دوره هفتم، شماره اول، بهار و تابستان ۱۳۹۳

مجله فناوری اطلاعات در طراحی مهندسی
Information Technology on Engineering Design
www.js.mshdiau.ac.ir/jte

پارتیشن‌بندی زمانی گراف برای خوشه‌بندی در سیستم‌های برچسب دار

علی‌اکبر الهداغی*[۱]، مهرداد جلالی[۲]، سید جواد سید مهدوی‌چابک[۳]

[۱،۲،۳]دانشگاه آزاد اسلامی، واحد مشهد، گروه کامپیوتر، مشهد، ایران

چکیده:

در عصر حاضر رشد اطلاعات در دنیای وب ۲.۰ به حدی بوده است که به علت حجم وسیع داده‌ها و اطلاعات و همچنین تغییر برخی مفاهیم در طی گذر زمان، اطلاعات غیرضروری و غیر مرتبط بسیاری با آنچه کاربران به دنبال آن هستند، به وجود آمده است.در این مقاله ما برای حل این مشکل، خوشه‌بندی زمانی برچسب‌ها در سیستم‌هایی که از برچسب به‌عنوان یک متاداده استفاده کردهو در طی زمان در حال تغییر می‌باشند را پیشنهاد می‌دهیم. این عمل از طریقپارتیشن‌بندی زمانی گراف برچسب‌هاست، به‌طوری‌که با تغییر دادن وزن‌های مشابهت برچسب‌ها در طی زمان، خوشه‌بندی ما هم دچار تغییر شده و خودش را با تغییرات وفق می‌دهد. برای نشان دادن کارایی این روش، ما آن را بر روی مجموعه داده سایت متافیلتر اجرا کرده و با روش‌های مشابه مقایسه کردیم. نتایج نشان‌دهنده این مطلب است که روش پیشنهادی ما،F-Measureمیانگین خوشه‌ها را ۲۴٪ نسبت به بهترین روش مشابه در طول زمان، بهبود بخشیده است و ازنظر مفهومی، هم با مفاهیم گذشته وهم با مطالب روز در ارتباط است.

واژه‌های کلیدی:

برچسب، خوشه‌بندی، Temporal, Graph Partitioning, Folksonomy

*عهده دار مکاتبات

دانشگاه آزاد اسلامی، واحد مشهد، گروه کامپیوتر، مشهد، ایران

پست الکترونیکی: a.a.aldaghi@mshdiau.ac.ir

۱- مقدمه

در عصر حاضر تبادل و رشد دانش به کمک فناوری اطلاعات در حال افزایش نمایی است.یکی از مهم‌ترین عوامل این افزایش اشتراک اطلاعات بین کاربران وب است. این اشتراک‌گذاری در زمان‌های اخیرتسهیل و افزایش‌یافته است که در اصطلاح، وب ۲.۰ نامیده می‌شود [۱]. وب ۲.۰ بر پایه چند اصل اساسی بناشده است. از این میان سه اصل زیر ماهیت وب ۲.۰ را که برخلاف فناوری‌های دیجیتال و وب ۱.۰ است را تشکیل می‌دهد.۱. وب یک جایگاه[1] است.۲. کاربر داده‌های خود را کنترل می‌کند. ۳. کاربر نقش مشارکتی دارد [۲]. ترکیب این سه اصل تولید برنامه‌های مبتنی بر وب را به شیوه‌ای ممکن ساخته است که وب را به یک وب مشارکتی تبدیل می‌کند.

وب ۲.۰ جنبه‌های متعددی دارد که نمونه‌های معمولی از مشارکت کاربران را انعکاس می‌دهد. این جنبه‌ها عبارت‌اند از: وبلاگ نویسی[2]، برچسب زدن[3]، نشانه‌گذاری‌های اجتماعی[4]، شبکه‌های اجتماعی[5]، پادکست و غیره [۱].در این میان، برچسب[6]ها به‌عنوان یک متاداده[7] در نمایش خلاصه محتوای هر سند بسیار حائز اهمیت هستند.

برچسب‌ها در ضمن سادگی، توسط کاربران برای توصیف و یا حاشیه‌نویسی از هر نوع منبع و یا سندی برای بازیابی در آینده اختصاص داده‌شده‌اند. برچسب‌های ارسال‌شده توسط گروهی از کاربران Folksonomy را تشکیل می‌دهد[۳]. هدف از برچسب‌گذاری در یک فُکسونومی این است که جستجو، حرکت بین اطلاعات و درک یک بدنهٔ اطلاعات، در طول زمان و به تدریج برای کاربران آسان شود. از آنجا که فُکسونومی‌ها عموماً در محیط‌های مبتنی بر اینترنت به وجود آمده‌اند، کاربران می‌توانند دریابند کهیک برچسب معین را چه کسی استفاده می‌کند و همچنین او از چه برچسب‌های دیگری استفاده می‌کند. به این ترتیب کاربران می‌توانند به درک متقابلی از مجموعهٔ برچسب‌هایی برسند که دیگرانبرای تفسیر،درونی سازی و درک یک بخش از اطلاعات نیاز دارند. این مورد می‌تواند در سیستم‌های مدیریت دانشانسان‌محور یک گام بزرگ روبه‌جلو باشد؛ چراکه قابلیت افراد در ساختاردهی صحیح دانش و همچنین یافتن محتواهای مرتبط را بالا می‌برد. این اصل، دریافتن اطلاعات مرتبط[8] نام دارد. بخش عمده‌ای از جاذبه فُکسونومی در توانمندی ذاتی و راحتی آن دریافتن اطلاعات نسبت به ابزارهای سنتی جستجو مانند موتورهایجستجوست [۴].

ساختار این مقالهبدین صورت است:

بخش ۲. مرور کارهای انجام‌شده در زمینهٔ خوشه‌بندی متون و سیستم‌های برچسب دار: در این بخش کارهای گذشتگان را که مرتبط با زمینه خوشه‌بندی اطلاعات صفحات وب با رویکرد استفاده از برچسب‌ها است را بررسی کرده و مورد تجزیه و تحلیل قرار می‌دهیم.

بخش ۳. روش پیشنهادی: در این بخش روش پیشنهادی خود را مطرح کرده و شیوه پیاده‌سازی و الگوریتم‌های موجود و نوآوری خود را شرح خواهیم داد.

بخش ٤. نتایج و ارزیابی: در این بخش نتایج و مقایسه‌های خود را با روش‌های موجود مشابه بررسی و ارزیابی خواهیم کرد.

بخش ٥: نتیجه‌گیری و پیشنهادها:در اینبخش به نتیجه‌گیری‌ها و پیشنهادهای خود اشاره خواهیم نمود.

[1]- Platform
[2]- Blogging
[3]- Tagging
[4]- Social Bookmarking
[5]- Social Networking
[6]- Tag
[7]- Metadata
[8]- Pivot Browsing

۲- مروری بر کارهای گذشته

ازجمله مقالاتی که از برچسب به عنوان داده برای خوشه‌بندی خود استفاده کرده‌اند، می‌توان به کارهای [۳،۵،۶،۷،۸،۹،۱۰،۱۱ و ۱۲] اشاره نمود که آنها را در ۳ قالب ذیل تقسیم‌بندی می‌کنیم:

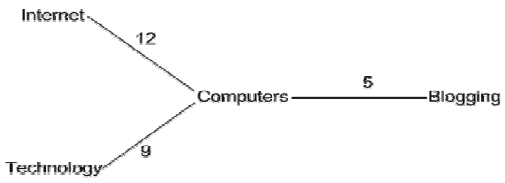
۱. کارهایی که مربوط به خوشه‌بندی برچسب‌ها و یا استفاده از برچسب‌ها برای دسته‌بندی کردن اسناد و یا کمک به موتورهای جستجو می‌باشند، مانند [۳، ۵، ۷و ۸].

۲. مجموعه کارهایی که علاوه بر خوشه‌بندی برچسب‌ها به پارامتر زمان آن‌هم توجه داشته‌اند و در خوشه‌بندی از آن متأثر بوده‌اند، ازجمله مقالات و کارهای [3، ۵ و ۷].

۳. کارهایی که مربوط به خوشه‌بندی اسناد متنی است؛ مانند [۵، ۱۱ و ۱۲] که در این بین مقالات [5 و ۱۱] از برچسب‌ها نیز کمک گرفته و یا به نوعی کار خود را با آن مقایسه کرده‌اند.

۲-۱- خوشه‌بندی برچسب‌ها

در [5]نویسندگان ابتدا به تفاوت‌های بین وب ۱.۰ و وب ۲.۰ اشاره کردند و خاطرنشان کردند که در وب ۲.۰ خلاف وب ۱.۰-که اکثراً خدمات دهنده‌ها مطالب را در اختیار کاربران می‌گذاشتند- این کاربران هستند که خود برای سایر کاربران مطلب تولید می‌کنند. در ادامه کارهای مرتبط با خودشان را مقایسه کردند و اشاره به این موضوع داشتند که اکثراً در حجم وسیع دچار نفرین ابعاد می‌شوند. کار اصلی این مقاله مبتنی بر برچسب‌هایهر پست داخل وبلاگ-هاست؛ به طوری که بر اساس ارتباط بین برچسب‌های گراف وزن‌داری را ایجاد می‌کند و وزن یال‌ها در واقع برابر با تعداد پستی است که دو برچسب موجود در دو طرف یال را در بین برچسب‌های خود آورده است شکل (۱). سپس بر اساس پارتیشن‌بندی گراف خوشه‌بندی خود را انجام می-دهد و نام این کار را خرد جمعی می‌نامد.البته آنها این کار را با چند روش دیگر مورد بررسی قرار می‌دهند.



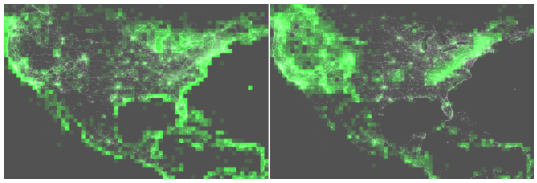
شکل (۱): یک نمونه از گراف ارتباط برچسب‌ها [۱]

در کار بعدی سوین انفینسن و همکاران در [۳]، به کاربرد ابزارهای وب ۲.۰ در کتابخانه‌های دیجیتال اشاره نمودند که باعث می‌شود مفهوم کتابخانه ۲.۰ پدید آید. آنها نشان دادند که با اشتراک‌گذاری عناوین کتاب‌ها و تعیین برچسب‌ها برای هر یک توسط اساتید، ناشرین و دانش‌آموزان چطور می‌توان به دسترسی سریع‌تر کاربران برای جستجوی یک کتاب خاص کمک کرد. در [۱۱]Brooksو Montanez تأثیر استفاده از برچسب‌گذاری توسط انسان را برای خوشه‌بندی پست‌های وبلاگ تحلیل کردند. آنها دریافتند که هر چند برچسب‌ها برای طبقه‌بندی پست‌های وبلاگ به دسته گسترده مؤثر هستند، ولی آنها کمتر در محتوای خاصی از پست، مؤثر نشان داده می‌شوند. سپس آنها استخراج یک مجموعه متناهی از بیشترین کلمات مهم از مقاله خودشان را به وسیله معیار *tf-idf* پیشنهاد دادند و ذکر کردند که استفاده از آن‌ها برای طبقه‌بندی بازده‌ای بهتری دارد.

۲-۲- خوشه‌بندی زمانی برچسب‌ها

هایپن ژانگ و همکارانش در [6] از روی ارتباط برچسب‌ها با مکان و زمان،خوشه‌بندی‌های خود را انجام دادند. مجموعهدادهٔ آنها شامل تصاویری است که هر تصویر

علاوه بر برچسب‌هایی برای معرفی آن تصویر، دارای عنوان، زمان ثبت و موقعیت جغرافیایی می‌باشد.



شکل (۲): توزیع جغرافیایی برای برچسب "beach" (تصویر چپ) و "mountains"(تصویر راست)[6]

آنها به ۵ روش مختلف برچسب‌های موجود را خوشه‌بندی می‌کنند.

۱. اولین روش آنها استفاده از توزیع جغرافیایی برچسب‌های تصاویر است. در شکل (۲) نمونه‌ای از توزیع جغرافیایی تصاویر گرفته‌شده با دو برچسب مختلف نشان داده شده است. آنها از روی این توزیع یک بردار ویژگی تحت عنوان بردار ویژگی Geo-spatial ایجاد می‌کنند.

۲. دومین روش ایجاد یک بردار ویژگی زمانی از روی توزیع بازه‌های زمانی استفاده از برچسب‌هاست.

۳. روش دیگر خوشه‌بندی آنها استفاده از دو خاصیت توزیع مکانی و زمانی، یعنی ترکیب هر دو روش قبلی است.

۴. استفاده از ویژگی وقوع مشترک زوج برچسب‌ها است. شیوهٔ محاسبه آن هم استفاده از رابطه(۱) است:

$$\text{mutual\_info}(t_1, t_2) = \log\left(\frac{\text{co\_occur}(t_1, t_2)}{\text{occur}(t_1) \times \text{occur}(t_2)}\right) \qquad (۱)$$

۵. فقط استفاده از $\text{co\_occur}(t_1, t_2)$

آنها برای ارزیابی کار خود سایتی طراحی کرده و از مردم دعوت کردند که به نتایج خوشه‌بندی‌هایایجادشدهٔ هر روش امتیاز بدهند؛ به عبارت دیگر، ارتباط برچسب‌ها را از طریق نظرسنجی عمومی حاصل کردند. برای مقایسه هم از معیارهای متریک precision و recall استفاده کردند.

۳-۲- خوشه‌بندی‌های متنی

در [۱۲] نویسنده به بررسی چند روش مفید برای خوشه‌بندی وبلاگ‌ها با توجه به متن موجود در آنها می‌پردازد. این روش‌ها عبارتند از: LSI[1], VSM[2], K-Means و FCM[3]. کار اصلی این پروژه ارائه یک اصلاح الگوریتم بهینه انتخابی برای خوشه‌بندی وبلاگ است. در انتها در یک جدول الگوریتم‌های فوق را با همدیگر مقایسه می‌کند. در جمع‌آوری داده از تعدادی وبلاگ‌هایی که قبلاً به صورت آفلاین گرفته شده است، استفاده می‌کند. سپس در مرحله تبدیل‌ها ضمن حذف برچسب‌های HTML، کلمات ربطی مانند "the", "on", "is", "of" را از داده‌ها حذف می‌کند. همچنین معادل‌سازی را برای برخی کلمات که در جمله از شکل‌های دیگر آن استفاده شده است، انجام می‌دهد. مثلاً کلمات iron معادل irons است و یا به‌جای damaged از شکل ساده آن یعنی damage استفاده می‌کند.

برای وزن دهی به کلماتی که در متن ظاهرشده‌اند از روش TF-IDF استفاده می‌کند.

۳- روش پیشنهادی

در این بخش روش کار خود که پارتیشن‌بندی گراف زمانی برای خوشه‌بندی در سیستم‌های برچسب دار است را شرح می‌دهیم. قبل از شروع، ذکر این نکته لازم است که خوشه‌بندی ما در این سیستم‌ها مستقل از کاربر است و فقط با منابع و برچسب‌گذاری سایر کاربران ارتباط دارد و به

[1]- Latent Semantic Indexing
[2]- Vector Space Model
[3]- Fuzzy C-Means

نوعی باخرد جمعی همراه است. ماژول‌های کاری به ترتیب ذیل خواهند بود:

۱. ماژول خزنده[1] صفحه‌های وب (جمع‌آوری مستندات برچسب دار)

۲. ماژول آنالیز سند شامل: استخراج متن، پیش‌پردازش اولیه، استخراج برچسب‌ها

۳. ماژول خوشه‌بندی برچسب‌ها

۴. ماژول سازمان‌دهی مجدد خوشه‌ها

۳-۱- ماژول خزنده صفحه‌های وب

در ابتدا لیستی از سایت‌های مرجع تهیه می‌شود که در آن کاربران ضمن گذاشتن منبع (لینک، متن، تصویر و ...) برای آنها برچسب‌هایی را نیز قرار می‌دهند. سپس مشابه [5] خزنده‌ای طراحی شده است که با رجوع به ریشهٔ تک‌تک سایت‌های مرجع، به صورت اول سطح شروع به خزیدن می‌کند.

۳-۲- ماژول آنالیز سند

ورودی این ماژول فایل حاوی تگ‌های HTML است و خروجی آن برچسب‌های موجود در آن صفحه. در این مرحله ما از پیش مکان قرار گرفتن برچسب‌ها را می‌دانیم و علت این است که قبلاً یک فرد خبره ساختار صفحات مرجع را بررسی کرده و از مکان‌های قرار گرفتن مشخصات منبع مطلع است. به همین جهت این ماژول با دریافت صفحه وب مورد نظر به سراغ تگ‌های html از پیش تعیین‌شده می‌رود. سپس نام و مشخصات آن منبع را به همراه برچسب-های آن، در بانک اطلاعاتی با زمان ثبت خزیدن ذخیره می-کند.

۳-۳- ماژول خوشه‌بندی برچسب‌ها

در خوشه‌بندی برچسب‌ها مطابق [5، ۸ و ۹] ابتدا روش مشابهت خود را تعیین می‌کنیم. ما روشی را انتخاب می‌کنیم که قرار است مشابهت بین دو برچسب را نشان دهد و در ضمن نرمالیزه نیز باشد؛ لذا از معادله (۲)استفاده می‌کنیم.

$$Sim(Tag_i, Tag_j) = \frac{|A \cap B|}{|A \cup B|} \qquad (۲)$$

در معادله(۲) A مجموعه سندهایی است که دارای Tag_i و Bمجموعه سندهایی است که دارای Tag_j می‌باشند. در ادامه برای انجام خوشه‌بندی، برچسب‌هایی را که قرار است در خوشه‌بندی شرکت کنند، انتخاب می‌کنیم. روش انتخاب هم بدین صورت است که بر اساس فراوانی استفاده از برچسب‌ها، آنها را به صورت نزولی مرتب کرده و K-تایابتدای آن را انتخاب می‌کنیم. پس از انتخاب برچسب‌ها جهت خوشه‌بندی، ماتریس مشابهت پایین مثلثی‌ای بر اساس معادله(۲) تشکیل می‌دهیم.روش خوشه‌بندی ما استفاده از پارتیشن‌بندی گراف است، به طوری که پس از تشکیل گراف از روی ماتریس فوق، عددτ را به عنوان حد آستانه در نظر گرفته و یال‌های کمتر از آن حد را حذف می‌کنیم. حال به دو روش Clique یا Single Link می‌توانیم عمل پارتیشن‌بندی را روی گراف انجام دهیم.

۳-٤- ماژول سازمان‌دهی مجدد خوشه‌ها

تا این مرحله ما یک مرحله خوشه‌بندی کامل را انجام دادیم ولی پس از این نیاز است که خوشه‌ها به‌روز نگه داشته شوند که به روش ذیل عمل می‌کنیم:

۱. در هر بازه زمانی T کسر مشخص ازیال‌های گراف مطابق معادله(۳) کم می‌کنیم. این عمل به علت آن است که

[1]- Crawler

حضور برچسب‌هایی که درگذشته‌ای دور بوده‌اند، در این مرحله، خوشه‌بندی کمرنگ شود.

$$Sim(Tag_i, Tag_j) = \frac{\beta \times Sim(Tag_i, Tag_j)}{\alpha + \beta} \quad \forall (Tag_i, Tag_j) \in M, i < j \qquad (۳)$$

که در معادله فوق، M؛ ماتریس مشابهت اصلی، α؛ ضریب به‌روزرسانی و β ؛ ضریب به یادآوری گذشته است.

۲. عملیات خزنده را مجدداً از سر گرفته و مراحل آن را تکرار می‌کنیم تا این دفعه برچسب‌های جدیدی به دست آیند. اگر تعداد برچسب‌های جدید از حد عددی مانند xبیشتر بودند به مرحلهٔ بعد می‌رویم؛ وگرنه به مرحلهٔ ۵ می‌رویم.

۳. در این مرحله – مشابه ماژول خوشه‌بندی اولیه – K_2تا از پرکاربردترین برچسب‌ها که فراوانی آنها از حد عددی xبیشتر باشد را انتخاب کرده و درجه شباهت آنها را مطابق معادله (۴) محاسبه می‌کنیم:

$$Sim_{new}(Tag_u, Tag_v) = \frac{|A_t \cap B_t|}{|A_t \cup B_t|} \qquad (۴)$$

که در آن A_t؛ مجموعه سندهای جدید کشف‌شده در زمان t است که دارای Tag_u و به همین ترتیب B_t؛ مجموعه سندهای جدید کشف‌شده در زمان t است که دارای Tag_v می‌باشند.

۴. مقادیر درایه‌هایی ماتریس جدید را با ماتریس اصلی مطابقت می‌دهیم. اگر زوج برچسبی هم در ماتریس اصلی حضور داشت و هم در ماتریس جدید، آنگاه درایه‌ای ماتریس اصلی را مطابق معادله(۵) به‌روز می‌کنیم.

$$Sim(Tag_i, Tag_j) = Sim(Tag_i, Tag_j) + \frac{\alpha \times Sim_{new}(Tag_u, Tag_v)}{\alpha + \beta} \quad \forall (Tag_i, Tag_j) \in M, \forall (Tag_u, Tag_v) \in M_{new} \; and \; Tag_i == Tag_u, Tag_j == Tag_v, \; i < j, \; u < v \qquad (۵)$$

که در آن:

M_{new}؛ ماتریس مشابهت جدید به دست آمده توسط معادله (۴) است. M؛ ماتریس مشابهت اولیه است.

Tag_iو Tag_j؛ زوج برچسب‌های موجود در ماتریس Mاست. Tag_uو Tag_v؛ زوج برچسب‌های موجود در پیمایش جدید هستند.

اگر چنانچه زوج برچسب جدیدی نسبت به ماتریس اولیه پدید آمده بود با توجه به معادله (۶)آن را به ماتریس اولیه اضافه می‌کنیم.

$$Sim(Tag_i, Tag_j) = \frac{\alpha \times Sim_{new}(Tag_u, Tag_v)}{\alpha + \beta} \quad \forall (Tag_i, Tag_j) \notin M, \forall (Tag_u, Tag_v) \in M_{new}, u < v \qquad (۶)$$

۵. پس از تغییر ماتریس مشابهت، گراف مشابهت را تشکیل داده و با پارتیشن‌بندی مجدد آن با توجه به حد آستانه τ،عمل خوشه‌بندی رابرای تشکیل و یا حذف خوشه-ها بررسی می‌کنیم که این شیوه را پارتیشن‌بندی گراف زمانی می‌نامیم.

۶. این عملیاتدائماً می‌تواند در حال اجرا باشد که دلیل آن هم توقف نیافتن رشد اطلاعات جدید در وب است.

٤. نتایج و ارزیابی

٤-١- انتخاب مجموعه داده و استخراج برچسب‌ها

مجموعه داده ما برگرفته از سایت MetaFilter[۱۳] می‌باشد. این مجموعه داده شامل شماره لینک به همراه برچسب‌های ثبت‌شده برای آن توسط کاربر و همچنین زمان ثبت آن‌که بین بازه سال ۲۰۰۵ تا ۲۰۱۲ میلادی می‌باشد، است. درجدول (۱)جزییات این مجموعه داده شرح داده شده است و در جدول (۲) تعدادی از پرکاربردترین برچسب‌ها آورده شده است.

جدول (۱): جزییات مجموعه داده MetaFilter

تعداد لینک‌ها	۱۰۲۹٤٥
تعداد برچسب‌ها	۱۲۰۷۰٥
تعداد برچسب‌های نسبت داده‌شده به لینک‌ها	٥۷٥۱٤۱

جدول (۲): تعدادی از پرکاربردترین برچسب‌ها در مجموعه داده MetaFilter

برچسب	تعداد لینک نسبت داده‌شده	تعداد ماه میلادی مورد	تعداد روز مورد استفاده
brokenlink	8450	27	255
art	4810	85	1934
history	2741	85	1414
video	2427	85	1450
film	1933	85	1141
music	5844	85	2054
War	1836	85	913
politics	3618	85	1390
iraq	1718	80	776
internet	1461	85	708

٤-٢- پیش‌پردازش اولیه

پس از بررسی اولیه داده‌ها متوجه شدیم که یکی از برچسب‌ها با وجود پر تکرارترین برچسب، ولی گویا مفهوم و استفاده خاصی ندارد. اگر به جدول (۲) مجدداً نگاهی بیندازید متوجه خواهید شد که برچسب "brokenlink" که به معنی لینک خراب یا شکسته است، ارتباطی با سایر برچسب‌ها ندارد و خود سایت اصلی در صورتی که لینک مورد نظر خراب بوده، این برچسب را قرار داده است.لذا ما نیز از این برچسب در تمامی مراحل عملیات خوشه‌بندی صرف‌نظر می‌کنیم.

٤-٣- خوشه‌بندی اولیه از طریق پارتیشن‌بندی گراف مشابهت

همان‌طور که بیان شد، خوشه‌بندی ما قرار است مبتنی بر تغییرات زمان عمل کند؛ لذا ابتدا خوشه‌بندی خود را به دو دوره تقسیم می‌کنیم:

۱. دورهٔ اول مجموعه داده‌تا انتهای سال ۲۰۰۷،

۲. دوره ۲تا ۵۰ مجموعه داده از ابتدای سال ۲۰۰۸ تا ماه اول ۲۰۱۲ به تفکیک هرماه یک دوره.

عملیات اولیه خوشه‌بندی ما بر روی مجموعه داده اولی یعنی ماقبل ۲۰۰۸ خواهد بود و برای خوشه‌بندی زمانی از ترکیب مجموعه اول و دوم و با حرکت بازه‌های زمانی ماه به ماه استفاده می‌کنیم. منطق این کار بدین دلیل است که فرض می‌کنیم یک مجموعه داده‌ای به ما داده شده است و حال از این به بعد قرار است با طی زمان، خوشه‌بندی را انجام دهیم.برای شروع کار ما K_1=200 برچسب از پرکاربردترین برچسب‌های دوره اول را انتخاب می‌کنیم.در این مجموعه داده، برخی مشخصات درجه تشابه به دست آمده برای ۲۰۰ برچسب مورد نظر مطابق جدول (۳) است.

جدول (۳): برخی مشخصات درجه تشابه به دست آمده برای کل مجموعه داده

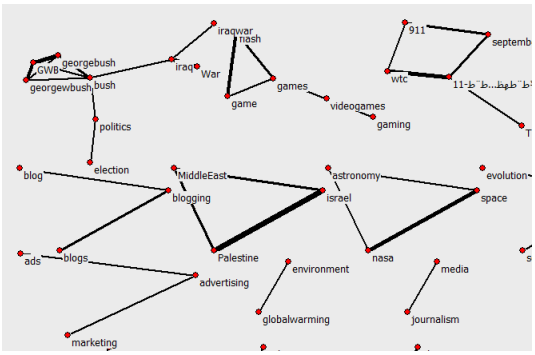
Min:	**0.0001**
Max:	**0.3326**
Mean:	**0.0046**
Variance:	**0.0001**
Std.:	**0.0107**

در ادامه از روی ۲۰۰ برچسب انتخابی یک ماتریس مشابهت پایین مثلثی تشکیل داده و درایه‌های آن را با توجه به معادله (۲)به دست می‌آوریم. برای نیل به این عمل برای هر زوج برچسب هم باید مجموع تعداد لینک‌هایی که دو برچسب را به عنوان برچسب خوددارند و هم مجموع تعداد لینک‌هایی که یا برچسب ۱ یا برچسب ۲ را داشته باشند به دست آورد. در جدول (۴)نمونه محاسبه درجه شباهت برای چند برچسب از همان ۲۰۰ برچسب آورده شده است.

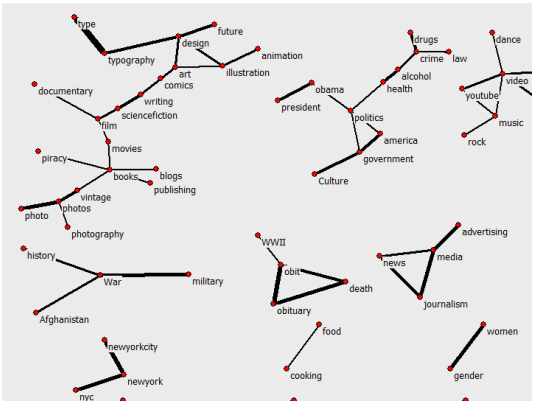
جدول (٤):چندنمونهمحاسبه‌درجهمشابهت

برچسب۱	برچسب۲			
movie	election	1	617	0.001
news	journalis	44	850	0.051
games	cars	2	877	0.002
children	books	19	879	0.021
parody	marketing	1	401	0.002
music	flash	59	3852	0.015
games	microsoft	7	990	0.007
photogra	photos	123	1613	0.076

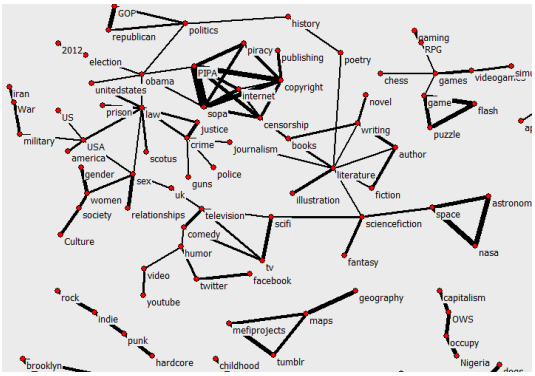
پس از تشکیل ماتریس مشابهت لازم است که عملیات خوشه‌بندی انجام گیرد. برای انجام عملیات خوشه‌بندی ما از پارتیشن‌بندی گراف استفاده می‌کنیم. برای هر دو شیوه تعیین حد آستانه τ ضروری است؛ لذا ما مشابه کار [۵] با آزمایش‌های مختلف حد آستانه‌های مناسب را پیدا کردیم. برای نشان دادن نتیجهٔ بصری این روش خوشه‌بندی، از نرم‌افزار Pajek 2.00 [۱۴] استفاده نموده‌ایم. از شکل (۳) تا شکل (۵) نمونه‌ای از این روش خوشه‌بندی را برای مجموعه داده مورد نظر می‌توان مشاهده نمود. ضخامت یال‌ها نشان‌دهنده درجه تشابه است.



شکل (۳):گرافتشکیل‌شدهبرای حداکثر ۲۰۰ برچسب دورهاولباحدآستانه=0.07



شکل (٤):گرافتشکیل‌شدهبرای حداکثر ۲۰۰ برچسبدوره ۲۶ (ماه اول سال ۲۰۱۰)باحدآستانه=0.07



شکل (۵):گرافتشکیل‌شدهبرای حداکثر ۲۰۰ برچسبدوره ۵۰ (تا ماه اول سال ۲۰۱۲)باحدآستانه=0.07

٤-٤- خوشه‌بندی مجدد داده‌ها در بازه‌های زمانی

همان طور که از شکل (۳) تا شکل (۵) مشخص است، ارتباط برچسب‌ها در طی زمان دستخوش تغییر می‌شوند؛

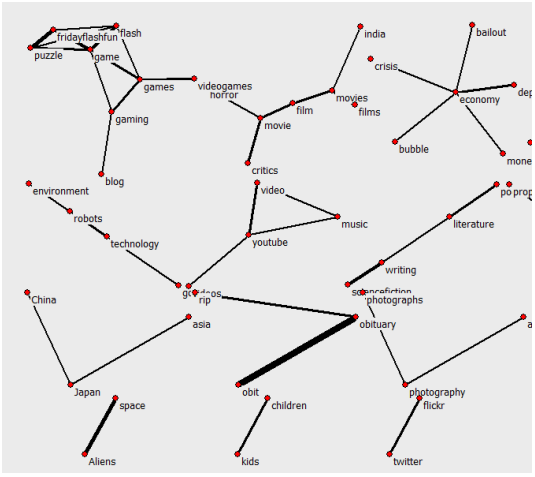
یعنی مثلاً اگر تا سال ۲۰۰۸ واژه war با Iraq و usa در ارتباط بوده، می‌بینیم که این ارتباط در شروع سال ۲۰۱۲ تغییر کرده و به جای Iraq، iran اضافه شده است. این به منزله این است که کاربران با توجه به مسائل روز لینک‌ها و مطالبی که در سایت‌ها قرار می‌دهند، به روز می‌شود و مطابق با همین وضعیت برچسب‌های آنها نیز شکل دیگری به خود می‌گیرد. همان طور که اشاره شد، هدف ما در این مقاله نیز به‌روز بودن خوشه‌بندی در طی زمان با برچسب‌های درج شده است.

برای نیل به هدف به‌روز نگه‌داشتن خوشه‌ها، همان طور که گفتیم، مجموعه داده به دو بخش تقسیم شده بود. ابتدا ما خوشه‌بندی را بر روی مجموعه دادهٔ دوره اول (از سال ۲۰۰۵ تا انتهای ۲۰۰۷) انجام می‌دهیم. سپس برای هر بازه (ما در اینجا ماهانه فرض گرفتیم؛ البته محدودیت در انتخاب بازه وجود ندارد) مطابق معادله‌های (۳) تا(۶) ماتریس مشابهت اصلی را به‌روزرسانی می‌کنیم.

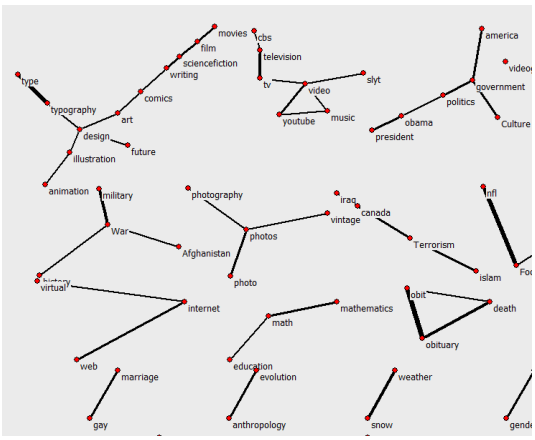
جدول (۵): مقادیر پارامترهای روش پیشنهادی

		K_2	K_1		
ماه	۰/۰۵	۲۰۰	۲۰۰	۲	۱

حال کافی است که از روی این ماتریس در هر بازه گراف را تشکیل داده تا پس از پارتیشن‌بندی آن خوشه‌بندی جدید حاصل شود. طبیعتاً انتظار ما از خوشه‌های به دست آمده این‌طور خواهد بود که ضمن حفظ ارتباطات مفاهیم گذشته، با مفاهیم روز نیز شدیداً در ارتباط باشد. ما برای خوشه‌بندی در اینجا نیز از Maximal Clique استفاده می‌کنیم.همچنین در جدول (۵) مقادیر پارامترهای روش پیشنهادی ذکر شده است.



شکل (٦):گراف جدید برای دوره ١٤ (ماه اول سال ۲۰۰۹)باحدآستانه>0.07 با (با حفظ ارتباط با دوره‌های گذشته و جدید)



شکل (۷):گراف جدید برای دوره ٢٦ (ماه اول سال ۲۰۱۰)باحدآستانه>0.07 (با حفظ ارتباط با دوره‌های گذشته و جدید)

برای مقایسه عملکرد،ما روش خود را Temporal Co-occurrence نامیده و با دو روش ذیل مقایسه می‌کنیم:

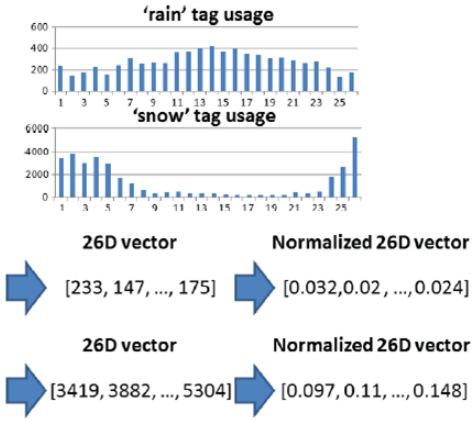
۱. روش Co-Occurrence در [3، ۵، ۶، ۷ و ۸]

۲. روش Temporal Vector در [6]

روش اول ما خوشه‌بندی را به این طریق انجام خواهیم داد که در هر دوره زمانی T، ۲۰۰ برچسب پرکاربرد را پیداکرده و برای آنها ماتریس مشابهت را تشکیل دهیم و پس از تشکیل گراف از روی ماتریس فوق و تعیین یک حد

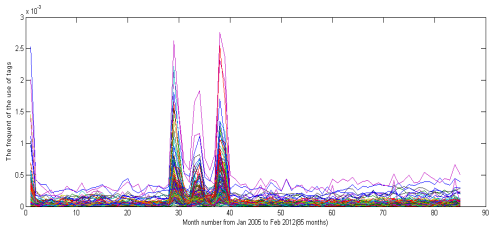
آستانه ثابت برای تمام زمان‌ها، پارتیشن‌بندی و در نتیجه خوشه‌بندی را انجام دهیم.

برای روش دوم برای هر دوره بردار ویژگی را برای ۲۰۰ برچسب پرکاربرد تا آن دوره را مطابق شکل (۸) تشکیل می‌دهیم. برای این کار تعداد تکرار هر برچسب را از اولین دوره تا آن دوره شمارش می‌کنیم و برای نرمال شدن، آن را بر مجموع تکرارهای کل برچسب‌های دوره تقسیم کرده و سپس برای خوشه‌بندی از روش *k-means* با *k=20* استفاده می‌کنیم.



شکل (۸): محاسبه بردار ویژگی زمانی برچسب‌ها[٦]

ما برای این بخش دوره‌های زمانی (بازه‌های زمانی) را ماهیانه و از اولین ماه سال ۲۰۰۵ تا اولین ماه سال ۲۰۱۲ در نظر گرفتیم که مجموعاً ۸۵ دوره شده است. در شکل (۹) توزیع ماهانه این برچسب‌ها قابل مشاهده است.



شکل (۹):توزیع ۲۰۰ برچسب از پرکاربردترین برچسب‌ها از ابتدای ۲۰۰۵ تا ابتدای ۲۰۱۲ به تفکیک ماه

پس از انجام خوشه‌بندی در سر دوره‌های زمانی مشخص‌شده با ۳ روش فوق ما، لازم است معیاری برای مقایسهٔ این روش‌ها اتخاذ کنیم.

۵-٤- تعیین معیار ارزیابی

برای محاسبه عددی ارزیابی روش‌های مورد مقایسه، ما از معیارهای Recall و Precision استفاده می‌کنیم. معیار Precision به ما می‌گوید که چه کسری از تعداد نمونه‌های به دست آمدهٔ[1] مورد نظر در خوشه، با یکدیگر مرتبط[2] هستند. (صحت) و Recall یعنی چه کسری از تعداد نمونه‌های مرتبط واقعی در خوشه قرار دارد. همچنین از معیار F-Measure برای نمایش ارتباط بین Recall و Precision [۵ و ۱۵] نیز استفاده خواهیم کرد که معادله آنها به طریق ذیل است:

که در آن R ، مجموعه برچسب‌های مرتبط و G؛ مجموعه برچسب‌های موجود در هر خوشه است.

لازم به ذکر است در معیارهای Recall و Precision نکته قابل تأمل نحوهٔ محاسبهٔ آنهاست. در [۶]برای تصمیم‌گیری برچسب‌های مرتبط از طریق طراحی یک سایت، از مردم نظرسنجی کرده‌اند، ولی به علت محدودیت این شیوه و یا عدم اطمینان از صحت درستی نظرات مردم، ما به دو روش آنها را برای هر خوشه محاسبه می‌کنیم. سپس در بازهٔ زمانی مورد نظر، میانگین معیارهای محاسبه‌شده‌خوشه‌های هر روش خوشه‌بندی را به دست می‌آوریم و با یکدیگر مقایسه می‌کنیم.

[1]- Retrieved instances
[2]- Relevant

$$\text{Precision} = \frac{|R \cap G|}{|R|} \qquad (۷)$$

$$\text{Recall} = \frac{|R \cap G|}{|G|} \qquad (۸)$$

$$\text{F} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (۹)$$

برای هر دو روش محاسبه Precision و Recall در هر خوشه داریم:

t: بازه‌ی زمانی (شماره ماهی که از ماه اول سال 2008 گذشته است) که $1 \le t \le 50$

Tag_i: هر برچسب موجود در خوشه،

G: تمامبرچسب‌هایخوشه به غیر از Tag_i

A: مجموعهٔ محاسبه‌شده ماتریس‌های مشابهت برای هر دورهٔ زمانی که درجهٔ مشابهت آنها از حد آستانهٔ ذکرشده بیشتر باشد و دارای ۴ فیلد است. برچسب اول، برچسب دوم، درجه تشابه، عدد بازه زمانی.

B: زیرمجموعه‌ای از Aکه دارای شرایط ذیل باشد:

- عدد بازهٔ زمانی آن ازtکمتر باشد.
- برچسب اول با Tag_i برابر باشد.

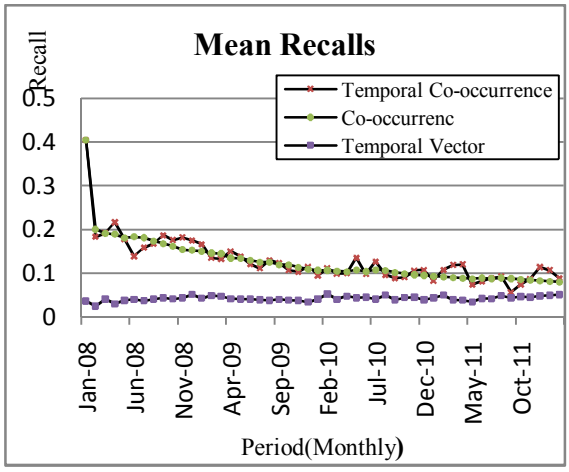
۱-۵-٤- روش اول محاســـــبهPrecision و Recall مستقل از زمان

در این روش در مجموعه Bبیشترین تکرار شماره بازه زمانی را پیدا می‌کنیم (peak در بردار زمانی برای این برچسب مطابق شکل (۹) -یعنی بیشترین برچسب‌هایی که با برچسب Tag_i مستقل از زمان شباهت زیادی داشتند- و آن راfمی‌نامیم. سپس تعریف می‌کنیم:

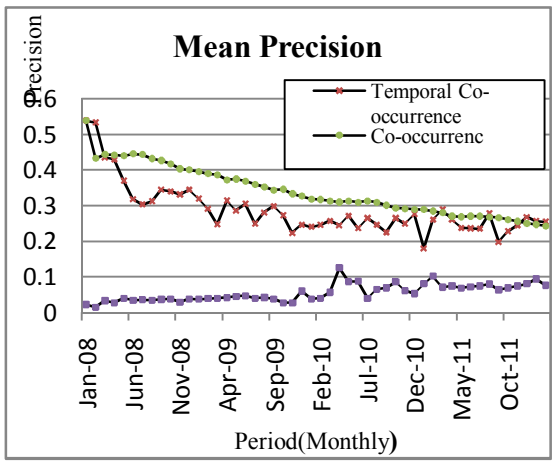
R: زیرمجموعه‌ای از برچسب‌هایBکه عدد بازه زمانی آن برابر با fباشد.

RelevantSize: اندازهٔ اشتراک مجموعه Rبا مجموعهٔ G

با توجه به این محاسبه، در شکل (۱۰ مقایسه میانگین Recall خوشه‌ها برای هر سه روش گفته‌شده و در شکل (۱۱ مقایسه میانگین Precision خوشه‌ها برای همان سه روش آورده شده است. در جدول (۶) نیز مقایسهٔ میانگین مقادیر Recall، Precisionو F-Measure برای تمام خوشه‌ها در تمامی بازه‌های زمانی جهت مقایسه روش پیشنهادی آورده شده است.



شکل (۱۰):مقایسه Mean Recalls خوشه‌ها در بازه‌های زمانی مختلف (روش محاسبه Recall مستقل از زمان)



شکل (۱۱):مقایسه Mean Precisionsخوشه‌ها در بازه‌های زمانی مختلف (روش محاسبه Precisionمستقل از زمان)

جدول (٦): مقایسهٔ معیارهای Recall، Precision و F-Measure(روش محاسبه Recallو Precisionمستقل از زمان)

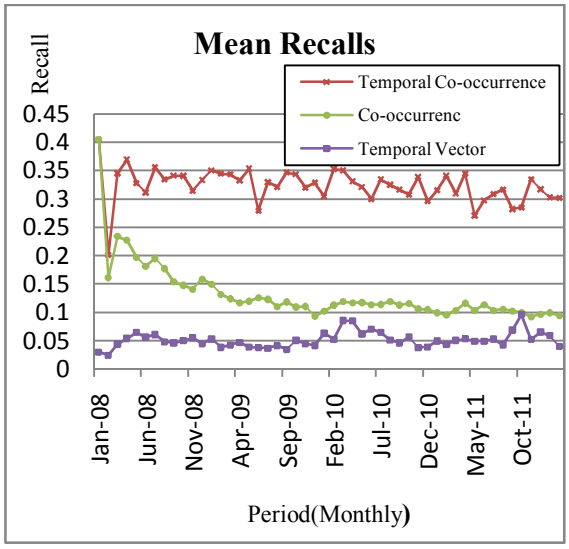
Method	Recall	Precision	F-Measure
Temporal Co-occurrence	0.122651	0.287737	0.17199
Co-occurrenc	0.121229	0.340581	0.178811
Temporal Vector	0.041395	0.055401	0.047385

۲-۵-٤- روش دوم محاسبه Precision و Recall وابسته به اخیرترین زمان

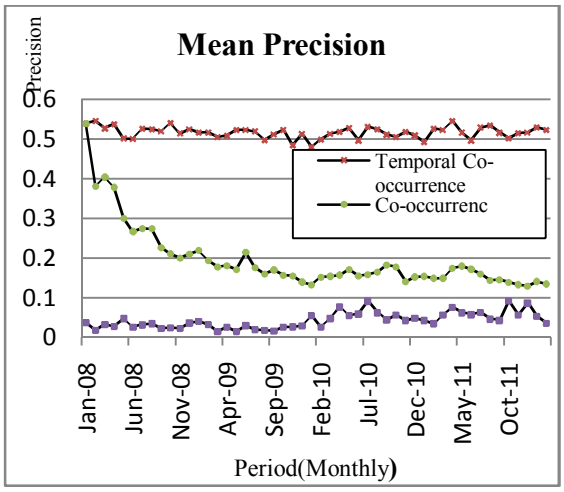
در این روش در مجموعه B به جایبیشترین تکرار شماره بازه زمانی، بزرگ‌ترین عدد بازه زمانی را پیدا می‌کنیم – یعنی اخیرترین برچسب‌هایی که با برچسب Tag_i شباهت زیادی داشتند– و آن را f می‌نامیم. سپس تعریف می‌کنیم:

R: زیرمجموعه‌ای از برچسب‌های B که عدد بازه زمانی آن برابر با f باشد.

$RelevantSize$: اندازهٔ اشتراک مجموعه R با مجموعه G مشابه نحوه محاسبه اول. در اینجا هم ضمن محاسبه معیارها با این روش، در شکل (۱۲) مقایسه میانگین Recall خوشه‌ها و در شکل (۱۳) مقایسه میانگین Precision خوشه‌ها قابل مشاهده است. در جدول (۷) نیز مقایسه میانگین مقادیر Recall، Precision و F-Measure برای تمام خوشه‌ها آورده شده است.



شکل (۱۲):مقایسه Mean Recallsخوشه‌هادربازه‌های زمانی مختلف (روش محاسبه Recallوابسته به زمان)



شکل (۱۳):مقایسه Mean Precisionsخوشه‌هادربازه‌های زمانی مختلف (روش محاسبه Precisionوابسته به اخیرترین زمان)

جدول (۷): مقایسه معیارهای Recall، Precision و F-Measure(روش محاسبه Recallو Precisionوابسته به زمان)

Method	Recall	Precision	F-Measure
Temporal Co-occurrence	0.323629	0.516722	0.397991
Co-occurrenc	0.131678	0.195546	0.157379
Temporal Vector	0.051034	0.042432	0.046337

به ازای هر دو روش محاسبه، شکل‌ها و جداول آورده شده نشان دهندهٔ مقایسه بین روش پیشنهادی با سایر روش‌هاست.

همان‌طور که در مقایسهٔ شکل (۱۰) با شکل (۱۲) و همچنین مقایسهٔ شکل (۱۱) با شکل (۱۳) مشخص است، می‌بینیم که در محاسبه معیارهای Precision و Recall به روش مستقل از زمان، به غیر از بازه‌هایی، زیاد شاهد بهبود نیستیم و علت آن هم بدین دلیل است که ما از دو ضریب α و β استفاده کردیم و به تدریج تأثیر گذشته را کمرنگ‌تر کردیم؛ حتی اگر در گذشته دور خیلی با هم ارتباط داشتند. البته اگر $\beta = 0$ و $\alpha = 1$ انتخاب می‌شد، نتیجه با روش Co- occurrenceبرابر می‌شد؛ بنابراین اگر چه ممکن است دو برچسب در یک بازه زمانی به اوج شباهت خود رسیده باشند، ولی در آینده تضمینی برای حفظ این ارتباط وجود ندارد. درحالی‌که در محاسبه معیارهای Precision و Recall به روش وابسته به زمان می‌بینیم که روش پیشنهادی نسبت به دو روش دیگر نتیجه بهتری داشته است.

نکته دیگری که می‌توان استنتاج کرد این است که در روش Co-occurrence در ابتدا نتایجی خوبی را از خود نشان می‌دهد، ولی به تدریج رو به کاهش خواهد بود؛ بنابراین اگر چندین دورهٔ تناوب زمانی دیگر روش‌ها را تکرار می‌کردیم مطمئناً در روش محاسبه معیارهای Precision و Recall به روش مستقل از زمان، روش پیشنهادی بهتر از Co-occurrence می‌شد؛ هر چند انتهای همین بازه‌های مشخص‌شده، گویای این ادعاست. در نهایت با توجه به ارزیابی‌هایانجام‌شده شاهد آن هستیم که نتیجه کار پیشنهادی به علت تطبیق‌پذیری با زمان نتایج مطلوب‌تری نسبت به سایر روش‌ها از خود به‌روز داده است.

۵- نتیجه‌گیری و پیشنهادها

در عصر حاضر با ظهور وب ۲.۰ما شاهد رشد چشمگیر فعالیت‌های کاربران در سطح وب هستیم. یکی از عمده‌ترین این فعالیت‌ها به اشتراک گذاشتن منابع با سایر کاربران است. در بعضی از خدمات وب ۲.۰ از فکسونومی برای رساندن مفهوم بهتر و خلاصه‌تر، یعنی امکان برچسب‌گذاری منابع توسط کاربران قرار داده شده است که در چند سال اخیر نیز استقبال خوبی از آن شده است؛ لذا با توجه به حجم وسیع منابع، این برچسب‌ها ملاک بسیار خوبی برای خوشه-بندی و دسته‌بندی جهت دسترسی کاربران به منابع مرتبط می‌باشد. از طرف دیگر ممکن است روزانه برچسب‌های جدیدی به بانک برچسب‌ها اضافه شود و یا حتی ارتباطات برچسب‌ها در طی زمان تغییر کنند؛ بنابراین نیاز به روش (هایی) است که خوشه‌بندی‌ها را با توجه به تغییرات زمانی به‌روز نگه دارد.

ما در این مقاله روش‌های مختلف خوشه‌بندی برای برچسب‌ها را در سیستم‌های برچسب‌گذاری مستقل از زمان و وابسته به زمان، شرح دادیم. سپس روش پیشنهادی خود که وابستگی برچسب‌ها نسبت به زمان را خوشه‌بندی می-کرد، شرح دادیم و در پایان ضمن مقایسه با سایر روش‌های مشابه شاهد بهبود کیفی و کمی بهتر خوشه‌های به وجود آمده بودیم و دیدیم که روش پیشنهادی،F-Measureمیانگین خوشه‌ها را در حالت وابسته به زمان نسبت به بهترین روش ۲۴٪ بهبود بخشید. در پایان، پیشنهاد ما برای کارهای آینده استفاده از ترکیب کلمات کلیدی استخراج‌شده از منابعی است که متنی هستند و وزن دهی مناسب به آنها و همچنین استفاده از آنتولوژی WordNet برای استخراج ریشه کلمات کلیدی استخراج‌شده و اضافه کردن آنها به مجموعه برچسب‌ها برای افزایش صحت خوشه‌ها می‌باشد. همچنین پیشنهاد استفاده از نتایج خوشه-

هـای بـه دسـت آمـده از ایـن روش را بـرای سیسـتم‌هـای پیشـنهاددهنده ودسـته‌بنـدی منـابعی کـه بـرای آنهـا برچسب‌گذاری انجام‌شده، داریم.

٦- مراجع

1- Svein Anfinnsen, Gheorghita Ghinea, Sergio de Cesare, "**Web 2.0 and folksonomies in a library context**"*International Journal of Information Management 31 (2011) 63–70*

2- O'Reilly, T. (2006). "**Web 2.0 compact definition: Trying again**". Available at:http://radar.oreilly.com/archives/2006/12/web 20 compact.html.

3- Golder, S., and Huberman, B. 2006. "**Usage patterns of collaborative tagging systems**". *Journal of Information Science* 32(2):198–208.

4- http://kmtoolbox.ir/?p=58

5-Nitin Agarwal, Magdiel Galan, H. L., and Subramanya, S. "WisColl: Collective wisdom based blog clustering". Information Sciences 180 (2010) 39–61, journal homepage: www.elsevier.com/locate/ins

6- Haipeng Zhang, Mohammed Korayem, Erkang You, David J. Crandall, "Beyond co-occurrence: discovering and visualizing tag relationships from geo-spatial and temporal similarities", WSDM '12 Proceedings of the fifth ACM international conference on Web search and data mining Pages 33-42-2012

7-Micah Dubinko,Ravi Kumar, Joseph Magnani, Jasmine Novak,Prabhakar Raghavan,Andrew Tomkins, "**Visualizing tags over time**"*WWW '06 Proceedings of the 15th international conference on World Wide Web.Pages 193 - 202, 2006*

8- Edwin Simpson, HP Laboratories, Bristol, "**Clustering Tags in Enterprise and Web Folksonomies**", *International Conference on Weblogs & Social Media, Seattle, March 31st, 2008*

9- Grigory Begelman, Philipp Keller, Frank Smadja, "**Automated Tag Clustering: Improving search and exploration in the tag space**"*ICWE'11 Proceedings of the 11th international conference on Web engineering*

10-Alberto Pérez García-Plaza a, Arkaitz Zubiaga, Víctor Fresno, Raquel Martínez, "**Reorganizing clouds: A study on tag clustering and evaluation**", *journal of Expert Systems with Applications 39 (2012) 9483–9493*

11-Brooks, C. H., and Montanez, N. "**Improved annotation of the blogosphere via autotagging and hierarchical clustering**".*Proceedings of the15th international conference on World Wide Web, New York, USA (2006),pp. 625-632.*

12-Jaiswal, Mayank Prakash, "**Clustering Blog Information**"*(2007).Master's Projects. Paper 36,http://scholarworks.sjsu.edu/etd_projects/36*

13- http://stuff.metafilter.com/infodump/

14-http://pajek.imfm.si/

15- http://en.wikipedia.org/wiki/Precision_and_recall