namrata Mahender. "Marathi WordNet Development." International Journal Of Engineering And Computer Science ISSN: 2319–7242 3, no. 8 (2014): 7622-7624.
[20] GU, Vasanthakumar, Vanitha Raj KC, Asha Rani BR, P. Deepa Shenoy, and Venugopal KR. "PTMIBSS: PROFILING TOP MOST INFLUENTIAL BLOGGER USING SYNONYM SUBSTITUTION APPROACH." ICTACT Journal on Soft Computing 7, no. 2 (2017).
[21] Markonis, Dimitrios, et al. "The Parallel Distributed Image Search Engine (ParaDISE)." arXiv preprint arXiv:1701.05596(2017).