علوم و مهندسی کامپیوتر
نشریه علمی پژوهشی انجمن کامپیوتر ایران
مجلد ۱۲، شماره ۱ (الف)، ۱۳۹۳
صفحات ۳۱-۳۷
مقاله عادی

ارائه‌ی یک روش بهبود یافته مبتنی بر آنالیز معنایی برای دستیابی به اطلاعات در شبکه‌های اجتماعی

محمد ستاری کامران زمانی‌فر ناصر نعمت‌بخش

دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه اصفهان، اصفهان، ایران

چکیده

امروزه ابزارهای بسیاری برای دسترسی سریع و آسان به اطلاعات در شبکه‌های اجتماعی معرفی شده‌اند که مهمترین و پرکاربردترین آنها، ابر برچسب است. یکی از مهمترین مسائل در مورد ابر برچسب[1]، نحوه‌ی انتخاب برچسب[2] برای آن است. تاکنون سه روش برای مشخص کردن چگونگی این انتخاب ارائه شده است. اولی مبتنی بر پرکاربرد بودن، دومی ترکیبی از آنالیز نحوی و خوشه‌بندی معنایی (روش مبتنی بر آنالیز نحوی) و سومی بهبود یافته‌ی روش دوم است. با این وجود، این روش‌ها هم با مشکلاتی مواجه هستند. برای حل این مشکلات، در این مقاله، یک روش ارائه شده است که در آن از آنالیز معنایی به جای آنالیز نحوی استفاده شده است. روش ارائه شده و روش‌های قبلی را روی بخشی از دو پایگاه داده‌ی دلیشس و بیبسونومی پیاده‌سازی شده‌اند. نتایج حاکی از بهبود معیار پوشش و عدم‌تغییر محسوس معیار اشتراک در روش ارائه شده نسبت به دو روش قبلی بوده است.

کلمات کلیدی: شبکه‌های اجتماعی، ابر برچسب، آنالیز معنایی، دستیابی به اطلاعات، خوشه‌بندی، بازیابی اطلاعات.

۱- مقدمه

امروزه ما با خدمات زیادی از وب ۲ سر و کار داریم مانند بلاگ‌ها، فروم‌ها، گروه‌ها، شبکه‌های اجتماعی و غیره. اما محبوب‌ترین و پرکاربردترین سرویس در بین خدمات وب ۲، شبکه‌های اجتماعی است که میلیون‌ها یا حتی میلیاردها کاربر را از سراسر جهان به سوی خود جلب کرده است. به طور مثال در حال حاضر شبکه‌ی اجتماعی فیس بوک بیش از یک میلیارد کاربر فعال دارد.

یک شبکه‌ی اجتماعی، ساختاری اجتماعی شامل افرادی است که توسط یک یا چند وابستگی به هم متصل شده‌اند [۱۰]. یکی از بخش‌های اصلی شبکه‌های اجتماعی، برچسب‌های اجتماعی هستند. برچسب‌ها، کلمات کلیدی هستند که از نظر شکل به کارگیری آزاد (بی‌قاعده) بوده و برای منابع اطلاعاتی مانند متن، عکس و ویدیو به کار می‌روند [۳]. یکی از کاربردهای مهم برچسب‌های اجتماعی، دستیابی به اطلاعات می‌باشد. ابزارهای متعددی برای دستیابی به اطلاعات با استفاده از برچسب مطرح شده‌اند، اما هیچ یک به اندازه‌ی ابر برچسب پرکاربرد و مشهور نشدند. دلیل مشهور شدن و پرکاربرد بودن ابر برچسب، استفاده در شبکه‌های اجتماعی مشهور مانند فلیکر، تکنوراتی و همچنین بکارگیری خصوصیات بصری متفاوت جهت ارائه‌ی محتوای اطلاعاتی است. در ادامه با این ابزار بیشتر آشنا می‌شویم.

ابر برچسب، ابزاری است که از آن برای بازیابی اطلاعات استفاده می‌شود. این ابزار، یک نمایش بصری از برچسب‌های تولید شده‌ی کاربری ارائه می‌دهد که موجب تسهیل و تسریع دسترسی به اطلاعات در شبکه‌های اجتماعی می‌شود. در ادامه عملکرد این ابزار شرح داده خواهد شد.

ابر برچسب تعداد محدودی از برچسب‌های به کار برده شده در یک شبکه‌ی اجتماعی را برای نمایش انتخاب می‌کند که هر کدام از این برچسب‌ها با چندین منبع در ارتباط هستند. کاربر می‌تواند با کلیک بر روی هر یک از این برچسب‌ها به منابع مرتبط با آنها شامل صفحات وب، عکس و غیره دست یابد.

در ابر برچسب از خصوصیات بصری متفاوتی مانند اندازه، رنگ و غیره می‌توان استفاده کرد که این تنوع استفاده از خصوصیات بصری یکی از دلایل مشهوریت ابر برچسب است. به عنوان مثال اندازه‌ی فونت هر برچسب می‌تواند نشان‌دهنده‌ی میزان کاربرد آن باشد.

از ابر برچسب علاوه بر شبکه‌های اجتماعی می‌توان برای خلاصه کردن نتایج جستجو در موتورهای جستجو [۶] و پرس‌وجوهای پایگاه داده [۴، ۵] استفاده کرد.

تحقیقات در زمینه‌ی ابر برچسب در شبکه‌های اجتماعی را می‌توان به دو تقسیم کرد:

۱- جنبه‌های بصری

۲- انتخاب برچسب برای ابر برچسب

بخش اول تحقیقات در زمینه‌ی ابر برچسب، جنبه‌های بصری است. این تحقیقات بیشتر بر روی نحوه‌ی آرایش برچسب‌ها در ابر برچسب متمرکز است [۲]. در این تحقیقات، سه نوع آرایش مورد مطالعه قرار گرفته شده است: ۱- الفبایی، ۲- تصادفی و ۳- معنایی. نتایج این بررسی‌ها حاکی از بهتر بودن آرایش معنایی نسبت به دو آرایش دیگر است.

بخش دوم تحقیقات در زمینه‌ی انتخاب برچسب برای ابر برچسب است که تمرکز این مقاله هم بر روی این بخش است. در ادامه کارهای انجام شده در این زمینه بررسی خواهند شد.

۱-۱- کارهای مرتبط

ابر برچسب با هدف نمایش برچسب‌های پرکاربرد در شبکه‌ی اجتماعی اولین بار در سال ۲۰۰۴ توسط استوارت باترفیلد (یکی از اعضای سازنده‌ی شبکه‌ی اجتماعی اشتراک عکس فلیکر) در شبکه‌ی اجتماعی فلیکر به کار برده شد و سپس با استفاده در شبکه‌های اجتماعی دیگر مانند تکنوراتی محبوب شد. پس از اینکه شبکه‌های اجتماعی مختلفی این ابزار را به کار بردند، بحث استفاده از آن به عنوان ابزاری برای بازیابی و دستیابی به اطلاعات مطرح شد.

در سال ۲۰۰۶، میلن [۹] با بررسی‌هایی که روی سیستم‌های بوک مارک اجتماعی انجام داد بر در نظر گرفتن معیار پرکاربرد بودن در ابر برچسب صحه گذاشت و به این نتیجه رسید که ۱۰ برچسبی در ابر برچسب بیشتر کلیک می‌شوند، جز ۱۰ برچسب پرتکرار در آن شبکه‌ی اجتماعی هستند.

در سال ۲۰۰۷، ریواندریا [۱۱] ابر برچسب را به عنوان ابزاری معرفی کرد که به مردم در کشف تصادفی اطلاعات کمک می‌کند. در همین سال، طبق بررسی سینکلیر [۱۲] مشخص شد که از ابر برچسب با توجه به دانه‌بندی درشت آن فقط برای مرور اطلاعات کلی می‌توان استفاده کرد و برای مرور اطلاعات خاص کاربردی ندارد.

سپس در سال ۲۰۱۱، پتروس ونتی [۱۴] با استفاده از معیارهای پوشش و اشتراک، مرتبط بودن و غیره یک چهارچوب جهت ارزیابی الگوریتم‌های انتخاب برچسب برای خلاصه کردن نتایج موتورهای جستجو معرفی کرد که از برخی از این معیارها مانند پوشش و اشتراک می‌توان برای ارزیابی الگوریتم‌های انتخاب برچسب در شبکه‌های اجتماعی استفاده کرد.

در سال ۲۰۱۲، مارتین [۸] معتقد بود که ابر برچسب مبتنی بر معیار پرکاربرد بودن (انتخاب برچسب‌ها با بیشترین تکرار) مشکل اشتراک بالا را دارد. براساس دیدگاه او، برچسب‌های پرکاربرد مجموعه‌ی بزرگی از اسناد را پوشش می‌دهند که این باعث افزایش اشتراک در برچسب‌های انتخابی می‌شود.

همچنین او مشکل وجود برچسب‌های هم‌خانواده را برای ابر برچسب مبتنی بر معیار پرکاربرد بودن ذکر کرد و جهت حل این مشکل، یک روش ترکیبی از آنالیز نحوی و خوشه‌بندی معنایی ارائه داد. بررسی‌های او نشان داد که این روش نسبت به روش مبتنی بر پرکاربرد بودن از لحاظ دو معیار پوشش و اشتراک بهتر عمل می‌کند.

در سال ۲۰۱۳، ستاری [۱] مشکل خوشه‌بندی ناصحیح کلمات هم‌خانواده را در آنالیز نحوی روش مارتین [۸] ذکر کرد و یک روش بهبود یافته برای حل این مشکل ارائه داد. نتایج، حاکی از بهبود دو معیار پوشش و اشتراک در این روش نسبت به روش مارتین بوده است.

۲- معیارهای ابر برچسب در شبکه‌ی اجتماعی

معیارهای ابر برچسب در شبکه‌های اجتماعی را می‌توان به دو دسته تقسیم کرد: ۱- معیارهایی که برای ارزیابی ابر برچسب به کار می‌روند. ۲- معیارهایی که جهت انتخاب برچسب برای ابر برچسب به کار می‌روند.

۲-۱- معیارهای ارزیابی ابر برچسب

معیارهای گوناگونی برای ارزیابی ابر برچسب در شبکه‌های اجتماعی در نظر گرفته شده‌اند. مهمترین و شناخته‌شده‌ترین این معیارها عبارتند از: ۱- پوشش: مشخص می‌کند که ابر برچسب چه نسبتی از کل منابع موجود در یک شبکه‌ی اجتماعی را در برگرفته است. ۲- اشتراک: میزان منابع مشترک در بین هر جفت برچسب در ابر برچسب را مشخص می‌کند.

در این مقاله مانند مقالات قبلی [۱، ۸]، این دو معیار به عنوان معیارهای ارزیابی ابر برچسب در نظر گرفته می‌شوند.

۲-۲- معیارهای انتخاب برچسب برای ابرچسب

معیارهایی که تاکنون جهت انتخاب برچسب برای ابر برچسب ارائه شده‌اند، عبارتند از: ۱- پرکاربرد بودن: برچسب‌هایی انتخاب شوند که بیشترین منابع را پوشش می‌دهند. ۲- پوشش زنجیره‌ای: برچسب‌هایی انتخاب شوند که با منابع بیشتری (بدون در نظر گرفتن منابع پوشش داده شده‌ی قبلی) مرتبط باشند. مزیت این معیار این است که دو معیار مستقل ارزیابی ابر برچسب (پوشش و اشتراک) را بهینه می‌کند، در حالی که معیار پرکاربرد بودن فقط پوشش را بهینه کرده و اشتراک را در نظر نمی‌گیرد.

۳- روش‌های متداول در انتخاب برچسب برای ابر برچسب

یکی از مسائل مهمی که در ابر برچسب مطرح است این است که چگونه و بر اساس چه معیارهایی، برچسب‌ها را جهت نمایش در ابر برچسب انتخاب کنیم. تاکنون سه روش جهت انتخاب برچسب برای ابر چسب ارائه شده است: ۱- روش مبتنی بر پرکاربرد بودن ۲- روش مبتنی بر آنالیز نحوی ۳- روش بهبود یافته مبتنی بر آنالیز نحوی.

۳-۱- روش مبتنی بر پرکاربرد بودن

در این روش، برچسب‌هایی که بیشترین کاربرد (فراوانی) را در شبکه‌های اجتماعی دارند، انتخاب خواهند شد. اما این روش مشکلات فراوانی دارد. یکی از این

مشکلات، وجود اشکال مختلف نحوی (کلمات هم‌خانواده) و مشکل دیگر، گسترده بودن منابع مرتبط با برچسب‌های پرکاربرد است. این مشکلات در نهایت منجر به افزایش اشتراک در برچسب‌های موجود در ابر برچسب می‌شود.

در این مقاله با توجه به اینکه در مقاله‌ی مارتین [۸] نشان داده شد که این روش از لحاظ دو معیار پوشش و اشتراک، کارایی لازم را ندارد، در ارزیابی این مقاله مدنظر قرار نمی‌گیرد.

۳-۲- روش مبتنی بر آنالیز نحوی

این روش [۸] که ترکیبی از آنالیز نحوی و خوشه‌بندی معنایی است از چهار بخش تشکیل شده است:

۱- آنالیز نحوی
۲- خوشه‌بندی معنایی با استفاده از تکنیک خوشه‌بندی سلسله مراتبی
۳- مرتب‌سازی خوشه‌ها براساس معیار پوشش زنجیره‌ای به صورت نزولی
۴- انتخاب برچسب‌ها با بیشترین پوشش زنجیره‌ای از هر خوشه

۳-۲-۱- آنالیز نحوی

بخش اول این روش شامل آنالیز نحوی است بدین صورت که برچسب‌هایی که براساس معیار فاصله‌ی لونه استین (تعداد حروفی که باید تغییر کنند تا یک برچسب به برچسب دیگر تبدیل شود) مشابه هستند، در یک خوشه قرار می‌گیرند.

پس از خوشه‌بندی برچسب‌ها با استفاده از معیار فاصله‌ی لونه استین، در هر خوشه برچسبی که بیشترین کاربرد (بیشترین پوشش منابع) را دارد به عنوان نماینده‌ی آن خوشه در نظر گرفته شده و بقیه‌ی برچسب‌های آن خوشه حذف می‌شوند.

۳-۲-۲- خوشه‌بندی معنایی با استفاده از تکنیک سلسله مراتبی

در این بخش از تکنیک خوشه‌بندی سلسله مراتبی استفاده شده است. این تکنیک بدین صورت عمل می‌کند که هر برچسب در ابتدا به عنوان یک خوشه در نظر گرفته می‌شود. سپس برای هر خوشه، معیار مشابهت با دیگر خوشه‌ها براساس فرمول (۱) محاسبه می‌شود:

$$DICE(t_i,t_j)=\frac{2.cocr(t_i,t_j)}{ocr(t_i)+ocr(t_j)} \qquad (۱)$$

$cocr(t_i,t_j)$: تعداد منابع مشترک بین برچسب t_i و t_j

$ocr(t_i)$: تعداد منابع پوشش داده شده توسط برچسب t_i

$ocr(t_j)$: تعداد منابع پوشش داده شده توسط برچسب t_j

پس از محاسبه‌ی مشابهت، خوشه‌ی که با خوشه‌ی در نظر گرفته شده، بیشترین مشابهت را دارد با آن خوشه ادغام شده و تشکیل یک خوشه‌ی جدید را می‌دهند. این ادغام تا زمانی ادامه پیدا می‌کند که تعداد خوشه‌ها برابر با تعداد برچسب‌های مورد نیاز برای ابر برچسب شود.

۳-۲-۳- مرتب‌سازی خوشه‌ها براساس معیار پوشش زنجیره‌ای به صورت نزولی

در این بخش، خوشه‌ها براساس ملاک پوشش زنجیره‌ای (میزان پوشش برچسب بدون در نظر گرفتن منابع پوشش داده شده‌ی قبلی) به صورت نزولی مرتب می‌شوند.

۳-۲-۴- انتخاب برچسب‌ها با بیشترین پوشش زنجیره‌ای از هر خوشه

این بخش از خوشه با بیشترین مقدار پوشش زنجیره‌ای شروع می‌کند و سپس تعداد برچسب‌های انتخابی برای هر خوشه را براساس مقدار پوشش زنجیره‌ای آن و اندازه‌ی ابر برچسب (تعداد برچسب‌های ابر برچسب) محاسبه می‌کند. فرض کنید این تعداد برابر با n باشد.

در ادامه مقدار پوشش زنجیره‌ای n تا از برچسب‌هایی که بیشترین پوشش زنجیره‌ای را در آن خوشه دارند به صورت جداگانه با مقدار آستانه (T) مقایسه شده، هر کدام که بزرگتر بود برای نمایش در ابر برچسب انتخاب می‌شود. این کار تا زمانی که تعداد برچسب‌های انتخابی برابر با تعداد برچسب‌های مورد نظر برای ابر برچسب شود، ادامه می‌یابد.

۳-۳- روش بهبودیافته مبتنی بر آنالیز نحوی

این روش که در واقع نسخه‌ی بهبودیافته‌ی روش قبلی [۸] است از چهار بخش تشکیل شده است:

۱- آنالیز نحوی بهبود یافته
۲- خوشه‌بندی معنایی با استفاده از تکنیک خوشه‌بندی سلسله مراتبی
۳- مرتب‌سازی خوشه‌ها براساس معیار پوشش زنجیره‌ای به صورت نزولی
۴- انتخاب برچسب‌ها با بیشترین پوشش زنجیره‌ای از هر خوشه

تنها تفاوت بین این روش و روش قبلی در بخش آنالیز نحوی می‌باشد. در واقع این روش، یک مشکل را در بخش آنالیز نحوی روش قبلی حل کرده است در ادامه به این مشکل و راه‌حلی که این روش به کار برده است، اشاره خواهیم کرد.

آنالیز نحوی در روش قبلی بدین صورت بود که معیار فاصله‌ی لونه استین برای هر جفت برچسب محاسبه می‌شد و جفت برچسب‌هایی که مقدار فاصله‌ی لونه استین برای آنها کوچکتر یا مساوی یک مقدار آستانه (D) (بیشترین تغییرات ممکن برای تبدیل یک برچسب به برچسب دیگر) بود در یک خوشه قرار می‌گرفتند. سپس در هر خوشه برچسبی که بیشترین کاربرد (بیشترین پوشش منابع) را دارد به عنوان نماینده‌ی آن خوشه در نظر گرفته شده و بقیه‌ی برچسب‌های آن حذف می‌شدند. (در این مقاله، با توجه به اینکه کلمات هم‌خانواده معمولا در زبان انگلیسی حداکثر در ۴ حرف اختلاف دارند، D را ۴ در نظر می‌گیریم.)

اما با توجه به اینکه هدف آنالیز نحوی قرار دادن کلمات هم‌خانواده در یک خوشه است می‌توان گفت که آنالیز نحوی به کار رفته در این روش در یک حالت خاص مغایر با هدف آنالیز نحوی است. بدین صورت که در بخش آنالیز نحوی این روش، معیار هم‌خوشه بودن دو برچسب تعداد حروفی است که باید تغییر کنند تا یکی از برچسب‌ها به یک برچسب دیگر تبدیل شود که این تعداد حروف باید کمتر از یک مقدار آستانه (D) باشد (D=4). اما این معیار همواره صحیح عمل نمی‌کند و با مشکلاتی مواجه است. یکی از این مشکلات این است که جفت برچسب‌هایی

که ۴ حرفی یا کمتر از ۴ حرفی هستند اما هم خانواده نیستند را هم خانواده در نظر می‌گیرد مثلا good و iran.

۳-۳-۱- آنالیز نحوی بهبود یافته

آنالیز نحوی روش بهبودیافته مشابه روش قبلی است. فقط با این تفاوت که جفت برچسب‌هایی که ۴ یا کمتر از ۴ حرفی هستند در آنالیز نحوی نادیده گرفته می‌شوند. به عبارت دیگر معیار فاصله‌ی لونه استین برای آن‌ها اعمال نمی‌شود و در نتیجه در یک خوشه نحوی قرار نمی‌گیرند. سه بخش بعدی این روش کاملا مشابه روش قبلی است.

۴- روش ارائه شده (روش مبتنی بر آنالیز معنایی)

روشی که در این مقاله ارائه شده است بر مبنای روش‌های قبلی [۱، ۸] است فقط با این تفاوت که از آنالیز معنایی به جای آنالیز نحوی در آن استفاده شده است. بنابراین این روش از چهار بخش زیر تشکیل شده است:

۱- آنالیز معنایی

۲- خوشه‌بندی معنایی با استفاده از تکنیک خوشه‌بندی سلسله مراتبی

۳- مرتب‌سازی خوشه‌ها براساس معیار پوشش زنجیره‌ای به صورت نزولی

۴- انتخاب برچسب‌ها با بیشترین پوشش زنجیره‌ای از هر خوشه

دراین روش جهت حل مشکلی که در بخش آنالیز نحوی دو روش قبلی وجود دارد، از آنالیز معنایی به جای آنالیز نحوی استفاده شده است. در ادامه این مشکل و راه حل آن را شرح خواهیم داد.

همانطور که ذکر شد، دو روش قبلی از معیار فاصله‌ی لونه استین برای خوشه‌بندی برچسب‌ها استفاده می‌کردند، اما با توجه به اینکه نمی‌توان گفت دو برچسب که اختلاف حروف آنها کمتر از یک مقداری خاص است، لزوما هم خانواده هستند، این معیار نمی‌تواند معیار مناسبی برای خوشه‌بندی کلمات هم‌خانواده باشد و ممکن است دو برچسب که هم‌خانواده نیستند را به اشتباه هم‌خانواده در نظر بگیرد یا بالعکس. به عنوان مثال اختلاف دو برچسب wanted و wanton کمتر از ۴ حرف است و هر دو بیش از ۴ حرف دارند، ولی هم‌خانواده نیستند، در حالی که هر دو روش قبلی این جفت برچسب را هم‌خانواده در نظر می‌گرفت.

۴-۱- آنالیز معنایی

این بخش تنها بخش تغییریافته‌ی روش مبتنی بر آنالیز معنایی نسبت به روش‌های مبتنی بر آنالیز نحوی و بهبود یافته مبتنی بر آنالیز نحوی می‌باشد. در این بخش به جای استفاده از معیار فاصله‌ی لونه استین، از پایگاه داده‌ی ورد نت به عنوان مبنای روابط معنایی استفاده شده است.

یکی از دلایل استفاده از پایگاه داده‌ی ورد نت، این است که این پایگاه در مقالات متعددی در زمینه‌ی بازیابی اطلاعات از جمله [۹، ۱۳] استفاده شده است. دلیل دیگر استفاده از این پایگاه داده، جامع بودن و و پشتیبانی از روابط معنایی مختلف در این پایگاه داده می‌باشد.

این بخش بدین صورت عمل می‌کند که در ابتدا برچسب‌ها به کلمات موجود در پایگاه داده‌ی ورد نت نگاشت می‌شود. سپس برچسب‌های نگاشت‌یافته که براساس رابطه‌ی معنایی اشکال مرتبط اشتقاقی به هم مرتبطاند، در یک خوشه‌ی معنایی قرار می‌گیرند. اگر برچسبی معادل هیچ یک از کلمات موجود در پایگاه داده‌ی ورد نت نباشد، به صورت یک خوشه‌ی تک‌عضوی در نظر گرفته می‌شود.

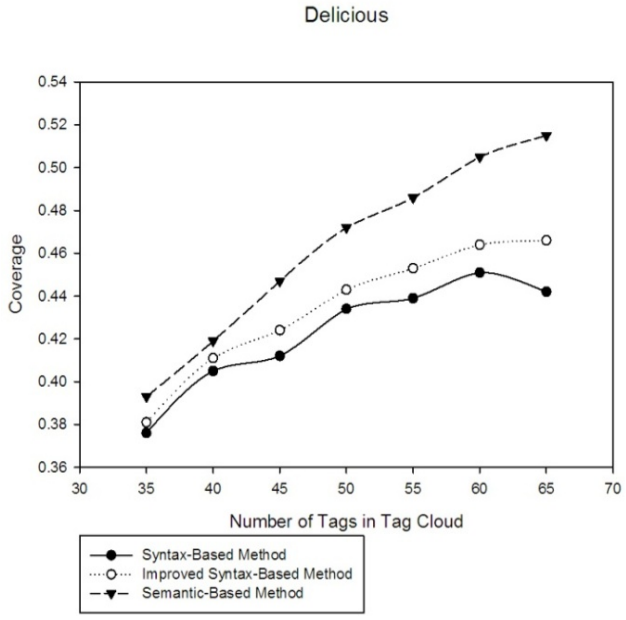
پس از خوشه‌بندی معنایی برچسب‌ها با استفاده از پایگاه داده‌ی ورد نت، در هر خوشه برچسبی که بیشترین کاربرد (بیشترین پوشش منابع) را دارد به عنوان نماینده‌ی آن خوشه در نظر گرفته شده و بقیه‌ی برچسب‌های آن خوشه حذف می‌شوند. سه بخش بعدی این روش کاملا مشابه روش‌های قبلی است.

۵- ارزیابی

جهت مقایسه روش ارائه شده (روش مبتنی بر آنالیز معنایی) با دو روش قبلی (روش مبتنی بر آنالیز نحوی و روش بهبودیافته مبتنی بر آنالیز نحوی)، هر سه روش را بر روی بخشی از دو پایگاه داده‌ی دلیشس و بیبسونومی پیاده‌سازی شده‌اند. بخش مرتبط با پایگاه داده‌ی دلیشس شامل ۳۳۱۱ عنصر (منبع) اطلاعاتی و ۲۷۰ برچسب و بخش پایگاه داده‌ی بیبسونومی شامل ۲۹۰۵ عنصر (منبع) اطلاعاتی و ۲۳۹ برچسب است. نتایج این ارزیابی در قالب اشکال و جداول در ادامه آمده است.

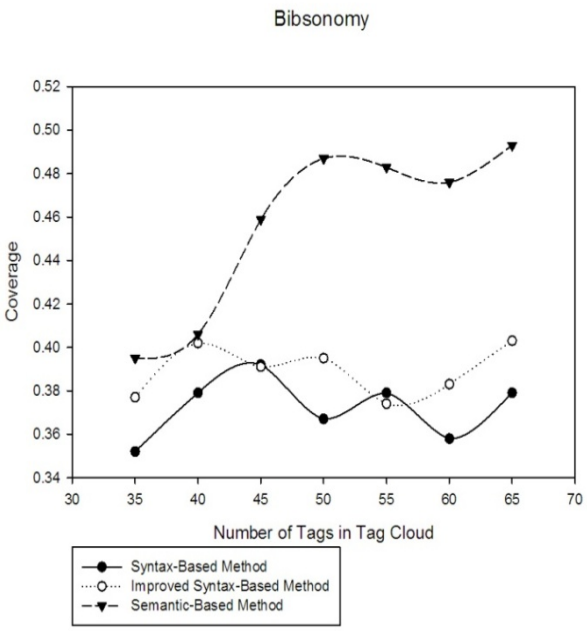
۵-۱- اشکال

نتایج در شکل ۱ و ۲ نشان می‌دهد که در پایگاه داده‌ی دلیشس در هر سه روش همراه با افزایش تعداد برچسب‌های انتخابی، پوشش هم‌افزایش می‌یابد در حالی که در پایگاه داده‌ی بیبسونومی، این برقرار نیست. در این پایگاه داده، روش ارائه شده تقریبا یک روند صعودی دارد، ولی نمودار دو روش قبلی با نوسان همراه است.

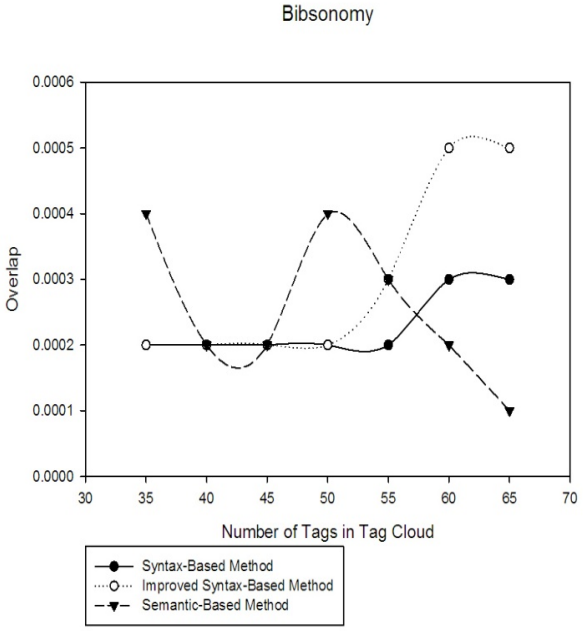


شکل ۱- نتایج پوشش برای روش‌های مبتنی بر آنالیز معنایی، بهبود یافته مبتنی بر آنالیز نحوی و مبتنی بر آنالیز نحوی در پایگاه داده‌ی دلیشس

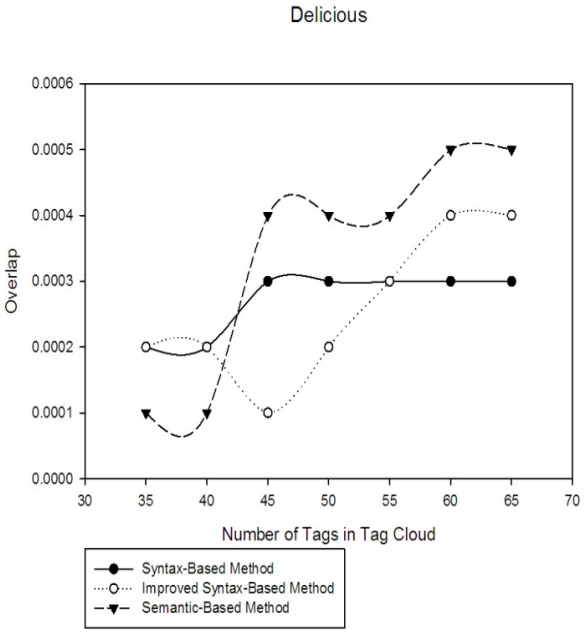
دلیل تفاوت در روند نمودارها در پایگاه داده‌های بیبسونومی و دلیشس را می‌توان تفاوت در خصوصیات برچسب‌های این دو پایگاه داده ذکر کرد. این خصوصیات می‌توانند تعداد کلمات هم‌خانواده و همچنین تاثیر برچسب‌های که دو روش قبلی (روش مبتنی بر آنالیز نحوی و روش بهبود یافته مبتنی بر آنالیز نحوی) به اشتباه در یک خوشه‌ی نحوی قرار می‌دهند، باشند.



شکل ۲- نتایج پوشش برای روش‌های مبتنی بر آنالیز معنایی، بهبود یافته مبتنی بر آنالیز نحوی و مبتنی بر آنالیز نحوی در پایگاه داده‌ی بیبسونومی



شکل ۴- نتایج اشتراک برای روش‌های مبتنی بر آنالیز معنایی، بهبود یافته مبتنی بر آنالیز نحوی و مبتنی بر آنالیز نحوی در پایگاه داده‌ی بیبسونومی



شکل ۳- نتایج اشتراک برای روش‌های مبتنی بر آنالیز معنایی، بهبود یافته مبتنی بر آنالیز نحوی و مبتنی بر آنالیز نحوی در پایگاه داده‌ی دلیشس

اشکال ۳ و ۴ میزان اشتراک روش‌های ارائه شده و قبلی را بررسی می‌کند. که نمودارهای موجود در آنها نشان می‌دهد که روند روش ارائه شده در پایگاه داده‌ی دلیشس افزایشی است، در حالی که در پایگاه داده‌ی بیبسونومی روند آن کاهشی است. همچنین در روش بهبودیافته مبتنی بر آنالیز نحوی در هر دو پایگاه داده، روند افزایشی است و در مورد روش مبتنی بر آنالیز نحوی در هر دو پایگاه داده تقریبا می‌توان گفت که روند ثابت است.

به طور کلی از نمودارهای شکل ۲ می‌توان نتیجه‌گیری کرد در هر سه روش و در هر دو پایگاه داده تغییرات اشتراک ناچیز بوده و در حد یک ده‌هزارم است.

۵-۲- محاسبه‌ی مقدار میانگین

در جدول ۱، مقدار میانگین معیار پوشش برای هر سه روش در دو پایگاه داده‌ی دلیشس و بیبسونومی محاسبه شده است. نتایج نشان می‌دهد که پوشش به طور میانگین در روش ارائه شده (مبتنی بر آنالیز معنایی) نسبت به دو روش قبلی در هر دو پایگاه داده افزایش یافته است. به عنوان مثال به طور میانگین پوشش روش مبتنی بر آنالیز معنایی نسبت به روش بهبودیافته مبتنی بر آنالیز نحوی در پایگاه داده‌ی دلیشس حدود ۳ درصد و در پایگاه داده‌ی بیبسونومی حدود ۷ درصد افزایش یافته است.

در جدول ۲، مقدار میانگین معیار اشتراک برای هر سه روش در دو پایگاه داده‌ی دلیشس و بیبسونومی محاسبه شده است. نتایج نشان می‌دهد که تفاوت مقدار اشتراک در هر دو پایگاه داده بسیار ناچیز و در حد یک ده‌هزارم می‌باشد.

جدول ۱- مقادیر میانگین پوشش برای روش‌های مبتنی بر آنالیز معنایی، بهبود یافته مبتنی بر آنالیز نحوی و مبتنی بر آنالیز نحوی

روش‌ها	دلیشس	بیبسونومی
روش مبتنی بر آنالیز نحوی	۰.۴۲۲	۰.۳۷۲
روش بهبود یافته مبتنی بر آنالیز نحوی	۰.۴۳۴	۰.۳۸۹
روش مبتنی بر آنالیز معنایی	۰.۴۶۲	۰.۴۵۷

جدول ۲- مقادیر میانگین اشتراک برای روش‌های مبتنی بر آنالیز معنایی، بهبود یافته مبتنی بر آنالیز نحوی و مبتنی بر آنالیز نحوی

روش‌ها	دلیشس	بیبسونومی
روش مبتنی بر آنالیز نحوی	۰.۰۰۰۲	۰.۰۰۰۲
روش بهبود یافته مبتنی بر آنالیز نحوی	۰.۰۰۰۲	۰.۰۰۰۳
روش مبتنی بر آنالیز معنایی	۰.۰۰۰۳	۰.۰۰۰۲

۵-۳- پیچیدگی زمانی

همانطور که اشاره شد، تنها تفاوت روش‌های مبتنی بر آنالیز معنایی، بهبود یافته مبتنی بر آنالیز نحوی و مبتنی بر آنالیز نحوی در چگونگی خوشه‌بندی کلمات هم‌خانواده می‌باشد.

در روش مبتنی بر آنالیز نحوی از معیار فاصله‌ی لونه استین جهت این خوشه‌بندی استفاده می‌شود. در این مقاله برای پیاده‌سازی این معیار از تابع LDistance (string s, string t) استفاده شده است که پیچیدگی زمانی آن به صورت زیر محاسبه می‌شود:

$$Time\ Complexity_1 = O(\binom{n}{2}(r*m)) \quad (۲)$$

r: تعداد حروف رشته‌ی s
m: تعداد حروف رشته‌ی t
n: تعداد کل برچسب‌ها

در روش بهبود یافته مبتنی بر آنالیز نحوی نیز مشابه روش اول از معیار فاصله‌ی لونه استین استفاده می‌شود فقط با این تفاوت که برای جفت برچسب‌های ۴ یا کمتر از ۴ حرفی، این معیار اعمال نمی‌شود. فرض کنید تعداد این حالتی که هر دو برچسب ۴ یا کمتر از ۴ حرفی باشند برابر با c باشد. لذا پیچیدگی زمانی خوشه‌بندی کلمات هم‌خانواده برابر است با:

$$Time\ Complexity_2 = O((\binom{n}{2}(r*m) - c(r*m-1)) \quad (۳)$$

ولی در روش مبتنی بر آنالیز معنایی از رابطه‌ی معنایی اشکال مرتبط اشتقاقی که در پایگاه داده‌ی وردنت تعریف شده است به جای معیار فاصله‌ی لونه استین برای خوشه‌بندی کلمات هم‌خانواده استفاده می‌شود که پیاده‌سازی خوشه‌بندی کلمات هم‌خانواده در روش سوم با استفاده از یک پرس‌وجو در پایگاه داده‌ای وردنت به ازای هر جفت برچسب انجام می‌پذیرد، بنابراین پیچیدگی زمانی این روش برابر است با:

$$Time\ Complexity_3 = O(\binom{n}{2}) \quad (۴)$$

پیچیدگی زمانی بقیه قسمت‌های هر سه روش که مشابه یکدیگر می‌باشند برابر است با $O(n^2)$. لذا جدول پیچیدگی زمانی به صورتی می‌باشد که در جدول ۳ آمده است.

جدول ۳ نشان می‌دهد که زمان روش مبتنی بر آنالیز معنایی نسبت به دو روش دیگر بهتر است. این به این دلیل است که این روش از پرس‌وجوی پایگاه داده‌ای به جای پردازش بر روی حروف برچسب‌ها استفاده می‌کند. در واقع هر چه تعداد حروف برچسب‌ها بیشتر باشد، عملکرد روش مبتنی بر آنالیز معنایی از لحاظ پیچیدگی زمانی نسبت به دو روش دیگر بهتر خواهد بود.

به طور کلی از نمودارها و جداول این بخش می‌توان نتیجه گرفت که روش مبتنی بر آنالیز معنایی نسبت به دو روش بهبود یافته مبتنی بر آنالیز نحوی و مبتنی بر آنالیز نحوی بهتر عمل می‌کند. این می‌تواند به دلیل نحوه‌ی خوشه‌بندی کلمات هم‌خانواده در این سه روش باشد، چون تنها تفاوت این سه روش در این بخش می‌باشد.

جدول ۳- مقادیر پیچیدگی زمانی برای روش‌های مبتنی بر آنالیز معنایی، بهبود یافته مبتنی بر آنالیز نحوی و مبتنی بر آنالیز نحوی

روش‌ها	پیچیدگی زمانی
روش مبتنی بر آنالیز نحوی	$O(\binom{n}{2}(r*m))$
روش بهبود یافته مبتنی بر آنالیز نحوی	$O((\binom{n}{2}(r*m) - c(r*m-1))$
روش مبتنی بر آنالیز معنایی	$O(\binom{n}{2})$

۶- نتیجه‌گیری و کارهای آینده

در این مقاله، یک روش برای بهبود دو معیار پوشش و اشتراک ارائه شد. تغییری که این روش نسبت به روش‌های قبلی [۱، ۸] ایجاد کرد، استفاده از آنالیز معنایی به جای آنالیز نحوی بود. روش ارائه شده همراه با روش‌های قبلی در بخشی از دو پایگاه داده‌ی بیسونومی و دلیشس پیاده‌سازی شد که نتایج حاکی از بهبود در هر دو معیار پوشش و عدم‌تغییر محسوس اشتراک در روش ارائه شده نسبت به روش‌های قبلی بود.

جهت بررسی بیشتر بر روی این سه روش، کارهای که در آینده می‌تواند انجام شود شامل معرفی یک معیار انتخاب برچسب جدید جهت بهبود معیارهای پوشش و اشتراک و همچنین پیاده‌سازی روش ارائه شده در پایگاه داده‌های دیگر موجود در شبکه‌های اجتماعی می‌باشد.

مراجع

[۱] م. ستاری، ک. زمانی‌فر و ن. نعمت‌بخش، "ارائه یک روش بهبود یافته مبتنی بر آنالیز نحوی برای دستیابی به اطلاعات در شبکه‌های اجتماعی،" در مجموعه مقالات هجدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران، ص ۲۳۲-۲۲۴، ۱۳۹۱.

[2] S. Bateman, C. Gutwin, and M. Nacenta, "Seeing Things in the Clouds: The Effect of Visual Features on Tag Cloud Selections," *Proc. ACM Conf. Hypertext and Hypermedia*, pp. 193-202, 2008.

[3] M. Gupta, L. Rui, Z. Yin, and J. Han, "Survey on Social Tagging Techniques," *ACM SIGKDD Explorations Newsletter*, vol. 2, no. 1, pp. 58-72, 2010.

[4] P. Ken, and C. Russell, "Tag grid: Supporting Collaborative and Fuzzy Multidimensional Queries of Tagged Datasets," *Proc. IEEE Intl Conf. Information Reuse and Integration*, pp. 364-376, 2010.

[5] P. Ken, and C. Russell, *Tag grid: Supporting Multidimensional Queries of Tagged Datasets*, Vienna: Springer, 2012.

[6] B. Kuo, T. Hentrich, M. Good, and D. M. Wilkinson, "Tag Clouds for Summarizing Web Search Results," *Proc. IEEE Intl Conf. World Wide Web*, pp. 1203-1204, 2008.

[7] D. Laniado, D. Eynard, and M. Colombetti, "Using WordNet to turn a folksonomy into a hierarchy of concepts," *Proc. IEEE Intl Workshop Semantic Web Application and Perspectives*, pp. 752-758, 2007.

[8] M. Leginus, P. Dolog, L. Ricardo, and F. Durao, "Methodologies for Improved Tag Cloud Generation with Clustering," *Proc. IEEE Intl Conf. Web Engineering*, pp. 61-75, 2012,

[9] D. R. Millen, and J. Feinberg, "Using Social Tagging to Improve Social Navigation," *Proc. IEEE Intl Workshop on the Social Navigation and Community Based Adaptation Technologies*, pp. 61-65, 2006.

[10] K. Musial, and K. Przemyslaw, "Social networks on the internet," *Journal of World Wide Web*, vol. 16, no. 1, pp. 31-72, 2013.

[11] A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen, "Getting Our Head in the Clouds: Toward Evaluation Studies of Tag Clouds," *Proc. ACM Intl Conf. Human Factors in Computing Systems*, pp. 995-998, 2007.

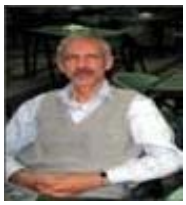
[12] J. Sinclair, and M. Cardew-Hall, "The Folksonomy Tag Cloud When Is It Useful," *Journal of Information Science*, vol. 13, no. 2, pp. 15-29, 2008.

[13] V. Snasel, M. Pavel, and J. Pokorny, "WordNet Ontology Based Model for Web Retrieval," *Proc. IEEE Intl Workshop on Web Information Retrieval and Integration*, pp. 20-24, 2005.

[14] P. Veneti, G. Kotrika, and H. Garcia-Molina, "On the Selection of Tags for Tag Clouds," *Proc. IEEE Intl Conf. Web Search and Data Mining*, pp. 835-844, 2011.

محمد ستاری در حال حاضر دانشجوی دکتری دانشگاه اصفهان می‌باشد. وی در سال ۱۳۸۹ مدرک کارشناسی مهندسی کامپیوتر گرایش نرم‌افزار خود را از دانشگاه آزاد اسلامی واحد نجف‌آباد دریافت کرد و در سال ۱۳۹۲ موفق به اخذ مدرک کارشناسی‌ارشد در همان گرایش از دانشگاه آزاد اسلامی واحد نجف‌آباد شد. وی در همان سال در مقطع دکتری پذیرفته شد و هم‌اکنون مشغول به تحصیل در این مقطع در گرایش نرم‌افزار در دانشگاه اصفهان می‌باشد. در حال حاضر فعالیت‌های پژوهشی‌اش متمرکز بر شبکه‌های اجتماعی، تکنیک‌های بازیابی اطلاعات و وب معنایی می‌باشد.
آدرس پست‌الکترونیکی ایشان عبارت است از:
mohammadsattari.38500658@gmail.com

کامران زمانی‌فر دانشیار گروه مهندسی کامپیوتر دانشگاه اصفهان می‌باشد. ایشان دکتری خود را در زمینه‌ی علوم کامپیوتری در سال ۱۳۷۵ از دانشگاه لیدز انگلستان اخذ نمودند و از همان سال به عنوان عضو هیئت علمی در دانشگاه اصفهان مشغول به کار شدند. زمینه‌های تحقیقاتی ایشان محاسبات ابری، محاسبات همه جا حاضر، سیستم‌های توزیع شده، محاسبات فراگیر و محاسبات نرم می‌باشند.
آدرس پست‌الکترونیکی ایشان عبارت است از:
zamanifar@eng.ui.ac.ir



ناصر نعمت‌بخش استادیار گروه مهندسی کامپیوتر دانشگاه اصفهان می‌باشد. ایشان از سال ۱۳۵۶ به عنوان عضو هیئت علمی در دانشگاه اصفهان مشغول به کار شدند و دکتری خود را در زمینه‌ی مهندسی کامپیوتر در سال ۱۳۶۸ از دانشگاه برادفورد انگلستان اخذ نمودند. زمینه‌های تحقیقاتی ایشان ارزیابی کارایی، مدل‌سازی قابلیت اطمینان، سیستم‌های مبتنی بر عامل و مهندسی نرم‌افزار می‌باشند.
آدرس پست‌الکترونیکی ایشان عبارت است از:
nemat@eng.ui.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۹۲/۱/۳۰
تاریخ اصلاح: ۹۲/۶/۵
تاریخ قبول شدن: ۹۲/۱۱/۲۵
نویسنده مرتبط: محمد ستاری، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه اصفهان، اصفهان، ایران.

[1]Tag Cloud
[2]Tag