DOR :20.1001.1.2538161.1399.5.4.3.0

تاریخ دریافت مقاله: ۹۸/۰۶/۱۴
تاریخ پذیرش مقاله: ۹۹/۰۲/۰۶

جاسازی عبارت پرس‌وجو بر اساس مدل‌سازی موضوعی

مریم بیابانی
دانشجوی کارشناسی‌ارشد- دانشکده مهندسی و علوم کامپیوتر- دانشگاه شهید بهشتی- تهران- ایران
پست الکترونیکی: m.biabani@mail.sbu.ac.ir

احمدعلی آبین*
استادیار- دانشکده مهندسی و علوم کامپیوتر- دانشگاه شهید بهشتی- تهران- ایران
پست الکترونیکی: a_abin@sbu.ac.ir

چکیده

افزایش حجم منابع متنی سبب شده است تا اهمیت فرآیند جستجو بیش از پیش در حوزه بازیابی اطلاعات آشکار گردد. این فرآیند امر چالش‌برانگیزی است زیرا در بسیاری موارد کلمات پرس‌وجوی به کار رفته توسط کاربران با کلمات موجود در متون تفاوت دارد و یا از عباراتی برای پرس‌وجو استفاده می‌شود که فرآیند جستجو را گمراه می‌نماید. بسیاری از روش‌های پیشین، عبارت پرس‌وجو را به‌عنوان کیسه‌ای از کلمات در نظر گرفته و محل قرار گرفتن و یا معنی کلمات را لحاظ نمی‌کنند. اخیرا روش‌های مبتنی بر جاسازی کلمات کارایی خود را در بسیاری از کاربردهای بازیابی اطلاعات به‌طور موثر نشان داده‌اند. در این تکنیک، بردار جاسازی‌شدهٔ عبارت پرس‌وجو از ترکیب بردار جاساز کلمات حاصل می‌شود. از این بردار جهت ادامهٔ فرایند جستجو استفاده می‌شود. در این پژوهش روشی مبتنی بر مدل‌سازی موضوعی برای جاسازی عبارت پرس‌وجو ارائه شده‌است که با لحاظ کردن موضوعات موجود در عبارات پرس‌وجو، عبارت پرس‌وجو را فارغ از تعداد و محل کلمات به یک نقطه در فضای جدیدی نگاشت کرده و از آن‌ها جهت پردازش‌های بعدی استفاده می‌نماید. روش پیشنهادی بر روی مجموعهٔ آزمون Stack Overflow ارزیابی و تحلیل شده‌است. نتایج به‌دست‌آمده نشان دهنده افزایش دقت روش ارائه‌شده در مقایسه با روش‌های موجود است.

واژه های کلیدی: بازیابی اطلاعات، جاسازی بردار، عبارت پرس‌وجو، سیستم‌های پاسخ به پرسش.

مقدمه

امروزه نمایش برداری کلمات یا به عبارتی دیگر جاسازی کلمات، گستره وسیعی از پژوهش های موجود در حوزه بازیابی اطلاعات و پردازش زبان طبیعی را تحت تاثیر خود قرار داده است و در بسیاری موارد نیز کارایی خود را نشان داده است. نمایش توزیع شده کلمات بر مبنای بردار چگال که به آن جاسازی کلمات گفته می‌شود کاربردهای بسیاری در پردازش زبان طبیعی و بازیابی اطلاعات دارد. برای نمونه می‌توان به کاربردهای جاسازی کلمات در بسط عبارت پرس‌وجو و تحلیل فاصله واژگانی اشاره کرد [۱، ۲].

* نویسندهٔ مسئول

تکنیک‌های مبتنی بر جاسازی کلمات، هر کلمه را به یک بردار در فضای برداری جدید نگاشت می‌کند به گونه‌ای که در این فضا کلماتی که از لحاظ معنایی یا نحوی به یکدیگر شبیه هستند به یکدیگر نزدیک‌تر هستند. روش‌هایی مانند Glove [٤]، Word2Vec [٣]، Bert [٦]و FastText [٥] را می‌توان از نمونه‌های موفق و پرکاربرد در این زمینه دانست.

چگونگی تبدیل عبارت پرس‌وجو به یک بردار در واقع اساسی‌ترین سوال مطرح در مورد تکنیک‌های مبتنی بر جاسازی عبارت است که امری چالش برانگیز می‌باشد. چرا که همان‌طور که در مطالعه [٧] ذکر شده است، عبارات پرس‌وجو در زمان آموزش بردار کلمات در دسترس نیستند و یا ممکن است عبارات پرس‌وجو شامل کلماتی باشند که در یک متن با هم ظاهر نشده‌اند. یکی از روش‌های پایه در حوزه جاسازی عبارت پرس‌وجو، گرفتن میانگین وزن‌دار از بردارهای کلمات عبارت پرس‌وجو می‌باشد که این بردار کلمات می‌توانند به کمک هر تکنیکی از قبل تولید شده باشند [٨، ٩، ١٠، ١١، ١٢]. از معایب این روش‌ها می‌توان این‌گونه عنوان کرد که ممکن است بردار تک تک کلمات عبارت پرس‌وجو در دسترس نبوده و یا بردار حاصل از میانگین‌گیری از لحاظ معنایی، معنای خاصی نداشته باشد. لذا در این پژوهش سعی داریم روشی ارائه دهیم که برای هر عبارت یک بردار متمایز نتیجه دهد به گونه‌ای که با تقریب خوبی نشان دهندهٔ معنای عبارت نیز باشد.

از کاربردهای جاسازی عبارت پرس‌وجو می‌توان به برچسب گذاری مطالب اشاره کرد. امروزه با گسترش وبگاه‌های مبتنی بر پرسش و پاسخ و شبکه‌های تعاملی، برچسب‌گذاری مناسب مطالب یک مسئلهٔ مهم و اساسی در جستجو و بازیابی اطلاعات می‌باشد. برچسب پیشنهاد شده توسط کاربر در بسیاری از موارد دقیق و اصولی نمی‌باشد و نیاز به بازبینی دارد. گاهی اوقات نیز فرد سوال کننده تسلط کافی بر روی موضوع نداشته به گونه‌ای که نمی‌تواند برچسب و یا موضوع مناسبی برای سوال مطرح شدهٔ خود پیشنهاد کند. همچنین در برخی موارد کاربران فراموش می‌کنند برای سوال مطرح شدهٔ خود برچسبی انتخاب نمایند. پیشنهاد برچسب مناسب می‌تواند هم در جهت دسته بندی موضوعی مطالب استفاده شود و هم می‌توان با استفاده از برچسب پیشنهاد شده فرد مناسب برای پاسخگویی سوال را انتخاب کرد. مثلا اگر فردی سوالی را در یکی از وبگاه‌های پرسش و پاسخ در رابطه با برنامه‌نویسی اندروید مطرح کند، سیستم به راحتی می‌تواند با استفاده از برچسب نهایی سوال مطرح شده را به فرد خبره در این حوزه بفرستد.

در این پژوهش یک روش مبتنی بر مدل‌سازی موضوعی برای جاسازی عبارت پرس‌وجو ارائه شده است که با لحاظ کردن موضوعات موجود در عبارت پرس‌وجو, عبارت پرس‌وجو را فارغ از تعداد و محل کلمات به یک نقطه در فضای جدیدی نگاشت کرده و از آن‌ها جهت پردازش‌های بعدی مانند بسط و دسته بندی عبارت پرس‌وجو استفاده می‌نماید. در واقع روش پیشنهادی با بهره‌گیری از میزان ارتباط هر عبارت پرس‌وجو با موضوعات مختلف، بردار آن عبارت را به‌دست آورده و با بهره‌گیری از مفهوم موضوع به گونه‌ای سعی دارد اطلاعات مخفی پشت کلمات عبارت پرس‌وجو را در حین جاسازی بردارهای عبارات پرس‌وجو لحاظ نماید. بدین منظور در این مقاله ابتدا کلمات مرتبط با هر موضوع را استخراج کرده و سپس با استفاده از این کلمات، میزان ارتباط عبارت پرس‌وجوی داده شده را با موضوعات مختلف به‌دست می‌آوریم. در پایان نیز با استفاده از این میزان ارتباط، بردار جاساز عبارت پرس‌وجو را به‌دست می‌آوریم و از آن در پردازش‌های بعدی استفاده می‌نماییم. روش پیشنهادی بر روی مجموعهٔ آزمون Stack Overflow ارزیابی و تحلیل شده‌است. نتایج به‌دست‌آمده نشان دهنده افزایش دقت روش ارائه‌شده در مقایسه با روش‌های موجود است.

در بخش ٢ روش پیشنهادی برای جاسازی عبارت پرس‌وجو به‌طور کامل بیان شده‌است. بخش٣ پیاده‌سازی

روش پیشنهادی را به‌همراه مجموعهٔ آزمون و پارامترهای مسئله بیان می‌کند. در بخش ٤ خروجی‌های روش پیشنهادی ارائه و تحلیل شده‌است و درنهایت مقاله در بخش ٥ با جمع‌بندی و نتیجه‌گیری پایان یافته‌است.

۲. مروری بر کارهای گذشته

امروزه با افزایش منابع متنی و کتابخانه‌ای و توسعهٔ وب، فرآیند جستجو به یکی از حوزه‌های پرکاربرد در بازیابی اطلاعات تبدیل شده است. برخی مواقع کاربران در هنگام جستجو از کلمات زائد و یا متفاوت با کلمات موجود در اسناد استفاده می‌کنند و همین امر سبب می‌شود که سامانه‌های جستجو در بازیابی نتایج مرتبط با عبارت پرس‌وجو چندان موفق نبوده و نتوانند رضایت کاربران را جلب نمایند. اخیرا از تکنیک جاسازی کلمات به‌طور گسترده در جستجو استفاده شده‌است بدین صورت که بردار کلمات عبارت پرس‌وجو ابتدا به یک فضای برداری نگاشت شده و سپس از این بردارها برای جستجوی عبارت پرس‌وجو استفاده می‌شود. بسیاری از روش‌های پیشین در این حوزه، عبارت پرس‌وجو را به‌عنوان کیسه‌ای از کلمات جاسازی شده در نظر می‌گیرند و محل قرار گرفتن یا معنی کلمات در عبارت پرس‌وجو را لحاظ نمی‌کنند. گروهی از این روش‌ها از تکنیک جاسازی عبارت پرس‌وجو استفاده می‌کنند. بدین صورت که بردار جاسازی شده کلمات موجود در عبارت پرس‌وجو را به گونه‌ای با هم ترکیب می‌کنند که یک بردار برای عبارت پرس‌وجو حاصل شود و از این بردار جهت ادامهٔ جستجو استفاده می‌کنند. در اغلب این روش‌ها هر کلمه به‌صورت یک بردار جداگانه در نظر گرفته می‌شود و برای محاسبهٔ بردار نهایی پرس‌وجو از میانگین وزن‌دار هر کلمه استفاده می‌کنند [۱۰، ۱۱، ۱۲]. برخی روش‌ها از درخت تجزیه برای ترکیب کردن بردار کلمات استفاده می‌کنند. این‌گونه روش‌ها را می‌توان تنها برای عبارات پرس‌وجویی که در قالب جمله هستند، قابل استفاده دانست[۱۳]. نالسینیک در [۱۰] به معرفی روش مدل دوگانه فضای جاساز پرداخته است. همان‌طور که می‌دانیم در روش word2vec تنها کلمات ورودی به‌صورت جاسازی شده نمایش داده می‌شوند. اما در مقاله [۱۰] کلمات عبارت پرس‌وجو به فضای ورودی و اسناد به فضای خروجی نگاشت می‌شوند و امتیاز شباهت بین همهٔ جفت کلمات عبارت پرس‌وجو-سند، با استفاده از فاصلهٔ کسینوسی محاسبه می‌شود.

لی و همکاران در [۹] نیز یک روش مبتنی بر شبکه عصبی برای تبدیل پارگراف به بردار ارائه نموده‌اند که در آن هر سند را با یک بردار متراکم نشان می‌دهد که برای پیش بینی کلمات در سند آموزش دیده است. زمانی و همکاران در [۷] روشی مبتنی بر مدل زبانی برای روش میانگین گیری بردار کلمات ارائه کرده و مدل ارائه شده را در کاربرد بسط عبارت پرس‌وجو ارزیابی کرده‌اند. آن‌ها همچنین در [١٤] روشی مبتنی بر شبکه عصبی برای جاسازی عبارت پرس‌وجو ارائه نموده اند. پلنگی و همکاران در [١٥] یک مدل RNN-LSTM معرفی کرده‌اند که تک تک کلمات را از جمله گرفته و اطلاعات آن را دریافت و در فضای معنا جاسازی می‌کند.

از کاربردهای جاسازی عبارت پرس‌وجو می‌توان به کاربرد آن در مسئلهٔ بازیابی سوال اشاره کرد. همان‌گونه که می‌دانیم، در سامانه‌های پرسش و پاسخ، سوالات بسیاری به‌صورت آرشیو قرار دارد که این سوالات منابع دانش و اطلاعات در وب هستند. همچنین مسئله تطبیق پرسش و پاسخ اهمیت بسیاری در استفاده مجدد از دانش ذخیره شده در این سامانه‌ها دارد. یکی از مهمترین مشکلات موجود در این سامانه‌ها وجود سوالات تکراری می‌باشد. بازیابی سوال[1] یا به عبارتی دیگر پیدا کردن سوالی که از لحاظ معنایی به سوال پرسیده شده نزدیک باشد را می‌توان یکی از کاربردهای جاسازی عبارت پرس‌وجو دانست. در [١٦] روش WEC[2] برای این کار پیشنهاد شده است. این روش با استفاده از بردار جاساز کلمات موجود در هر

1-question retrieval
2- Word Embedding based Correlation

عبارت پرس‌وجو احتمال همبستگی بین سوال جدید و سوالات آرشیو شده را محاسبه می‌کند. ونگ در [۱۷] برای حل این مسئله ابتدا یک مدل مبتنی بر شبکه عصبی برمبنای کیسه‌ای از کلمات از داده‌ها ساخته و سپس شباهت بین سوال جدید و کلیه سوالات و جواب‌هایی که در پایگاه داده وجود دارد را محاسبه می‌کند.

از دیگر کاربردهای جاسازی عبارت پرس‌وجو می‌توان به مسئله انتساب برچسب اشاره کرد. دلایل متعددی برای این که برخی سوالات در وبگاه‌های پرسش و پاسخ پس از مدت طولانی بدون پاسخی می‌مانند وجود دارد که یکی از آن‌ها این است که کاربر از کلمات کلیدی مناسبی به‌عنوان برچسب برای سوالات خود استفاده نکرده است که دامنه موضوع سوال را خلاصه کند [۱۸]. برچسب‌ها می‌توانند نقش مهمی در سازماندهی، نمایه سازی و طبقه‌بندی پست‌ها در وبگاه‌های پرسش و پاسخ داشته باشند. بانرجی و همکاران در [۱۹] یک رویکرد جدید برای ارزیابی انتخاب برچسب‌ها در هر پست ارائه داده‌اند که با تحلیل بر روی شبکه‌ای از برچسب‌ها یک رابطه بین برچسب‌ها به‌دست می‌آورد. مایتی و همکاران در [۲۰] یک روش مبتنی بر یادگیری عمیق برای معرفی برچسب مناسب برای پرسش‌ها در Stackoverflow پیشنهاد کرده‌اند. سیستم پیشنهاد شده نمایش محتوی را با استفاده از عنوان و بدنه سوال یاد می‌گیرد و سپس با استفاده از رابطهٔ میان برچسب‌های پیشین کاربر، برچسب نهایی را پیشنهاد می‌کند. بردار جاساز عبارت پرس‌وجو می‌تواند نقش کلیدی در بهبود روش‌های فوق الذکر ایجاد کند. تخمین دقیق بردار عبارت پرس‌وجو می‌تواند سرعت و دقت بسیاری از وظایف در بازیابی اطلاعات را افزایش دهد.

فرآیند پرس‌وجو در بسیاری موارد می‌تواند امری چالش برانگیز باشد. دراکثر موارد کلمات پرس‌وجوی به کار رفته توسط کاربران با کلمات موجود در متون تفاوت دارد. به عبارت دیگر یک عبارت پرس‌وجو می‌تواند شامل چندین کلمهٔ اصلی باشد که در سایر نوشته‌ها با هم ظاهر نشده‌اند. همچنین در برخی اوقات کاربران از کلماتی برای پرس‌وجو استفاده می‌کنند که تاثیری در نتیجه ندارد و یا فرآیند پرس‌وجو را گمراه می‌نماید. از دیگر چالش‌های موجود در این حوزه عدم وجود عبارت پرس‌وجو در زمان آموزش می‌باشد. همان‌طور که می‌دانیم در بیشتر روش‌های مطرح شده، مدل‌ها با استفاده از داده‌های عمومی مانند اخبار یا صفحات وب آموزش دیده‌اندکه نمی‌توان از آن‌ها به منظور جاساز عبارت‌هایی با موضوعات خاص به‌ویژه در وبگاه‌های پرسش و پاسخ استفاده نمود. به‌طور مثال کلمهٔ python در حالت کلی به معنای نوعی مار است اما در برنامه نویسی نام یک زبان برنامه‌نویسی است. چالش دیگری که در جاسازی عبارات پرس‌وجو وجود دارد این است که عبارت پرس‌وجو می‌تواند قالب‌های متفاوتی داشته باشد بدین صورت که می‌تواند به شکل یک یا چند کلمه، جملهٔ محاوره‌ای و یا جملهٔ استاندارد خبری یا پرسشی باشد.

بیشتر روش‌های موجود در حوزه جاسازی عبارت پرس‌وجو صرفا از بردارهای کلمات عبارت پرس‌وجو استفاده می‌کنند. برخی از روش‌ها نیز از ایده دسته‌بندی عبارات پرس‌وجو بهره برده‌اند بدین صورت که ابتدا عبارات پرس‌وجو را در دسته‌های مختلف دسته‌بندی کرده و سپس از خصوصیات هر گروه در کاربردهایی مانند جاسازی بردار، پیشنهاد برچسب، یافتن فرد خبره و مسیریابی سوالات در شبکه‌های پرسش پاسخ استفاده کرده‌اند. هیچ کدام از روش‌های موجود اهمیت موضوعات مرتبط با عبارت پرس‌وجو را در حین جاسازی بردار عبارت پرس‌وجو لحاظ نکرده‌اند. لحاظ کردن موضوعات در حین جاسازی بردارهای عبارت پرس‌وجو از این جنبه دارای اهمیت است که یک عبارت پرس‌وجو ممکن است به چندین موضوع مختلف مرتبط باشد که در این صورت استفاده از موضوعات که در واقع اطلاعات مخفی شده در پشت کلمات عبارت پرس‌وجو هستند می‌توانند به ما

در بازیابـــی بهتر اطلاعات کمک کنند. بـــرای مثال عبارت پرس‌وجو زیر را در نظر بگیرید:

"Can we share data through app directly to a google plus page?"

نگاه مبتنی بر موضوع به کلمات این جمله نشان می‌دهد که کاربر با پرســـیدن این عبـــارت پرس‌وجو موضوعاتی از قبیـــل web-api و android را هدف قرار داده اســـت در حالی که نـــگاه انفرادی به کلمات این عبارت پرس‌وجو این اطلاعات را منتقل نمی‌کند. در ادامه این مقاله تلاش می‌کنیم تـــا با اســـتفاده از تکنیک‌های مبتنی بر مبنای مدل‌ســـازی موضوعـــی[3]، موضوعات[4] نهفته در عبـــارات پرس‌وجو را اســـتخراج کنیم و به کمک آن گام موثری در راستای رفع این چالش برداریم.

۳. نوآوری مقاله

در بیشتر روش‌های موجود در حوزه جاسازی عبارت پرس‌وجو، از بردارهای جاســـازی شـــده کلمات به تنهایی برای جاســـازی بردار عبارت پرس‌وجو استفاده می‌شود. هیـــچ یک از روش‌هـــای موجود، اهمیـــت موضوع عبارت پرس‌وجو را در هنگام جاسازی عبارت پرس‌وجو در نظر نگرفته‌اند. اهمیت لحاظ کـــردن موضوع عبارت پرس‌وجو از آن جهت اســـت که یک عبارت پرس‌وجو ممکن است به چندین موضوع مختلف مرتبط باشد. در واقع موضوعات همان اطلاعات مخفی در پشـــت کلمـــات عبارت پرس‌وجو هســـتند که می‌توانند در بازیابی بهتر اطلاعات کمک کننده باشـــند. در این مقالـــه ما یک روش مبتنی بر مدل‌ســـازی موضوعـــی برای مســـئله جاســـازی بردارهـــای عبارت پرس‌وجو پیشنهاد داده‌ایم که با بهره‌گیری از میزان ارتباط هر عبارت پرس‌وجو با موضوعات مختلف، بردار آن عبارت پرس‌وجو را به‌دست می‌آورد. روش پیشنهادی سعی دارد با بهره‌گیری از مفهوم موضوع به گونه‌ای اطلاعات مخفی پشت کلمات عبارت پرس‌وجو در حین جاسازی بردارهای عبارت پرس‌وجو لحاظ نماید.

3- Topic modeling
4- Topics

۴. روش پیشنهادی

ایده اصلـــی روش پیشـــنهادی از آنجا نشـــئت گرفته اســـت که در عمل هرگاه انســـان به دنبـــال درک موضوع یک ســـند باشـــد ابتدا یک نگاه کلی به آن سند می‌اندازد و سپس کلمات کلیدی را استخراج کرده و با استفاده از آن‌ها موضوع ســـند را تشـــخیص می‌دهد. ما نیز در اینجا برای تعیین موضوع اصلی هر عبارت پرس‌وجو از همین روش اســـتفاده کرده‌ایم. بدین صورت کـــه ابتدا کلمات مرتبط با هر موضوع را اســـتخراج کرده ســـپس با اســـتفاده از این کلمـــات، میزان ارتبـــاط عبارت پرس‌وجوی داده شـــده را با موضوعات مختلف به‌دســـت می‌آوریم. در پایان نیز با اســـتفاده از این میزان ارتباط، بردار هر عبارت پرس‌وجو را به‌دســـت می‌آوریم. در ادامه گام‌های روش پیشـــنهادی به‌صورت دقیق‌تر بیان می‌شود.

٤-١ گام ۱: اســـتخراج کلمـــات مهم با اســـتفاده از مدل‌سازی موضوعی

در گام نخســـت کلمات کلیدی مرتبط بـــا هر موضوع را اســـتخراج می‌نماییـــم تا در گام‌های بعـــدی از اطلاعات آن‌هـــا به منظور جاســـازی عبـــارت پرس‌وجو اســـتفاده نماییـــم. در این گام از ابـــزار MALLET بـــه منظور انجام مدل‌سازی موضوعی اســـتفاده می‌کنیم [۲۳]. این ابزار با اســـتفاده از روش تخصیص پنهان دریکله، ســـندها را در تعداد موضوعات مورد نظر دسته‌بندی می‌نماید. در ابزار MALLET فایـــل ورودی باید در قالب خاصی که توســـط تولیدکنندگان آن تعریف شده‌است تهیه شود. قالب استفاده شـــده در این مدل بدین صورت است که هر سطر از فایل شـــامل یک عبارت پرس‌وجو و یک شناسه منحصر به فرد می‌باشـــد. این فایل بایـــد در اختیار ابـــزار MALLET قرار داده شـــود تا ســـندها در تعداد موضوعات مورد نظر (در این پژوهش ســـه موضوع) دسته‌بندی شوند. پس از اتمام مدل‌سازی موضوعی، به ازای هر موضوع، ۲۰ کلمهٔ کلیدی آن موضـــوع را به‌عنوان کلمات کلیدی آن موضوع در نظر می‌گیریم.

٤-٢ گام ۲: ایجاد بردار متناظر با هر موضوع

در این گام ابتدا بردار جاسازی شده کلمات کلیدی مرتبط با هر موضوع را که در مرحلهٔ قبلی به‌دست آمده است توسط روش شناخته شده Word2Vec محاسبه کرده و بر اساس آن‌ها بردار متناظر با هر موضوع را به کمک رابطهٔ زیر به‌دست می‌آوریم.

$$\overrightarrow{V_{T_i}} = \sum_{w_j \in W} \overrightarrow{V_{w_j}}\, p(w_j|T_i) \qquad (۱)$$

که $p(w_j|T_i)$ نشان دهندهٔ میزان ارتباط کلمهٔ کلیدی w_j به موضوع T_i می‌باشد و $\overrightarrow{V_{w_j}}$ بردار Word2Vec متناظر با کلمهٔ کلیدی w_j می‌باشد. در پایان این مرحله ما برای هر موضوع یک بردار $\overrightarrow{V_{w_j}}$ داریم که ابعاد آن متناظر با ابعاد کلمات در فضای جاسازی Word2Vec می باشد.

٤-٣ گام ۳: استخراج بردار برای هر عبارت پرس‌وجو

پس از آن که به ازای هر موضوع یک بردار بر اساس کلمات کلیدی آن تخمین زده شد، در این گام برای هر عبارت پرس‌وجو یک بردار بر اساس احتمال تعلق آن عبارت به موضوعات مختلف به‌دست می‌آوریم. برای تخمین احتمال تعلق یک عبارت پرس‌وجو به یک موضوع از الگوریتم مدل‌سازی موضوعی LDA استفاده می‌نماییم [۲٤]. لذا بردار متناظر با هر عبارت جست‌وجو را می‌توان از رابطهٔ زیر به‌دست آورد.

$$\overrightarrow{V_q} = \sum_{i=1}^{N} \overrightarrow{V_{T_i}}\, p(T_i|query) \qquad (۲)$$

که در آن $p(T_i|query)$ بیانگر توزیع احتمالاتی عبارت پرس‌وجو $query$ نسبت به موضوع T_i و $\overrightarrow{V_{T_i}}$ بیانگر بردار متناظر با هر موضوع می‌باشد. در پایان این گام به ازای هر عبارت پرس‌وجو، یک بردار در فضای جاسازی به‌دست می‌آید که از آن می‌توان در کاربردهای مختلف بازیابی اطلاعات استفاده نمود.

۵. پیاده‌سازی روش پیشنهادی

در این بخش ابتدا به نحوهٔ تولید مجموعهٔ آزمون و پارامترهای آن پرداخته می‌شود. سپس پارامترهای دخیل در انجام آزمایش‌ها و معیار ارزیابی معرفی و نحوهٔ مقداردهی آن‌ها بیان می‌شود. در نهایت نتایج حاصل از پیاده‌سازی مدل‌های پیشنهادی با روش‌های معروف موجود در این حوزه مقایسه خواهد شد.

جدول۱: اطلاعات داده‌ی آزمون

Topic	#query
Android	73381
Php	51564
JavaScript	47929

۵-۱ مجموعهٔ آزمون

وبگاه‌های مبتنی بر پرسش و پاسخ امروزه نقش بسزایی در رفع مشکلات و سوالات تخصصی کاربران دارند. یکی از معروف‌ترین این وبگاه‌ها Stack Overflow می‌باشد. مدل‌های پیشنهادشده در این مقاله بر روی مجموعه آزمون Stack Overflow که شامل ۲٤،۱۲۰،۵۲۳ پست در بازهٔ زمانی آگوست ۲۰۰۸ تا مارچ ۲۰۱۵ است، اجرا و ارزیابی شده‌است. در شبکهٔ Stack Overflow هر سوال یک یا چند برچسب مخصوص به خود دارد که طراح سوال با توجه به ذات پرسش به آن نسبت می‌دهد. برای ارزیابی دقیق‌تر در اینجا فقط عباراتی در نظر گرفته شده‌است که فقط یکی از سه برچسب JavaScript، PHP و Android را دارند و عنوان هر پست به‌عنوان یک عبارت پرس‌وجو در نظر گرفته شده‌است. اطلاعات دادهٔ آزمون ایجاد شده در جدول ۱ با جزییات بیشتر قابل مشاهده می‌باشد.

۵-۲ معیار ارزیابی

برای ارزیابی مدل‌های پیشنهادی و مدل‌های مبنا از معیارارزیابی MAP[5] استفاده شده‌است. معیار MAP به‌عنوان یک معیار ترکیبی که همزمان دو معیار دقت و فراخوانی را در نظر می‌گیرد یکی از روش‌های مناسب جهت ارزیابی می‌باشد. نحوه محاسبه MAP به‌صورت زیر می‌باشد.

5- Mean Average Precision

$$MAP = \frac{1}{|Q|}\sum_{q=1}^{|Q|}\left(\frac{1}{|\mathrm{Eq}|}\sum_{i=1}^{|R_q^{\cdot}|}1_{R_q^i \in E_q}\left(P@i\right)\right) \qquad (۳)$$

کــه در آن |Q| تعداد کل عبــارات پرس‌وجوی موجود در مجموعه آزمون، E_q مجموعه برچســب‌های بازیابی شده برای یک عبارت پرس‌وجو و Rq برچسب‌های صحیح برای عبــارت پرس‌وجوی مورد نظر اســت. مقدار $1_{condition} = 1$ اســت اگر condition = true باشد یعنی برچسب عبارت پرس‌وجو در برچسب‌های بازیابی شده وجود داشته باشد و P@i نشان‌دهندۀ دقت سیستم در خروجی شمارۀ i است.

۶. شبیه‌سازی و تحلیل نتایج

به منظور ارزیابی کارایی روش پیشنهادی در جاسازی عبارت پرس‌وجو، ابتدا یک تحلیل مبتنی بر رابطه همسایگی برای بردارهای جاســازی تمام عبارات پرس‌وجو خواهیم داشت. بدین صورت که برای تمام عبارات پرس‌وجو بردار جاساز آن‌ها را به کمک روش‌های مختلف به‌دست می‌آوریم و از آنجــا که انتظــار داریم عبارات پرس‌وجوی مشــابه، بردارهای مشــابه به هم داشته باشند بررسی می‌کنیم که کدام روش این شرط را بیشتر ارضا نموده‌است. لذا برای هر عبارت پرس‌وجوی موجود در مجموعه داده یک بردار به کمک روش پیشــنهادی به‌دســت آورده و ۵ برداری را کــه کمترین فاصله با بردار عبــارت پرس‌وجو حاصل را داشــته‌اند، پیدا می‌کنیم. خروجی حاصــل برای تعدادی از عبارات پرس‌وجو در جدول ۳ قابل مشاهده می‌باشد. همین کار را با اســتفاده از روش میانگین‌گیری از بردار کلمات[6] انجــام داده و ۵ بردار نزدیک به هــر عبارت پرس‌وجو را برای این حالت نیز به‌دست می‌آوریم. در پیاده سازی روش میانگین‌گیری از بردار کلمات، کلیۀ کلمات زائد و ایست‌واژه از عبارت پرس‌وجو حذف شده است و فقط از کلمات اصلی موجود در هر عبارت پرس‌وجو استفاده شده است. همچنین برای هر عبارت پرس‌وجو، ۵ عبارت پرس‌وجو با کمترین فاصلۀ انتقالی کلمه نیز اســتخراج و نتیجه آن در جدول ۳ لیست شده اســت. معیار فاصلۀ انتقالی کلمه روشی موثر برای محاســبه فاصله دو متن است که بر اساس کمترین فاصلۀ انتقال بردار جاســاز یک متــن به متن دیگر فاصله متــون را محاســبه می‌نماید [۲۱]. نگاهــی دقیق به جدول ۳ نشــان می‌دهد که در روش پیشــنهادی بر خلاف روش میانگیــن وزن‌دار عبارات پرس‌وجویی کــه معنای تقریبا یکســانی دارند به فضای یکســانی نیز نگاشت شده‌اند که این امر بیانگر یکی از نقاط قوت روش پیشنهادی می‌باشد.

همچنین برای ارزیابی بهتر تشابه عبارات پرس‌وجویی که برچســب یکســانی دارند ســعی کرده‌ایم تا از روش خوشه‌بندی اســتفاده نماییم. دلیل این امر هم این است که روش‌های خوشــه‌بندی به‌صورت بدون ناظر و بر اساس میزان شــباهت بردارهــا، آن‌هــا را در گروه‌های مختلف خوشــه‌بندی می‌کنند. لذا خوشــه‌بندی دقیق‌تر بردارهای عبارات پرس‌وجو به گونه‌ای بیانگر جاســازی بهتر آن‌ها و در نتیجه دقت بالاتر روش می‌باشد. بدین منظور کیفیت خوشــه‌بندی روش K-Means به کمــک دو معیار NMI و ARI برای دو روش پیشنهادی و میانگین‌گیری وزن‌دار در جدول شماره ۲ گزارش شده است[۲۵]. از آنجا که هدف از خوشه‌بندی، ایجاد دسته‌هایی با بیشترین شباهت درون-خوشه‌ای و کمترین شباهت بین-خوشه‌ای است، لذا هر یک از معیارهای ارزیابی اگر به یک نزدیک‌تر باشند بدین معنی اســت که بردارهایی که برچسب یکسانی دارند در فاصلۀ نزدیک‌تری به یکدیگر قرار گرفته اند. نتایج حاصل نشــان می‌دهد که عبارات پرس‌وجویی که دارای برچسب یکسانی هستند تقریبا به یک ناحیه از فضا نگاشت شده‌اند.

علاوه بر انجــام دو آزمایش قبلی برای ارزیابی روش پیشنهادی، کارایی روش پیشــنهادی را در کاربرد واقعی پیشنهاد برچســب نیز بررسی نموده ایم. امروزه پیشنهاد برچسب یا دســته‌بندی مناسب هر ســوال نقش بسزایی در کاربری وبگاه‌های پرســش و پاســخ دارد. به‌طور کلی برچسب‌دهی مناســب می‌تواند هم در پیدا کردن سوالات مشــابه و هــم در پیــدا کردن فــرد خبره مناســب جهت

6- Average Word Embedding

جدول ۲: ارزیابی کیفیت خوشه‌بندی بر روی بردارهای جاساز عبارات پرس‌وجو به کمک دو روش پیشنهادی و میانگین‌گیری وزن‌دار

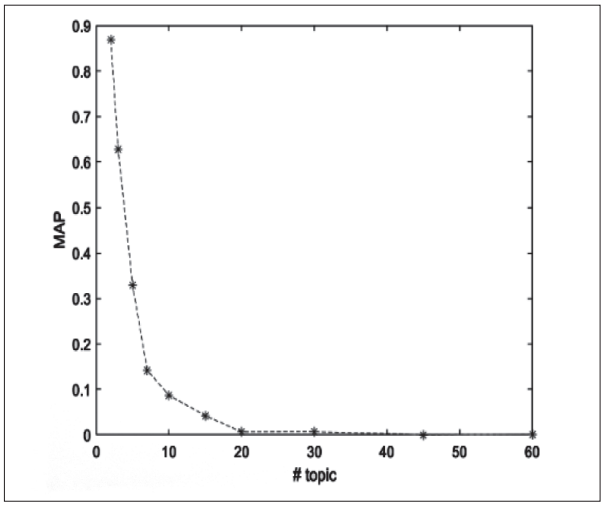
معیار ارزیابی	روش میانگین وزن‌دار کلمات	روش پیشنهادی
NMI	۰.۲۰۴۳	۰.۲۳۱۷
ARI	۰.۱۲۰۷	۰.۲۵۳۲

پاسخگویی به سوال مطرح شده، بسیار موثر باشد.

عموما شبکه‌های پرسش و پاسخ اینترنتی بر اساس نوع برچسب‌دهی به دو دسته کلی تقسیم می‌شوند. دسته نخست وبگاه‌هایی مانند YAHOO ANSWERS، وبگاه‌هایی هستند که دسته‌بندی سوالات از قبل مشخص شده و کاربر پس از مطرح کردن سوال خود یکی از برچسب‌های موجود در دسته‌بندی‌های موجود را به‌عنوان برچسب متناظر با سوال مطرح شده انتخاب می‌کند. در دستهٔ دوم وبگاه‌هایی مانند Stack Overflow هستند که برچسب متناظر با هر سوال توسط کاربر تعریف می‌شود. در این‌گونه وبگاه‌ها کاربر می‌تواند تعدادی کلمه را به‌عنوان برچسب به سوال مطرح شده نسبت دهد [۲۶].

در این مقاله برای ارزیابی کارایی روش پیشنهادی در کاربرد پیشنهاد برچسب، از سوالاتی استفاده کرده‌ایم که تنها دارای یکی از سه برچسب Android، JavaScript و یا PHP هستند. برای ارزیابی کارایی روش پیشنهادی در کاربرد پیشنهاد برچسب بدین صورت عمل می‌کنیم که ابتدا بردار جاسازی شده متناظر با هر عبارت پرس‌وجو را پیدا کرده و سپس K کلمهٔ نزدیک به این بردار را در مدل از پیش آموزش‌دیده Word2Vec به‌دست می‌آوریم. بدین منظور از کتابخانهٔ genism و بردارهای از پیش آموزش‌دیدهٔ google news[۷] استفاده کردیم. اگر برچسب متناظر با آن عبارت پرس‌وجو در بین k کلمه پیدا شده بود، مقدار p@k آن عبارت پرس‌وجو را یک و در غیر این‌صورت این مقدار را صفر در نظر می‌گیریم. این روش ارزیابی می‌تواند به گونه‌ای هر دو مدل روش برچسب‌دهی در وبگاه‌های پرسش‌وپاسخ را پشتیبانی نماید. همچنین دقت روش پیشنهادی در این کاربرد بر اساس معیار MAP

7- https://code.google.com/archive/p/word2vec/



شکل ۱: نمودار تغییرات MAP بر حسب تعداد موضوع در الگوریتم LDA

نیز گزارش شده است. برای محاسبه MAP، میانگین p@k را به ازای مقادیر مختلف k=1,2,3,...,100 محاسبه و نتیجه را با روش میانگین‌گیری وزن‌دار کلمات و روش ارائه شده در مقاله [۲۲] مقایسه نموده‌ایم. نتایج حاصل در جدول شماره ٤ گزارش شده است. همان‌طور که در جدول ٤ مشاهده می‌شود روش پیشنهادی بهبود قابل توجهی در این کاربرد خاص نسبت به دو روش مقایسه شده دارد.

۶-۱ تحلیل پارامترهای تاثیرگذار

در این بخش به بیان پارامترهای اصلی موجود در روش پیشنهادی می‌پردازیم. یکی از پارامترهای مهم در روش پیشنهادی تعیین تعداد موضوعات در الگوریتم LDA می‌باشد که این پارامتر تا حد زیادی وابسته به مجموعه داده می‌باشد. تغییرات این پارامتر در شکل ۱ قابل مشاهده است.

همان‌طور که در شکل ۱ مشاهده می‌شود با افزایش تعداد موضوعات میزان MAP کاهش می‌یابد. یکی از دلایل اصلی در این مسئله کوتاه بودن طول جملات عبارت پرس‌وجو می‌باشد و این که هر جمله حداکثر می‌تواند به تعداد کمی از موضوعات وابسته باشد.

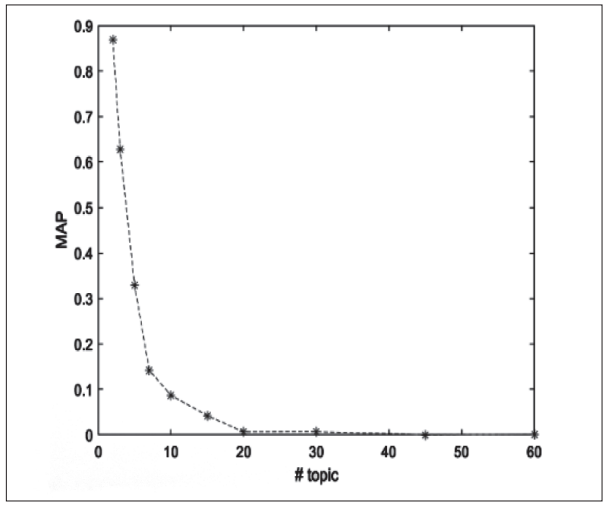
پارامتر دیگری که در کارایی روش پیشنهادی موثر می‌باشد تعداد کلمات کلیدی مورد استفاده در ساخت بردار متناظر با هر موضوع می‌باشد. تغییرات این پارامتر در شکل

جدول ۳ : پنج عبارت پرس‌وجو نزدیک به عبارت پرس‌وجو داده شده در فضای جاساز

Query	Method	Top 5 nearest query (self-query was omitted)
problem in creating temporary file in android	Proposed	1- how to find the path of database file in android emulator 2- how do i access the data files belongs to a application in android 3- how to create a file in android 4- how to config picture in android programming 5- how to extract file in android
	AWE [۹]	1-how to show long string in a text view without horizontal scroll 2- how to add image in android 3- maximize an alert dialog 4- experiences with android 5- how can i use jsr 172 in android
	WMD [۲۱]	1-how to create a file in android 2-how to open .png or .doc file in android 3-how to extract file in android 4-how to find the path of database file in android emulator 5-encrypting a db file in android
catching file download	Proposed	1-how to download file 2-how to store object in sqlite database 3-how to store data in android database 4-problem sending an email with an attachment programmatically 5-how to extract file in android
	AWE [۹]	1-catching file download 2-android as a commercial game platform 3-how to get android application id 4-android: is it possible to display video thumbnails 5-how to include and use zxing library in android with eclipse
	WMD [۲۱]	1-catching file download 2-how to create a file in android 3-how to extract file in android 4-how to encrypt database file 5-problem in creating temporary file in android
how to store object in sqlite database	Proposed	1- catching file download 2- how to encrypt database file 3- how to store data in android database 4- problem sending an email with an attachment programmatically 5- problem in creating temporary file in android
	AWE **[۹]**	1-how to extract file in android 2-animating view. scroll to(...) 3-android with e ink display 4-programmatically click views in android 5-how can i create 2 rows of buttons in android layout.xml file
	WMD **[۲۱]**	1-how to store data in android database 2-how to encrypt database file 3-how to find the path of database file in android emulator 4-how to query mms sender and text from mms database 5-how can i call a web service without using ksoap2 in android

جدول ۴: مقایسه کارایی روش پیشنهادی بر روی داده‌های stack overflow در کاربرد پیشنهاد برچسب

Tag	Method	p@5	p@5	p@10	MAP
Android	Proposed	0.52475	0.54212	0.54605	0.6280989
	AWE [۹]	0.04967	0.18018	0.20832	0.2500227
	MLE [۲۱]	0.1035	0.2186	0.238	0.271929
Php	Proposed	0.43251	0.49860	0.5179	0.6660854
	AWE [۹]	0.07784	0.23884	0.27734	0.33817
	MLE [۲۱]	0.26835	0.26835	0.35155	0.3871734
JavaScript	Proposed	0.477222	0.483847	0.53431	0.6756997
	AWE [۹]	0.07569	0.258909	0.2907	0.34014
	MLE [۲۱]	0.0902	0.283315	0.3438	0.374228



شکل ۲: نمودار تغییرات MAP برحسب پارامتر تعداد کلمات کلیدی در الگوریتم پیشنهادی

۲ نشـــان داده شده اســـت. همان‌طور که مشاهده می‌شود برای مجموعهٔ آزمون ایجاد شـــده هـــر چقدر تعداد کلمات کلیدی استفاده شده بیشتر باشد میانگین دقت متوسط در وظیفهٔ پیشنهاد برچسب کاهش می‌یابد. همان‌گونه که شکل ۲ نشان می‌دهد، با افزایش تعداد کلمات کلیدی، میزان MAP کاهش می‌یابد که می‌تواند به دلیل افزایش کلمات نامرتبط با موضوع در خروجی الگوریتم LDA باشد.

۷. نتیجه‌گیری

امروزه با افزایش منابع متنی و شـــبکه‌های پرســـش و پاسخ، جاسازی عبارات پرس‌وجو می‌تواند نقش مهمی در کاربردهای مختلف بازیابی اطلاعات ایفا نماید. در این مقاله یک روش مبتنی بر مدل‌ســـازی موضوعی برای جاسازی عبارت پرس‌وجو ارائه شـــده است که در مقایسه با سایر مدل‌های جاســـازی عبارت پرس‌وجو نتایج بهتری را ارائه کرده‌است. روش پیشنهادی در این مقاله بر روی مجموعه آزمون ایجاد شده از مجموعه پست‌های Stack Overflow آزمایـــش و ارزیابی شده‌اســـت. نتایج حاصـــل، بهبود در کاربردهایـــی مانند پیشـــنهاد برچســـب را در مقایســـه با روش‌های پیشین نشان می‌دهد.

مراجع

[1] Jastrzebski S, Leśniak D, Czarnecki WM. How to evaluate word embeddings? on importance of data efficiency and simple supervised tasks. arXiv preprint arXiv:1702.02170. 2017.

[2] Bakarov A. A survey of word embeddings evaluation methods. arXiv preprint arXiv:1801.09536. 2018.

[3] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013.

[4] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. InProceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp.1532-1543. 2014.

[5] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T., Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, pp.135-146. 2017.

[6] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv: 1810.04805. 2018.

[7] Zamani, H, and Croft, B. "Estimating embedding vectors for queries." In Proceedings of the 2016 ACM International

Conference on the Theory of Information Retrieval, pp. 123-132. ACM. 2016.
[8] Kiros, R., Zemel, R., and Salakhutdinov,R., "A multiplicative model for learning distributed text-based attribute representations." In Advances in neural information processing systems, pp. 2348-2356. 2014.
[9] Le, Quoc, and Mikolov, T., "Distributed representations of sentences and documents." In International conference on machine learning, pp. 1188-1196. 2014.
[10] E. Nalisnick, B. Mitra, N. Craswell, and R. Caruana. Improving Document Ranking with Dual Word Embeddings. In WWW '16, pp. 83–84. 2016.
[11] I. Vuli´c and M.-F. Moens. Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings. In SIGIR '15, pp. 363–372. 2015.
[12] G. Zheng and Callan, J. Learning to Reweight Terms with Distributed Representations. In SIGIR '15, pp. 575–584. 2015.
[13] Socher, R., Lin, C. C., Manning, C., & Ng, A. Y. Parsing natural scenes and natural language with recursive neural networks. In Proceedings of the 28th international conference on machine learning (ICML-11), pp. 129-136. 2011.
[14] Zamani, H, and Croft, B. Relevance-based word embedding. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 505-514. 2017.
[15] Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., and Ward, R. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 24(4),pp. 694-707, 2016.
[16] Shen, Y., Rong, W., Jiang, N., Peng, B., Tang, J., and Xiong, Z. Word embedding based correlation model for question/answer matching. In Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[17] Weng, S., Wu, C. K., Wang, Y. C., and Tsai, R. T. H. Question Retrieval with Distributed Representations and Participant Reputation in Community Question Answering. In International Journal of Computational Linguistics & Chinese Language Processing, Vol 22, No 2, Special Issue on Selected Papers from ROCLING XXIX. 2017.
[18] Roy, P. K., & Singh, J. P. A Tag2Vec Approach for Questions Tag Suggestion on Community Question Answering Sites. In International Conference on Machine Learning and Data Mining in Pattern Recognition (). Springer, Cham. pp. 168-182. 2018.
[19] Banerjee, R., Rajanala, S., & Singh, M. Evaluating the Choice of Tags in CQA Sites. In International Conference on Database Systems for Advanced Applications. Springer, Cham. pp. 625-640. 2019.
[20] Maity, S. K., Panigrahi, A., Ghosh, S., Banerjee, A., Goyal, P., & Mukherjee, A. DeepTagRec: A Content-cum-User Based Tag Recommendation Framework for Stack Overflow. In European Conference on Information Retrieval. Springer, Cham, pp. 125-131. 2019.
[21] Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. From word embeddings to document distances. In International conference on machine learning, pp. 957-966. 2015.
[22] Zamani, H., and Croft, W. B. Embedding-based query language models. In Proceedings of the 2016 ACM international conference on the theory of information retrieval. ACM. pp. 147-156. 2016.
[23] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit."http://mallet.cs.umass.edu. 2002.
[24] Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research, Vol. 3, pp. 993-1022. 2003.
[25] Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms .Society for Industrial and Applied Mathematics. pp. 1027-1035. 2007.
[26] Gupta, M., Li, R., Yin, Z. and Han, J. Survey on social tagging techniques. ACM Sigkdd Explorations Newsletter,Vol. 12,No. 1, pp.58-72. 2010.