بهبود کارایی یک معماری موتور جستجوی معنایی با استفاده از تشابه معنایی و در نظر گرفتن نزدیکی مفهومی

زهراجعفری[۱]، حمید رستگاری[*,۲]

۱- دانشکده مهندسی کامپیوتر، واحد نجف آباد، دانشگاه آزاد اسلامی ،نجف آباد ، ایران

۱- دانشکده مهندسی کامپیوتر، واحد نجف آباد، دانشگاه آزاد اسلامی ،نجف آباد ، ایران.

خلاصه

با پیشرفت وب و گسترش مطالب موجود در آن و استفاده از آن اهمیت موتورهای جستجو روز به روز بیشتر می‌شود. از طرف دیگر سطح انتظارات و نیازهای کاربران نیز روز به روز در حال گسترش است و کاربران تمایل به استفاده از سرویس‌هایی دارند که نه تنها درخواست آنها را بررسی کرده و پاسخ دهند، بلکه درخواست آنها را درک و تجزیه و تحلیل کرده و مرتبط ترین موارد را یافت کنند. به همین دلیل بحث استفاده از موتورهای جستجوی معنایی روز به روز پر اهمیت تر می‌شود. در این مقاله در جهت بهبود کارایی بازیابی اسناد مرتبط با نیازمندی‌های کاربر روشی پیشنهاد شده است که بر اساس آن علاوه بر درنظر گرفتن کلمات کلیدی و مفاهیم بر اساس اطلاعات زبان شناسی و تشخیص حوزه سند، تشابه معنایی کلمات(کلمات کلیدی مشابه) و نزدیکی مفهومی با استفاده از اطلاعات RDF در نظر گرفته شده است. روش پیشنهادی بر اساس چهارچوب معماری جدید SBIRS است و مؤلفه‌های این معماری با اضافه شدن دیتاست RDF و مؤلفه تشابه معنایی در راستای بهبود آن در نظر گرفته شده است. برای نشان دادن نزدیکی مفهومی از اطلاعات دیتاست RDF و اسناد استخراج شده مرتبط با حوزه آن و همچنین برای تشابه معنایی کلمات نیز از اطلاعات اسناد استفاده شده است. در تشابه معنایی از ماشین بردار پشتیبان برای یادگیری کلمات مشابه و نامشابه استفاده شده است. آزمایش‌ها روی قسمتی از اسناد استخراج شده مرتبط با حوزه RDF و جمع آوری شده از اینترنت انجام شده است تا بتوان از کارایی روش پیشنهاد شده اطمینان حاصل کرد و تحلیل درستی از نتایج ارائه داد. آزمایشات نشان می‌دهد که با استفاده از روش پیشنهادی به دلیل در نظر گرفتن بهتر ارتباط‌های معنایی و مفهومی بین کلمات در مقایسه با معماری پایه SBIRS معیارهای فراخوانی و معیار F بررسی گردیده و این نتیجه حاصل شدکه فراخوانی و معیار F افزایش بیشتری داشته است.

واژه های کلیدی: وب معنایی، جستجو معنایی، استخراج مفاهیم، شباهت معنایی، RDF، شباهت مفهومی

[*]Corresponding author Islamic Azad University, Najaf Abad Branch, Faculty of Computer Engineering,Zahra Jafari,09139591368
Email:taranom1225@gmail.com

۱. مقدمه

وجود موتورهای جستجویی که بتوانند بهترین پاسخ را به درخواست کاربر بدهند امری ضروری است با توجه به اینکه تعداد صفحات وب موجود در اینترنت دائماً در حال افزایش می‌باشد و پیدا کردن اسناد مناسب برای کاربران مشکل است.این واقعیت که امروزه نیاز به استفاده از موتورهای جستجوی پیشرفته برای یافتن صفحات وب در بین حجم عظیم صفحات وجود دارد واقعیتی انکار ناشدنی است. از طرفی بنا به دلایل متعدد موتورهای جستجوی سنتی نمی‌توانند همواره پاسخ‌های بهینه را ارائه کنند. از آن جمله می‌توان به عدم استفاده کاربر از کلید واژه‌های صحیح در بسیاری از جستجوها و یا عدم تطابق کلمات اشاره کرد. از این رو ضرورت استفاده از موتور جستجوی معنایی که با تجزیه و تحلیل درخواست کاربر و صفحات بتواند مرتبط‌ترین صفحات را ارائه کند ،محسوس است.

موتورهای جستجوی معنایی یکی از مواردی هستند که از لغت شناسی و امکانات جدیدی که وب در اختیار قرار داده است، استفاده می‌کنند. جستجوی کارامد، یکی از کاربردهای پیش روی نسل بعدی وب است که تحت عنوان وب معنایی شناخته می‌شود [۱].

روش‌های موجود جستجوی وب، اسناد را به عنوان یک مجموعه کلمه می بینند و تنها به تکرار و وجود آنها توجه می کنند. موتور جستجو های پیش رو دیگر محدود به تطبیق کلمات کلیدی بین درخواست کاربر و اسناد نیستند و سعی می کنند با استفاده از اطلاعات و آنالیز بیشتر جوابی دقیق تر و ساخت یافته تر به دست آورند. در این صورت دیگر تطبیق دقیق درخواست کاربر و کلمات کلیدی نیاز نبوده و موتور جستجو با تحلیل منظور کاربر، خود بهترین جوابها را پیشنهاد خواهد داد [۲].می‌توان در نظر گرفت که هر صفحه وب حاوی اطلاعات مفهومی است که می‌تواند برای بررسی بیشتر خود صفحه در نظر گرفته شود. این اطلاعات مفهومی که معمولا تحت عنوان لغت شناسی شناخته می شوند می توانند در تحلیل صفحات وب کمک بسزایی انجام دهند. به عنوان مثال در تشخیص مرتبط بودن موضوع یک صفحه وب با درخواست کاربر می توان از این اطلاعات استفاده کرد. ایده اصلی این مدل این است که در خصوص مربوط بودن یک مدل با توجه به مدل فکری کاربر که از اطلاعات به دست آمده است استفاده شود. در این صورت سندی که دقیقا منطبق بر درخواست کاربر است به عنوان سند هسته در نظر گرفته می‌شود و با بسط مفاهیم و بررسی های بیشتر سایر اسناد نیز استخراج می‌شوند [۳].

اما روشی که در اینجا استفاده شده است علاوه بر درنظر گرفتن کلمات کلیدی و مفاهیم بر اساس اطلاعات زبان شناسی و تشخیص حوزه سند، تشابه معنایی کلمات و نزدیکی مفهومی با استفاده از اطلاعات RDF را در نظر گرفته است. چهارچوب روش بر اساس معماری جدید SBIRS است دیتاست RDF استاندارد و مؤلفه تشابه معنایی در راستای بهبود آن گرفته شده است. برای استخراج نزدیکی مفهومی از اطلاعات دیتاست RDF استفاده شده است و در تشابه معنایی از ماشین بردار پشتیبان برای یادگیری کلمات مشابه و نامشابه استفاده شده است.

۲. معماری SBIRS

معماری SBIRS برای بازیابی اطلاعات از وب معنایی از نمایش‌های مفهومی محتوا در راستای کلیدواژه‌ها به‌عنوان پایگاه دانش استفاده می‌کند و نمایش‌های مفهومی نیازهای کاربر را ارائه می‌دهد[۴]. این معماری نمایش‌های مفهومی محتوا، بسط‌های پرس‌وجو، مطابقت معنایی، استخراج نتایج مرتبط بر اساس ارتباط مفاهیم را با کمک مؤلفه‌های پیمایشگر، پیش‌پردازشگر، مفسر معنایی، شاخص کننده معنایی، مبدل پرس‌وجوی معنایی، بازیاب کننده محتوای معنایی و رتبه دهنده معنایی انجام می دهد.این مؤلفه‌ها در لایه‌های مختلف معماری گروه‌بندی‌شده‌اند. لایه داده فیزیکی پایگاه داده وب را با مؤلفه‌های پیمایشگر و پیش‌پردازشگر ایجاد می‌کند. لایه مفسر معنایی پایگاه دانش را با مفسر معنایی و اندیس

معنایی ایجاد می‌کند. لایه مطابقت معنایی وظیفهٔ مطابقت بین محتوای معنایی و جستجوی معنایی را به عهده دارد. لایه بازیابی، نتایج بازیابی شده را رتبه‌بندی می‌کند و به لایه کاربرد تحویل می‌دهد.

۳. شباهت معنایی کلمات

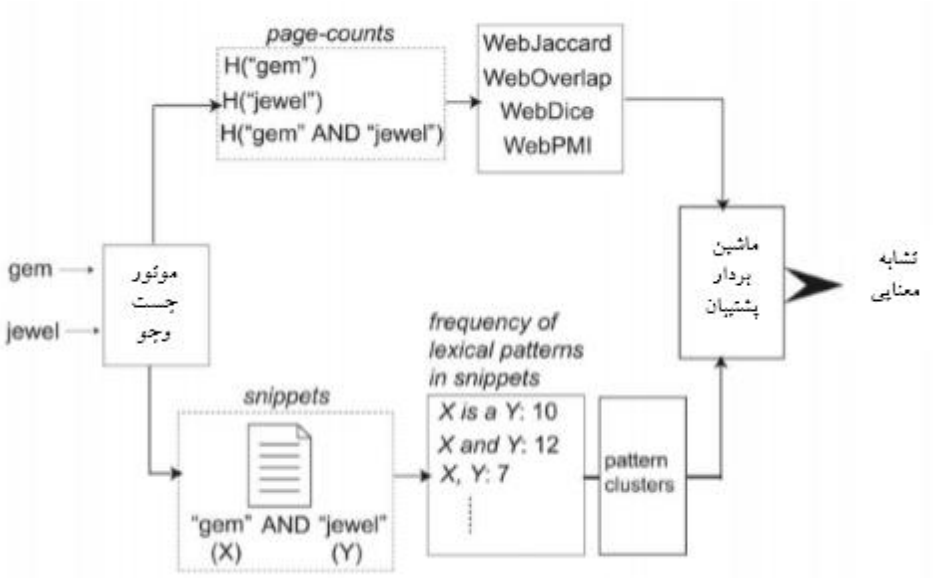
دو کلمه P و Q را در نظر بگیرید. ما مسئله اندازه‌گیری شباهت معنایی بین P و Q را به‌صورت ساختن یک تابع sim(P,Q) در نظر گرفته‌ایم، که مقداری را در رنج [۰، ۱] برمی‌گرداند. اگر P و Q به‌شدت مشابه باشند (مثلاً مترادف باشند)، ما انتظار داریم sim(P,Q) نزدیک به ۱ باشد. از طرف دیگر، اگر P و Q ازنظر معنایی شبیه نباشند، پس انتظار داریم sim(P, Q) به صفر نزدیک باشد. ما ویژگی‌های مختلفی که شباهت بین P وQ را بیان می‌دارند، با استفاده از شمارش صفحه و اسنیپت بازیابی شده از یک موتور جستجو برای دو کلمه داده‌شده تعریف می‌کنیم. با استفاده از این نوع نمایش ویژگی کلمات، ما یک ماشین بردار پشتیبانی دو-دسته‌ای* را جهت دسته‌بندی جفت واژه‌های مترادف و غیر مترادف آموزش می‌دهیم. سپس تابع sim(P, Q) با امتیاز اعتماد SVM آموزش داده‌شده تخمین زده می‌شود[۵].

شکل ۱ مثالی را جهت محاسبه شباهت معنایی بین دو کلمه gem و jewel نشان می‌دهد. ابتدا پرس‌وجوی خود را به موتور جستجو صادر می‌کنیم و شمارش صفحه را برای دو کلمه و برای عطف آن‌ها («gem»، «jewel» و « gem AND jewel») بازیابی می‌کنیم. در بخش بعدی ما چهار امتیاز شباهت با استفاده شمارش صفحه تعریف می‌کنیم. امتیازهای شباهت مبتنی برشمارش صفحه حضور همزمان کلی دو کلمه را در وب در نظر می‌گیرند. بااین‌وجود، آن‌ها متن محلی را که کلمات در آن به‌طور همزمان ظاهرشده‌اند، مدنظر قرار نمی‌دهند. از طرف دیگر، اسنیپت های برگردانده شده توسط یک موتور جستجو متن محلی که دو کلمه در آن ظاهرشده‌اند را ارائه می‌دهند. درنتیجه، ما میزان تکرار الگوهای معنایی لغوی متعددی را در اسنیپت ها برای پرس‌وجوی عطف دو کلمه پیدا می‌کنیم.

الگوهای لغوی به‌صورت خودکار استخراج می‌شوند. بااین‌وجود، لازم است به این نکته اشاره‌کنیم که یک رابطه معنایی می‌تواند با استفاده از چند الگوی لغوی بیان شود. گروه‌بندی الگوهای لغوی مختلف که رابطه معنایی یکسانی را حمل می‌کنند، ما را قادر به نمایش یک رابطه معنایی دقیق بین دو کلمه می‌سازد. آن‌ها برای این منظور، یک الگوریتم خوشه‌بندی الگوی متوالی ارائه نموده‌اند. هم امتیازات شباهت مبتنی برشمارش و هم خوشه‌های الگوی لغوی جهت تعریف ویژگی‌های مختلفی که رابطه بین دو کلمه را نمایش می‌دهند، مورداستفاده واقع‌شده‌اند. با استفاده از نمایش ویژگی جفت کلمات، یک ماشین بردار پشتیبانی دو کلاسه را آموزش داده‌اند.

* Two-class

شکل۱- نمایش کلی روش شباهت معنایی



۱،۳. معیارهای حضور همزمان مبتنی برشمارش صفحه

شمارش صفحه برای پرس‌وجوی P AND Q را می‌توان به‌صورت یک تقریب از وقوع همزمان دو کلمه (دو عبارت چندکلمه‌ای) P و Q در وب در نظر گرفت. بااین‌وجود، شمارش صفحه برای پرس‌وجوی P AND Q به‌تنهایی شباهت معنایی را به‌صورت دقیق بیان نمی‌دارد. برای مثال، گوگل ۱۱،۳۰۰،۰۰۰ کلمه را به‌عنوان شمارش صفحه برای پرس‌وجوی "car"AND"automobile" برمی‌گرداند. درحالی‌که برای پرس‌وجوی "car" AND "apple" این مقدار برابر با ۴۹۰۰۰۰۰۰ است. باوجوداینکه واژه automobile ازنظر معنایی شباهت بیشتری به واژه car دارد، اما شمارش صفحه‌ای که برای پرس‌وجوی "car" AND "apple" برگردانده شده است، تقریباً ۴ برابر بیشتر از شمارش صفحه پرس‌وجوی "automobile" AND "car" است. باید شمارش صفحه را نه‌تنها برای پرس‌وجوی P AND Q، بلکه برای P و Q به‌صورت جداگانه نیز در نظر گرفت تا بتوان شباهت معنایی بین P و Q را ارزیابی کرد.

بر این اساس چهار معیار وقوع همزمان رایج؛ جاکارد*، همپوشانی†، دایس‡ و اطلاعات دوجانبه نقطه به نقطه§ را جهت محاسبه شباهت معنایی با استفاده از شمارش صفحه در نظر گرفته‌شده است. در ادامه از H(P) برای نشان دادن شمارش صفحه برای پرس‌وجوی P در یک موتور جستجو استفاده‌شده است. ضریب جاکاردِ وب** بین دو کلمه (یا دو عبارت چندکلمه‌ای) P و Q، به‌صورت زیر تعریف می‌شود:

$$WebJaccard(P,Q) = \begin{cases} 0 & if\ H(P \cap Q) \leq C \\ \frac{H(P \cap Q)}{H(P)+H(Q)-H(P \cap Q)} & Otherwise \end{cases} \quad (۱)$$

در رابطه بالا، P ∩Q پرس‌وجوی عطف P AND Q را نشان می‌دهد. با در نظر گرفتن نویز و مقیاس در داده‌های وب، ممکن است دو کلمه باوجوداینکه رابطه‌ای ندارند در برخی صفحات همزمان حضور پیدا کنند. جهت کاهش تأثیرات منفی

---

*

† Overlap

‡ Dice

§ Pointwise mutual information (PMI)

** WebJaccard

چنین حضورهای همزمانی، در حالتی که شمارش صفحه برای پرس‌وجوی $P \cap Q$ کمتر از یک آستانهٔ c باشد، ضریب WebJaccard را برابر با صفر قرار داده‌ایم. به‌طور مشابه، WebOverlap به‌صورت زیر تعریف‌شده است:

(۲)

$$Web\ Overlap\ (P,Q) = \begin{cases} 0 & if\ H(P \cap Q) < C \\ \dfrac{H(P \cap Q)}{\min(H(P), H(Q))} & Otherwise \end{cases}$$

WebOverlap یک تغییر طبیعی در ضریب همپوشانی است. ما ضریب WebDice را به‌صورت گونه‌ای از ضریب Dice تعریف کرده‌ایم. WebDice به‌صورت زیر تعریف‌شده است:

$$Web\ Dice(P,Q) = \begin{cases} 0 & if\ H(P \cap Q) < C \\ \dfrac{2*H(P \cap Q)}{H(P)+H(Q)} & Otherwise \end{cases}$$ (۳)

اطلاعات دوجانبه نقطه‌به‌نقطه معیاری است که از تئوری اطلاعات نشأت‌گرفته است؛ جهت بازتاب وابستگی بین دو رخداد احتمالی ارائه‌شده است. ما WebPMI را به‌صورت گونه‌ای از معیار اطلاعات دوجانبه نقطه‌به‌نقطه با استفاده از شمارش صفحه تعریف کرده‌ایم:

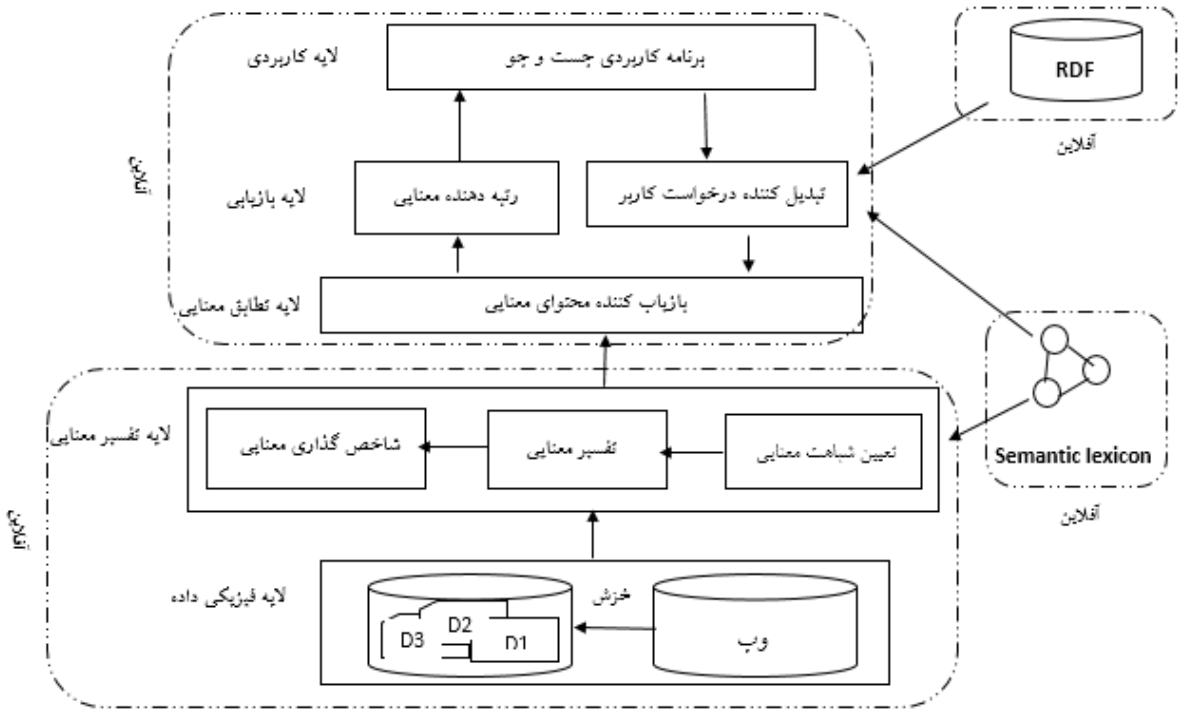
(٤)

$$Web\ PMI(P < Q) = \begin{cases} 0 & if\ H(P \cap Q) < C \\ log_2\left(\dfrac{\frac{H(P \cap Q)}{N}}{\frac{H(P)}{N}\frac{H(Q)}{N}}\right) & otherwise \end{cases}$$

۴. معماری موتور جستجوی بهبود داده شده

معماری بهبود داده شده سعی در اکتشاف حوزه اسناد و درخواست کاربر با استفاده از لغتنامه معنایی، استخراج نزدیکی معنایی کلمات از خود کلمات اسناد وب و نزدیکی مفهومی کلمات با استفاده از RDF را دارد. در این معماری پس از کشف نزدیکی معنایی و مفهومی در فرآیند جستجو، کلمات کلیدی استخراج‌شده را نیز مدنظر قرار می‌دهد تا بتواند دقت و کارایی جستجو را بهبود ببخشد. به‌علاوه با در نظر گرفتن نزدیکی حوزه اسناد و درخواست کاربر سعی می‌شود که اسنادی که به حوزه مدنظر کاربر نزدیک‌تر هستند در رتبه بالاتری نمایش داده شوند. در ادامه، ابتدا به توضیح موتور جستجوی مورداستفاده و معماری آن پرداخته شده است پس‌ازآن روش اکتشاف نزدیکی معنایی کلمات و نزدیکی مفهومی کلمات توضیح داده‌شده و شیوه به‌کارگیری آن در معماری پیشنهادی توضیح داده شده است. طرح کلی روش پیشنهادی در شکل ۲ نشان داده‌شده است.

شکل۲- طرح کلی معماری بهبود داده شده



۴.۱.دریافت اطلاعات

اولین مرحله در هر موتور جستجو، دریافت اطلاعات صفحات وب است. برای دریافت اطلاعات صفحات وب دو راهکار وجود دارد: راه اول پیاده‌سازی یک نرم‌افزار پیمایشگر* برای کاوش وب و استخراج اطلاعات هر یک از صفحات وب و راه دوم استفاده از دیتاست‌های موجود که قبلاً استخراج و پالایش‌شده هستند، است. به دلیل وجود دیتاست‌ها، در این پایان‌نامه نرم‌افزار پیمایشگر پیاده‌سازی نشده است و ازیک دیتاست RDF استاندارد به همراه اسناد استخراج شده مرتبط با آن استفاده شده است.

۴.۲. پیش‌پردازش

در این مرحله اطلاعات اسناد پردازش می‌شوند و داده‌های اضافی از آن‌ها حذف می‌شوند. ازجمله مواردی که در این مرحله انجام می‌شوند شامل:

حذف کلمات اضافی مانند am, is, are, has, have, ….

ریشه‌یابی کلمات مانند حذف ing یا s یا es از انتهای کلمات

* Crawler

۴.۳. تحلیل مفهومی و استخراج کلمات کلیدی

پس از استخراج اسناد نوبت به آنالیز مفهومی و استخراج کلمات کلیدی هر یک از آن‌ها می‌رسد. کلمات کلیدی با شناسایی کلمات منحصربه‌فردی که تکرارشان زیاد است، مشخص می‌شوند. به‌عبارت‌دیگر هر کلمه‌ای که تعداد تکرارش در سند از حد آستانه‌ای بیشتر باشد به‌عنوان کلمه کلیدی آن سند در نظر گرفته می‌شود. برای آنالیز مفهومی از دو نوع از داده‌ها استفاده می‌شود:

لغتنامه معنایی*

قالب توضیح منابع

لغتنامه معنایی به کلماتی گفته می‌شود که می‌توانند حوزه یک متن را مشخص کنند. به‌عنوان‌مثال در مورد حوزه بانکداری می‌توان به چک، سفته، وام، شماره‌حساب، موجودی حساب اشاره کرد. به‌این‌ترتیب درصورتی‌که یکی از این کلمات در متن یک سند و یا درخواست کاربر موجود باشد می‌توان حوزه مدنظر را مشخص کرد و به‌این‌ترتیب می‌توان با داشتن این اطلاعات به پالایش بهینه نتایج قابل‌نمایش برای کاربر پرداخت.

در اینجا برای بهبود موتور جستجو از آنتولوژی استفاده نموده‌ایم. روش‌های مختلفی برای آنالیز مفهومی با استفاده از آنتولوژی وجود دارند. در این خصوص کارهای وسیعی انجام‌شده است که به برخی از آن‌ها اشاره‌شده است. در مورد آنتولوژی کارهایی که انجام‌شده اغلب بر پایه دایره لغت است.

در کاربردی جدیدتر استفاده از لغتنامه معنایی پیشنهادشده است. به‌طوری‌که واژگانی وجود دارند که در صورت استفاده در یک متن خاص می‌توانند زمینه متن را با دقت خوبی نشان دهند. به‌عنوان‌مثال کلمات «وام، چک، سفته، موجودی » می‌توانند نشان دهند که متن موردنظر مربوط به موارد بانکی است. به‌این‌ترتیب با داشتن این دایره واژگان معنایی می‌توان ارتباط مفهومی بین درخواست کاربر و اسناد را به‌طور دقیق‌تری در نظر گرفت. برای این منظور در این پایان‌نامه از لغتنامه معنایی مرتبط با اسناد استفاده‌شده است و درصورتی‌که درخواست کاربر و اسناد کلمه خاصی از این واژگان را مشترک داشته باشند، با یک وزن بیشتر در امتیازدهی آن سند در بالاتر نشان داده می‌شود. به این دلیل که تشخیص داده می‌شود که این سند در حوزه مدنظر کاربر قرار دارد. توجه شود که لزوماً نباید کلمه درخواست کاربر و کلمه موجود در سند یکی باشند، بلکه مربوط به یک حوزه باشند کفایت می‌کند. به‌این‌ترتیب که ممکن است که کاربر کلمه وام» را در درخواست خود وارد کرده باشد و سند دارای کلمه «سفته» باشد. در این صورت نیز تشخیص داده می‌شود که حوزه یکسان است و سند مذکور در رتبه‌بندی بالاتر نشان داده می‌شود.

از سوی دیگر قالب توضیح منابع هم اطلاعات آنالیز جملات را در خود دارد که با استفاده از آن می‌توان اطلاعات بیشتر و دقیق‌تری از سند را در دست داشت.

۴.۴.استخراج شباهت معنایی کلمات

برای استخراج شباهت معنایی کلمات، روش پیشنهادی به این صورت کار می‌کند که ازآنجاکه بررسی دوبه‌دوی ارتباط تمام کلمات با یکدیگر آن‌هم در حجم داده‌های وب بسیار زمان‌بر است، ما سعی کردیم که با ارائه روشی تنها ارتباط دوبه‌دوی قسمتی از کلمات را مقایسه کنیم. به این منظور با این پیش‌فرض که کلماتی که معنای نزدیکی داشته باشند

* Semantic Lexicon
† Resource Description Framework
‡ Dictionary

به‌احتمال‌زیاد دریکی از متن‌های صفحات وب در کنار یکدیگر آمده‌اند و با یک کلمه عطف و یا فصل جداشده‌اند به استخراج تمامی کلماتی که عطف و یا فصلی بین آن‌ها بوده پرداختیم. جدول ۴-۱ نشان‌دهنده کلمات عطف و فصل مورداستفاده در این تحقیق را نشان می‌دهد.

پس از به دست آوردن کلمات، یک سری ملاک بین دو کلمه در بین تمام اسناد موجود به دست می‌آید. این پارامترها عبارت‌اند از:

تعداد دفعاتی که عطف بین دو کلمه آمده است

تعداد کلماتی که فصل بین دو کلمه آمده است

جدول ۱- کلمات عطف و فصل

نوع	لیست
کلمات عطف	yet, so, or, because, even, both,and, either
کلمات فصل	Not, Nor

ملاک‌های حضور کلمات باهم در اسناد مختلف شامل ملاک جاکارد، همپوشانی کلمات، دایس و PMI می باشد. در روابط ۱ تا ۴ به ترتیب فرمول ملاک های جاکارد، همپوشانی کلمات، دایس و PMI بیان شده است.

پس از استخراج این پارامترها، آن‌ها را به SVM برای آموزش می‌دهیم. به‌این‌ترتیب که SVM تلاش می‌کند که ارتباط معنایی این کلمات را با توجه به این شش ویژگی فرابگیرد. برای این منظور پس از استخراج اطلاعات کلمات از اسناد موجود، SVM را با برخی از کلماتی که خود میدانیم هم‌معنی هستند و برخی کلماتی که میدانیم هم‌معنی نیستند آموزش داده تا بعد بتواند در خصوص سایر کلمات تصمیم‌گیری کند.
تا این قسمت شباهت معنایی کلمات استخراج شد و نیز سعی شد که حوزه مفهومی درخواست کاربر و اسناد نیز بررسی شود.

۴.۵.شاخص کننده معنایی

در این مرحله وزن مربوط به هر یک از کلمات کلیدی و مفاهیم در هر سند مشخص می‌شود. این وزن‌ها tf وidf برای هر یک از کلمات کلیدی و مفاهیم است. tf تعداد تکرار هر کلمه کلیدی یا مفهوم در آن سند است. idf* تعداد تکرار هر کلمه کلیدی یا مفهوم در تمام اسناد است. این اطلاعات برای محاسبه رتبه اسناد به مراحل بعدی ارسال می‌شوند.

* document frequency

۴.۶.ورود و تبدیل درخواست کاربر

در این مرحله درخواست کاربر به‌صورت جمله یا کلمات کلیدی از ورودی دریافت می‌شود. در ادامه باید کلمات کلیدی و مفاهیم و همچنین معانی مشابه و نزدیکی مفهومی استخراج شوند

در مرحله بعد برای بررسی دقیق‌تر ارتباط مفهومی کلمات از دیتاست RDF استفاده شده است. در این دیتاست اطلاعات به‌صورت مفاهیم و ارتباط بین آن‌ها است. لذا با استفاده از استخراج کلماتی که درجاهای متعدد دارای ارتباط مفهومی با یکدیگر هستند سعی شده است علاوه بر استخراج شباهت معنایی کلمات و در نظر گرفتن حوزه به استخراج ارتباط مفهومی کلمات و در نظر گرفتن این مورد نیز در جستجو پرداخته شود. به‌عنوان‌مثال جمله «من سیب خوردم» را اگر در نظر بگیریم در سه‌تایی RDF بر اساس قوانین پردازش زبان طبیعی جمله واکاوی شده و به استخراج «سیب»، «خوردن» می‌پردزد. پس ما می‌توان فهمید که سیب با خوردن ارتباط دارد و به‌عنوان‌مثال اگر کاربر به دنبال خوردنی بود یک سندی که به سیب و چند مورد مشابه دیگر اشاره می‌کند می‌تواند مدنظر کاربر باشد، درصورتی‌که این سند در نگاه اول و توسط روش‌های سنتی استخراج نمی‌شود.

۴.۷.تطابق معنایی

در این مرحله جدول درهمی با توجه به کلمات کلیدی و کلمات مشابه و همچنین مفاهیم و نزدیکی مفهومی(مفاهیم مشابه در صورت وجود) به‌دست‌آمده از درخواست کاربر در مرحله قبلی، ساخته می‌شود. تطابق بین درخواست کاربر و اسناد با توجه به کلمات کلیدی و مفاهیم استخراج‌شده از آن‌ها است. جدول درهم (hash table) برای سند iام به‌صورت زیر محاسبه می‌شود(۴-۱).

$$hash_i = \sum_{j=1}^{NQ} f1(w_{ij}) + \sum_{j=1}^{NQ} \sum_{w \in similarity\ of\ w_{ij}} f2(w) + \sum_{j=1}^{NS} f3(c_{ij}) + \sum_{j=1}^{NS} \sum_{s \in similarity\ of\ c_{ij}} f4(c) \quad (۵)$$

$$f1(w_{ij}) = \begin{cases} W_{keyword} & if\ keyword\ j\ is\ document\ i \\ 0 & otherwise \end{cases}$$

$$f2(w_{ij}) = \begin{cases} W_{simKeyword} & if\ keyword\ j\ is\ document\ i \\ 0 & otherwise \end{cases}$$

$$f3(c_{ij}) = \begin{cases} W_{semantic} & if\ semantic\ j\ is\ document\ i \\ 0 & otherwise \end{cases}$$

$$f4(w_{ij}) = \begin{cases} W_{simSemantic} & if\ semantic\ j\ is\ document\ i \\ 0 & otherwise \end{cases}$$

که NQ تعداد کلمات کلیدی، NS تعداد مفاهیم، Wkeyword وزن تأثیر کلمات کلیدی مستقیماً واردشده در توسط کاربر، WsimKeyword وزن تأثیر کلمات مشابه با کلمات کلیدی درخواست کاربر، Wsemantic وزن تأثیر مفاهیم استخراج‌شده از درخواست کاربر، WsimSemantic وزن تأثیر مفاهیم مشابه با مفاهیم درخواست کاربر می‌باشند.

درنهایت لیست اسناد قابل‌نمایش با آستانه گذاری بر روی ضرایب hash به دست می‌آید(۶) یعنی:

$$lst = \{i \mid hash_i > T, 1 \le i \le N\} \quad (٦)$$

که T حد آستانه و N تعداد اسناد است. به‌عبارت‌دیگر اگر تأثیر کلمات کلیدی و مفاهیم با نزدیکی معنایی و مفاهیمی آن‌ها از حد آستانه‌ای بیشتر باشد، آنگاه آن سند برای رتبه‌بندی به مرحله بعد ارسال می‌شود. به‌عبارتی‌دیگر در پایان این مرحله لیست اسناد کاندید برای نمایش مشخص می‌شوند.

۴.۸.رتبه دهنده معنایی

در این مرحله رتبه هر سند (که از مرحله قبلی ارسال‌شده است) تعیین می‌شود و سپس اسناد برحسب رتبه از زیاد به کم برای کاربر نمایش داده می‌شوند. در ابتدا معکوس فراوانی سند را تعریف می‌کنیم که به‌صورت (۷) محاسبه می‌شود:

$$idf = \log\frac{N}{df} \qquad (۷)$$

که df فراوانی سند است. وزن کلمه کلیدی wj یا مفهوم cj در سند Di به‌صورت (۸)محاسبه می‌شود:

$$wt_{ij}(w_{ij}) = tf(w_{ij}) * idf(w_i) \qquad (۸)$$
$$wt_{ij}(c_j) = tf(c_{ij}) * idf(c_j)$$

درنهایت ضریب شباهت* بین درخواست کاربر و سند وب iام توسط رابطه (۹) به دست می‌آید:

$$sr(Q, D_i) = \sum_{j=1}^{t} wt_{qj} * wt_{ij} \qquad (۹)$$

در این مرحله اسنادی که امتیازشان خیلی کم باشد (پایین‌تر از حد آستانه از پیش تعیین‌شد باشد) از لیست حذف می‌شوند. لیست نهایی با مرتب‌سازی اسناد بر اساس این ضریب شباهت به کاربر نمایش داده می‌شود.

۵. پیاده سازی و نتایج

۵.۱. پارامترهای پیکربندی

پارامترهای استفاده شده برای وزن دهی اسناد و مرتب سازی آنها در جدول ۲ خلاصه شده است .

جدول ۲- پارامترهای پیکربندی

پارامتر	مقدار	توصیف
$W_{keyword}$	۱	وزن تأثیر کلمات کلیدی مستقیماً واردشده در درخواست کاربر
$W_{simKeyword}$	۰.۸	وزن تأثیر کلمات مشابه با کلمات کلیدی درخواست کاربر
$W_{semantic}$	۱	وزن تأثیر مفاهیم استخراج‌شده از درخواست کاربر
$W_{simSemantic}$	۰.۷	وزن تأثیر مفاهیم مشابه با مفاهیم درخواست کاربر
T	۲.۵	حد آستانه حداقل تعلق درخواست کاربر به اسناد
T_2	۱.۸۵	حد آستانه حداقل امتیاز اسناد لیست نهایی
c	۰.۰۵	حد آستانه احتمال حضور همزمان دو کلمه در معیارهای مختلف

* similarity coefficient

۵.۲. نحوه بدست آوردن داده ها

یکی از راهکارهای به دست آوردن اطلاعات صفحات وب برای موتورهای جستجو استفاده از اطلاعات استخراج‌شده از صفحات وب توسط پیمایشگر است. از آنجا که روش پیشنهادی بر مبنای استفاده از اطلاعات RDF و صفحات وب است، برای اینکه اطلاعات استخراج‌شده از RDF را جهت بهبود جستجوی صفحات وب معمولی استفاده کرد، به این منظور به‌صورت دستی به استخراج یک سری از اسناد صفحات وب در این حوزه که تا حدی مرتبط به RDF مورداستفاده در ارزیابی می‌شد، اقدام شد. بر این اساس از RDF درزمینهٔ music استفاده کرده و به استخراج اسناد وب درزمینهٔ مربوطه و مرتبط با RDF استفاده‌شده پرداخته‌شده است و تعدادی سند وب در این حوزه جمع آوری شد که بتوان از دیتاست استاندارد RDF نیز در کنار آن استفاده شود.

۵.۳. معیارهای ارزیابی

برای بررسی تأثیر هر گام، اول باید ملاک‌های ارزیابی و داده‌های مورداستفاده مشخص شود. ازآنجاکه مهم‌ترین ملاک‌های ارزیابی موتورهای جستجو (چه سنتی و چه معنایی) سه ملاک دقت*، بازیابی و ارزیابی F هستند، ما نیز این سه ملاک را برای ارزیابی کارایی استفاده کردیم. ملاک دقت نشان‌دهنده این موضوع است که چه درصدی از نتایج نشان داده‌شده توسط موتور جستجو، نتایج صحیح و مدنظر می‌باشند رابطه (۱۰) ،ملاک بازیابی بیانگر این موضوع است که چند درصد از اسناد مرتبط توسط موتور جستجو بازیابی شده و نمایش داده‌شده‌اند رابطه (۱۱) .

درنهایت ملاک ارزیابی F بیانگر ترکیبی از دو ملاک گفته‌شده است رابطه (۱۲)

$$\text{Precision} = \frac{\text{No.of relevant documents retrieved}}{\text{Total No.of documents retrieved}} \qquad (۱۰)$$

$$\text{Recall} = \frac{\text{No.of relevant documents retrieved}}{\text{Total No.of documents relevant}} \qquad (۱۱)$$

$$F = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (۱۲)$$

که Precision دقت و Recall بازیابی است. از طرف دیگر هرچند مجموعه داده‌هایی برای بررسی قدرت موتورهای جستجو موجود است ولی در اکثر مقالات این حوزه ذکرشده است که بررسی دقیق کارایی تنها با کمک فرد خبره ممکن است. به‌نحوی‌که فرد خبره در خصوص تک‌تک نتایج ارائه‌شده توسط الگوریتم نظر بدهد و به‌این‌ترتیب کارایی محاسبه شود. به‌عبارت‌دیگر برخلاف بسیاری از حوزه‌های دیگر مجموعه داده‌ها دارای برچسبی که برای مقایسه و محاسبه کارایی قابل‌استفاده باشد، نمی‌باشند. به‌این‌ترتیب سعی شده است تا در خصوص هر یک از موارد تست شده به‌دقت کارایی محاسبه شود.

* Precision
† Recall
‡ F-measure

۵.۳. طرح آزمایش

در بررسی تأثیر گام‌های روش پیشنهادی، سه حالت مختلف در نظر گرفته می‌شود: موتور جستجوی SBIRS، استفاده از شباهت معنایی، استفاده از شباهت مفهومی و شباهت معنایی به‌این‌ترتیب تأثیر افزوده شدن هر گام را به مراحل قبلی با استفاده از ملاک‌های ذکرشده بررسی شده است برای ارزیابی از تعدادی درخواست تصادفی استفاده‌شده است. به‌علاوه برای دقت بهتر ارزیابی تعدادی از اسناد مورد آزمایش نیز توسط تعدادی فرد خبره بررسی‌شده و میزان نزدیکی آن‌ها به هر یک از درخواست‌ها گردآوری‌شده است.

جدول ۳- انطباق بخشی از اسناد مورد آزمایش و درخواست های تصادفی کاربر

درخواست کاربر	نتیجه بررسی فرد خبره و میزان ارتباط اسناد مورد آزمایش با درخواست کاربر
۱	Elvis-Aaron-Presley ElvisWorldrestaurantreviews (could be) PublicPerformanceRevised (10% related) group1a_gi-project09 (Semantically and conceptually related)
۲	Christian-hip hop music (40% related) music-assesshb2011-2015 Atton_Popular_Music_Rev3 group1a_gi-project09 (50%) gradek_music_sol_strategies(10% preferably unrelated) JSTOR_Project_MUSE_Titles (30%)
۳	Elvis Elvis Elvis - 2014 Spring Tour Press Release (60%) Elvis The King Tribute Cruise (70%) Elvis-Aaron-Presley Leman Book Review (30%) Pressetext (30%) elvis aint dead worksheet (10%) ELVIS PRESLY (80%) ElvisWorldrestaurantreviews (10%) activity 2 (80%) 0608ElvisGospelService-press (10%)
۴	Atton_Popular_Music_Rev3 gradek_music_sol_strategies magini (40%) Christian-hip hop music (70%) group1a_gi-project09 (20%)
۵	Elvis-Aaron-Presley PublicPerformanceRevised (10% related) group1a_gi-project09 (20% semantically related)
۶	Elvis-Aaron-Presley group1a_gi-project09 (40%) Sacredandprofaneicon-workOulu (50%) activity 2 (30%)
۷	Elvis-Aaron-Presley coyne606-1 (70%) Leman Book Review group1a_gi-project09 (30%)

۵,۴. روش SBIRS

در جستجوی پایه اسناد بر مبنای روش SBIRS مورد مقایسه قرارگرفته است. در این گام نتایج اجرای این روش بر روی مجموعه داده مورد تست نشان داده‌شده است. جدول ۴ مقادیر دقت و بازیابی نشان داده شده است.

جدول ۴- دقت و بازیابی روش SBIRS

درخواست کاربر	Precision	Recall	F-measure
۱	۱	۰.۲۵	۰.۳۳
۲	۱	۰.۳۳	۰.٤۹
۳	۰.۷۵	۰.۹۱	۰.۸۲
٤	۱	۱	۱
٥	۱	۰.۳۳	۰.٤۹
٦	۱	۰.۲	۰.۳۳
۷	۱	۰.٦	۰.۷٥

۵,۵.افزودن شباهت معنایی

با افزودن گام بررسی شباهت معنایی کلمات، موتور جستجو علاوه بر جستجوی کلمات واردشده توسط کاربر به جستجوی کلمات مشابه باکلمات کلیدی وارد شده توسط کاربر نیز می‌پردازد و به‌این‌ترتیب شانس یافتن اسناد موردنظر کاربر افزایش پیدا می‌کند. به‌این‌ترتیب بازیابی موتور جستجو افزایش یافته است.به دلیل بازیابی اسناد بیشتر با توجه به این نکته که سندی اگر دارای کلمه کلیدی مشابه با درخواست کاربر باشد نیز بازیابی می شود، دقت مشابه با روش پایه به دست آمده است .مقادیر دقت و بازیابی این حالت نیز در جدول ۵ نشان داده‌شده است.

جدول ۵- دقت و بازیابی وF-measure با در نظر گرفتن شباهت معنایی

درخواست	دقت Precision	بازیابی Recall	F-measure
۱	۱	۰.٤	۰.٤۹
۲	۱	۰.۸۳	۰.۹۰
۳	۰.۷۵	۰.۹۱	۰.۸۲
٤	۱	۱	۱
٥	۱	۰.۳۳	۰.٤۹
٦	۱	۰.٤	۰.٥۷
۷	۱	۰.۸	۰.۸۸

۵,۶.در نظر گرفتن شباهت معنایی و شباهت مفهومی

در گام بعد با استفاده از اطلاعات استخراج‌شده از RDF ها و در نظر گرفتن شباهت مفهومی کلمات سعی در بهبود بیشتر کارایی الگوریتم شده است. در این گام علاوه بر در نظر گرفته شدن حوزه درخواست کاربر و تطبیق با اسناد مرتبط و همچنین در نظرگرفته شدن اسنادی که شامل کلمات کلیدی مشابه هستند تلاش شده است تا نزدیکی مفهومی از درخواست کاربر و اطلاعات اسناد نیز با استفاده از اطلاعات RDF استفاده شده در نظر گرفته شود و با استفاده از این راهکار نیز در برخی از موارد کارایی افزایش پیداکرده است. نتایج این حالت نیز در جدول ۶ نشان داده‌شده است.

جدول ۹- دقت و بازیابی و F-measure با در نظر گرفتن شباهت مفهومی و شباهت معنایی

درخواست	دقت Precision	بازیابی Recall	F-measure
۱	۱	۰.٦	۰.۷۵
۲	۱	۰.۸۳	۰.۹۰
۳	۰.۷۵	۰.۹۱	۰.۸۲
٤	۱	۱	۱
٥	۱	۰.۳۳	۰.٤۹
٦	۱	۰.٦	۰.۷۵
۷	۱	۱	۱

۶. نتیجه گیری

در این مقاله, تلاش شد تا روشی بر مبنای در نظر گرفتن شباهت معنایی و نزدیکی مفهومی کلمات و نیز اطلاعات زبان‌شناسی برای جستجوی مفهومی صفحات وب ارائه شود. در این روش ابتدا درخواست کاربر به‌صورت متنی دریافت می‌شود، سپس این درخواست بسط داده شده و کلمات کلیدی و مفاهیم از آن استخراج می‌شود. همچنین در این مرحله کلمات کلیدی مشابه با درخواست کاربر (شباهت معنایی) و مفاهیم نزدیک به آن (نزدیکی مفهومی) نیز استخراج می‌شوند. کلمات مشابه، کلماتی هستند که بین آن‌ها کلمات عطف یا فصل وجود داشته باشد و شباهت معنایی توسط ذخیره و آموزش ماشین بردار پشتیبان با کمک کلمات مشابه و غیرمشابه به دست می‌آید. نزدیکی مفهومی توسط RDF انجام می‌شود که شامل اطلاعات مفهومی است.این روش سعی در بهبود روش SBIRS برای یافتن کلمات کلیدی مشابه و در نظر گرفتن مفاهیم مشابه در اسناد به منظور بالابردن معیار فراخوانی و یافتن اسناد مرتبط بیشتر با نیازمندی های کاربر دارد.

مراجع

1. E. Motta, and M. Sabou, "Next generation semantic web applications," The Semantic Web–ASWC 2006, pp. 24-29: Springer, 2006.
2. B. Aleman-Meza, B. Arpinar, M. V. Nural, and A. P. Sheth, "Ranking documents semantically using ontological relationships." pp. 299-304.
3. V. Jindal, S. Bawa, and S. Batra, "A review of ranking approaches for semantic search on Web," Information Processing & Management, vol. 50, no. 2, pp. 416-425, 2014.
4. M. Thangaraj, and G. Sujatha, "An architectural design for effective information retrieval in semantic web," Expert Systems with Applications, vol. 41, no. 18, pp. 8225-8233, 2014.
5. Bollegala, D., Y. Matsuo, and M. Ishizuka, "A web search engine-based approach to measure semantic similarity between words". Knowledge and Data Engineering, IEEE Transactions on,Vol 23(7), pp 977-990, 2011.