اولین کنفرانس ملی فناوری در مهندسی کامپیوتر و فناوری اطلاعات
1st National Innovation Conference on Computer Engineering and Information Technology
SHAFAGH UNIVERSITY 12/DECEMBER/2013
مؤسسه آموزش عالی غیردولتی غیرانتفاعی شفق تنکابن زمان: ۲۹ آذر ۱۳۹۲

تحلیل محتوا، رویکردی نوین در بهبود کارایی تشخیص اجتماع

علی ریحانیان[۱]، حسین علیزاده[۲]، بهروز مینایی[۳]

[۱] دانشکده فناوری اطلاعات، دانشگاه علوم و فنون مازندران، بابل
areihanian@ustmb.ac.ir

[۲] دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران
halizadeh@iust.ac.ir

[۳] دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران
b_minaei@iust.ac.ir

چکیده

یکی از مباحث مهم در زمینه‌ی تحلیل شبکه‌های پیچیده، مبحث تشخیص اجتماع می‌باشد. اکثر روش‌هایی که در زمینه‌ی تشخیص اجتماع پیشنهاد شده‌اند، این عمل را تنها با درنظر گرفتن ساختار گرافی شبکه انجام می‌دهند. اما در سال‌های اخیر، تلاشهایی در زمینه‌ی تحلیل محتوا و استفاده از نتایج آن به منظور بهبود کارایی تشخیص اجتماع صورت گرفته است. در این مقاله برآنیم تا با معرفی برخی از این رویکردهای جدید و پیاده سازی روش پیشنهادیشان، به بررسی نتایج حاصل از بکارگیری تحلیل محتوا در عمل تشخیص اجتماع بپردازیم. نتایج حاصله نشان می‌دهند که بکارگیری محتوا به صورت ملموسی به بهبود کارایی عمل تشخیص اجتماع کمک می‌کند.

کلمات کلیدی

شبکه پیچیده، اجتماع، تشخیص اجتماع، تحلیل محتوا

۱- مقدمه

شبکه‌های پیچیده[۱]، نامی است که به یک حوزه‌ی علمی و تحقیقاتی بین رشته‌ای اطلاق می‌شود که این حوزه، مفاهیم مربوط به رشته‌هایی نظیر فیزیک، ریاضیات و آمار، علوم کامپیوتر و علوم اجتماعی را گرد هم می‌آورد. یک شبکه‌ی پیچیده، نمایانگر تعاملات بین موجودیت ها[۲] یا گروهی از افراد می‌باشد. به عنوان نمونه‌هایی از شبکه‌های پیچیده می‌توان به شبکه اینترنت، شبکه جهانی وب[۳]، شبکه‌های اجتماعی[۴]، شبکه تراکنش بر خط[۵]، شبکه تماس‌های تلفنی[۶]، شبکه تعامل پروتئین-پروتئین[۷] و شبکه ارجاعات[۸] اشاره کرد.

یکی از مباحث مهم در زمینه‌ی تحلیل شبکه‌های پیچیده، بحث تشخیص اجتماع می‌باشد. نقش تشخیص اجتماع، جستجو برای پیدا کردن اجتماعات است [۱].

پژوهش‌گران مختلف، الگوریتم‌های گوناگونی را به منظور تشخیص اجتماع پیشنهاد داده‌اند که این الگوریتم‌ها، تنها با در نظر گرفتن ساختار گرافی شبکه به عمل تشخیص اجتماع می‌پردازند. در سال‌های اخیر، تلاش‌هایی در زمینه‌ی تحلیل محتوای[۹] یک شبکه و استفاده از نتایج آن، در بهبود عمل تشخیص اجتماع انجام گرفته است. به عنوان نمونه‌ای از این تلاش‌ها، می‌توان به رویکردی که تحت عنوان تشخیص اجتماع مبتنی بر موضوع[۱۰] مطرح شده است، اشاره کرد. در این رویکرد، ابتدا خوشه‌های موضوعی[۱۱] در شبکه یافت می‌شوند و سپس، یکی از الگوریتم‌های تشخیص اجتماع بر روی این خوشه‌های موضوعی اعمال می‌شود. به عنوان نمونه‌ی دیگری از این تلاشها، می‌توان به رویکردی که تحت عنوان تشخیص اجتماع بر روی شبکه‌ی معنایی[۱۲] مطرح شده است، اشاره کرد. در این رویکرد، ابتدا محتوای رد و بدل شده بین عناصر موجود در شبکه، تحلیل و بررسی می‌شوند و بر اساس نتایج این تحلیل، یک شبکه‌ی معنایی ترسیم می‌شود و عمل تشخیص اجتماع بر روی این شبکه‌ی معنایی انجام می‌پذیرد. در این مقاله، این

دو رویکرد به طور مبسوط بررسی شده و روش‌های پیشنهادیشان پیاده‌سازی می‌شوند.

ساختار مقاله بدین صورت می‌باشد که در بخش بعدی، به بررسی برخی از مفاهیم اولیه در شبکه‌های پیچیده پرداخته می‌شود. در بخش سوم، دو روشی که با دو رویکرد متفاوت به تحلیل محتوا به منظور بهبود کارایی تشخیص اجتماع پرداختند، بررسی می‌شوند. در بخش چهارم، روش‌های مطرح شده در بخش قبلی پیاده‌سازی شده و نتایج حاصل از بکارگیری این دو روش تحلیل می‌شوند. همچنین در این بخش اثبات می‌شود که استفاده از نتایج تحلیل محتوا، موجب بهبود کارایی عمل تشخیص اجتماع می‌گردد. در بخش آخر نیز به بیان نتایج حاصل از این پژوهش و کارهای آینده پرداخته می‌شود.

۲- تعریف مفاهیم اولیه مورد نیاز

در این بخش، به بررسی مفاهیم اولیه مورد نیاز جهت مطالعه‌ی شبکه های پیچیده پرداخته می‌شود[۲]:

۲-۱- تعریف ریاضی شبکه

به عنوان یک نمایش ریاضی از یک شبکه، می‌توان آن را به صورت یک گراف $G=\{V,E\}$ نمایش داد. این گراف، شامل مجموعه‌ای از گره‌ها به صورت $V=\{v_1,v_2,...,v_n\}$ و مجموعه‌ای از یال‌ها به صورت $E=\{e_1,e_2,...,e_m\}$ می‌باشد. در این گراف، هر یال e_j بصورت $e_j=\{v_{j1},v_{j2}\}$ تعریف می‌شود که در آن $v_{j1}\neq v_{j2}$ و $v_{j1},v_{j2}\in V$ می‌باشد[۳].

۲-۲- نحوه نمایش شبکه

می‌توان شبکه‌ها را به ۲ صورت زیر نمایش داد[۴و۱]:

۱. **به صورت گراف:** در این نوع نمایش از شبکه، ساختارهای گرافی، وظیفه متصل کردن گره‌ها و یال‌ها به یکدیگر را بر عهده دارند.

۲. **به صورت ماتریس:** برای نمایش گراف شبکه، می‌توان از ماتریس نیز استفاده کرد. به این نوع ماتریس، ماتریس مجاورتی[۱۳] گراف گفته می‌شود. ماتریس مجاورتی، یک ماتریس مربعی می‌باشد. ابعاد این ماتریس، برابر تعداد گره-های گراف متناظر با آن بوده و در صورت وجود یالی بین دو گره، در درایه‌ی متناظر با دو گره در ماتریس مجاورتی، مقدار ۱ لحاظ شده و در غیر اینصورت در این درایه، مقدار ۰ لحاظ می‌شود.

۲-۳- تعریف اجتماع

در مقالات و تحقیقات علمی مختلف، تعاریف متفاوتی از اجتماع ارائه شده است که در ادامه به ذکر برخی از این تعاریف پرداخته می‌شود:

یک اجتماع شبکه که گاهی از اوقات به آن پیمانه[۱۴] یا خوشه[۱۵] گفته می‌شود، به عنوان یک گروه از گره‌ها در نظر گرفته می‌شود که تعاملات بهتر و بیشتری بین اعضای این گروه نسبت به اعضای این گروه و دیگر اعضای شبکه وجود دارد[۵].

تعریف دیگری بیان می‌کند که اجتماع، یک زیرشبکه‌ی[۱۶] متراکم[۱۷] در داخل یک شبکه‌ی بزرگتر می‌باشد، مثل یک گروه از دوستان صمیمی در یک شبکه اجتماعی یا گروهی از صفحات وب به هم پیوند داده شده در شبکه جهانی وب[۶].

در [۷] آمده است که در یک اجتماع، چگالی[۱۸] یال‌های داخل آن بیشتر از چگالی یال‌های بین آن با سایر گره‌های شبکه است.

۲-۴- تشخیص اجتماع

با فراگیر شدن شبکه‌های اجتماعی مثل facebook، تحلیل داده‌های چنین شبکه‌هایی به یک موضوع تحقیقاتی با اهمیت تبدیل شده است[۸]. ساختار اجتماع[۱۹] چنین شبکه‌های پیچیده‌ای، هم سازمان-دهی این شبکه‌ها و هم ارتباطات پنهان بین اجزای تشکیل دهنده‌ی آن‌ها را آشکار می‌سازد[۹]. بنابراین تشخیص اجتماع، به عنوان یک موضوع مهم در تحلیل شبکه‌های اجتماعی مطرح می‌شود. نقش تشخیص اجتماع، جستجو برای پیدا کردن اجتماعات است[۱]. احتمال این‌که افرادی که در این شبکه‌ها به اجتماعات مشترکی متعلقند، دارای سرگرمی‌های مشترک، عملکردهای اجتماعی مشترک، دیدگاه-های مشترک و ... باشند، بیشتر وجود دارد[۸]. بنابراین، اجتماعات شناسایی شده در چنین شبکه‌هایی، می‌توانند در پیشنهاد مشارکتی[۲۰]، انتشار اطلاعات[۲۱]، به اشتراک‌گذاری دانش[۲۲] و دیگر کاربردها مورد استفاده قرار گیرند[۲].

۲-۵- تابع پودمانی[۲۳]

مشهورترین معیار ارزیابی اجتماعات تشخیص داده شده توسط یک الگوریتم تشخیص اجتماع، پودمانی می‌باشد. پودمانی که با Q نمایش داده می‌شود، اولین بار توسط خانم گیروان[۲۴] و آقای نیومن[۲۵] مورد استفاده قرار گرفت. در حال حاضر، بسیاری از محققین، از پودمانی برای ارزیابی روش‌های تشخیص اجتماعی که در تحقیقاتشان بکار می-گیرند، استفاده می‌کنند. پس از این‌که تمامی اجتماعات یک شبکه شناسایی شدند، پودمانی به آن شبکه اعمال می‌شود. ورودی تابع پودمانی، شبکه و تمامی اجتماعات آن بوده و خروجی آن، یک عدد حقیقی بین ۱- و ۱ می‌باشد. مقادیر پودمانی نزدیک به ۱ برای یک

روش تشخیص اجتماع، بیانگر مناسب بودن آن روش می‌باشد. پودمانی به صورت زیر تعریف می‌شود[۱۰]:

$$Q = \sum_{i} (e_{ii} - a_i^2) \qquad (۱)$$

در رابطه‌ی بالا، i بیانگر اجتماعات تشخیص داده شده می‌باشد و e_{ii}، نسبت تعداد یال‌هایی که گره‌های داخل اجتماع را به هم متصل می‌کنند به کل یال‌های گراف را مشخص می‌کند. به منظور توضیح بیشتر در مورد a_i، بهتر است که ابتدا به رابطه‌ی زیر نگاهی انداخته شود[۱۱]:

$$e_{ij} = \begin{cases} \frac{1}{2}\left(\frac{a}{c}\right) & i \neq j \\ \left(\frac{b}{c}\right) & i = j \end{cases} \qquad (۲)$$

در رابطه‌ی بالا، i و j اندیس اجتماعات می‌باشند، a برابر تعداد یال‌هایی می‌باشد که اجتماع i را به اجتماع j متصل می‌کنند، b برابر تعداد یال‌های درون اجتماع i می‌باشد و c برابر تعداد کل یال‌ها می‌باشد. با در نظر گرفتن مجموع ۲ حالت بیان شده در رابطه‌ی بالا، به تعریف a_i می‌رسیم. a_i، بیانگر نسبت تعداد یال‌هایی که حداقل یک گره آن در اجتماع i است، به کل یال‌های گراف می‌باشد. اگر تمامی رئوس در یک اجتماع قرار گرفته باشند (یعنی کل گراف شامل تنها یک اجتماع باشد) و یا گره‌ها بصورت تصادفی در بین اجتماعات قرار گرفته باشند، مقدار Q برابر ۰ خواهد بود. اگر روش تشخیص اجتماع، هر راس گراف را در یک اجتماع مجزا قرار دهد، در صورتی که شبکه واقعا چنین ساختاری را نداشته باشد، آنگاه مقدار Q نزدیک به ۱- خواهد بود.

روش‌های تشخیص اجتماع قدرتمند، دارای مقادیر پودمانی مثبت نزدیک به ۱ خواهند بود. اما باید به این نکته اشاره کرد که مقدار پودمانی مناسب برای شبکه‌های مختلف، متفاوت می‌باشد. برای مثال، شاید مقدار پودمانی برای یک شبکه، هیچگاه بیشتر از ۰.۴ نشود. بنابراین، نمی‌توان گفت که اگر پودمانی مقدار ۰.۴ را دارد، تشخیص اجتماع به درستی انجام نشده است. ممکن است که ۰.۴، بیشترین مقداری باشد که پودمانی بتواند در آن شبکه به آن برسد. در سال‌های اخیر، روش‌های جدیدی ارائه شده‌اند که سعیشان بر بیشینه‌سازی مقدار پودمانی[۲۶] است.

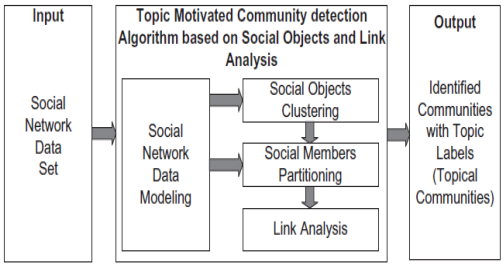
۳- تشخیص اجتماع به کمک تحلیل محتوا

همانطور که در بخش‌های قبلی بیان شد، اکثر الگوریتم‌هایی که در مقالات و پژوهش‌های علمی مختلف به منظور تشخیص اجتماع پیشنهاد شده‌اند، تنها با در نظر گرفتن ساختار گرافی شبکه به عمل تشخیص اجتماع می‌پردازند. برخی از این پژوهش‌ها نیز به بررسی و تحلیل محتوا پس از اعمال الگوریتم‌های تشخیص اجتماع بر روی شبکه و شناسایی اجتماعات، میپردازند. بدین صورت که، با درنظر گرفتن تک تک گره‌های موجود در برخی از اجتماعات شناسایی شده و تخلیل محتوای مربوط به این گره‌ها، بررسی می‌کنند که آیا این گره‌ها از لحاظ محتوایی نیز با یکدیگر شباهت دارند یا خیر.

اما در سال‌های اخیر، پژوهش‌هایی در زمینه‌ی بکارگیری نتایج حاصل از تحلیل محتوای مربوط به هر یک از گره‌های شبکه، به منظور بهبود عمل تشخیص اجتماع انجام شده است. در ادامه‌ی این بخش، به دو نمونه از مهم‌ترین تلاش‌ها در این زمینه پرداخته می‌شود.

۳-۱- تشخیص اجتماع مبتنی بر موضوع

در مقاله‌ای که در سال ۲۰۱۲ توسط ژائو و همکاران نوشته شد[۸]، یک روش تشخیص اجتماع مبتنی بر موضوع پیشنهاد شده است. این روش، می‌تواند اجتماعات موضوعی را که به طور همزمان منعکس-کننده‌ی موضوعات و قدرت اتصالات می‌باشند، شناسایی کند. یعنی اجتماعات تشخیص داده شده توسط روش پیشنهادی این مقاله، دارای گره‌هایی خواهد بود که این گره‌ها هم دارای موضوعات یکسان و هم دارای ارتباطات تنگاتنگ با یکدیگر خواهند بود. چارچوب روش ارائه شده در این مقاله، در شکل زیر نشان داده شده است:



شکل (۱): چارچوب روش تشخیص اجتماع مبتنی موضوع ارائه شده توسط ژائو و همکاران[۱]

کل فرایند شناسایی انجمن در این روش، توسط ۴ ماژول کلیدی انجام می‌شود. این ماژول‌ها عبارتند از: مدل‌سازی داده‌های شبکه اجتماعی[۲۷]، خوشه‌بندی اشیاء اجتماعی[۲۸]، افرازبندی اعضای اجتماعی[۲۹]، اجتماعی[۲۹]، تحلیل پیوند[۳۰]. در ادامه، به بررسی ماژول‌های ذکر شده پرداخته می‌شود.

ماژول اول، ماژول مدل‌سازی داده‌های شبکه اجتماعی می‌باشد. هدف این ماژول، ساختاربندی مجموعه داده به مدل‌های فرمال[۳۱] برای پردازش می‌باشد. این مقاله، با در نظر گرفتن اشیاء اجتماعی[۳۲]، یک مدل گرافی فرمال برای توصیف شبکه اجتماعی پیشنهاد می‌دهد.

ماژول‌های خوشه‌بندی اشیاء اجتماعی و افرازبندی اعضای اجتماعی دومین و سومین ماژول می‌باشند. به طور کلی، مردم فعالیت-های اجتماعی را بر روی اشیاء اجتماعی مثل ایمیل‌ها و وبلاگ‌ها انجام می‌دهند. این اشیاء، اغلب نمایانگر عناوینی که مردم به آنها علاقه‌مند هستند، می‌باشند. در این مقاله، تنها، اشیاء اجتماعی متنی[۳۳] در نظر گرفته شده‌اند.

چهارمین ماژول، ماژول تشخیص اجتماع موضوعی (یا تحلیل پیوند) می‌باشد. این ماژول به روی هر خوشه‌ی موضوعی اعمال می-شود. اعضای هر کدام از خوشه‌های موضوعی، اغلب به یکدیگر با شدت‌های[۳۴] مختلفی متصلند. برای شناسایی اعضایی که بسیار تنگاتنگ به هم متصلند[۳۵]، بایستی تحلیل اتصالات[۳۶] روی هر خوشه‌ی موضوعی انجام شود. در این مقاله به این فرایند، شناسایی انجمن موضوعی یا تجزیه و تحلیل اتصالات گفته می‌شود. در این فرایند، روش‌های تشخیص اجتماعات زیادی می‌توانند بکار گرفته شوند. در این مقاله، برای تجزیه و تحلیل اتصالات، از روش‌های بر پایه‌ی بیشینه‌سازی پودمانی[۳۷] استفاده شده است.

با توجه به موارد ذکر شده، روش ارائه شده توسط این مقاله، در ۳ مرحله انجام می‌شود: خوشه‌بندی اشیاء اجتماعی[۳۸]، افرازبندی کاربر بر مبنای موضوع[۳۹] و تشخیص اجتماع بر پایه‌ی اتصال[۴۰].

از آنجایی که این مقاله، مباحث موضوع و پیوند را با یکدیگر ترکیب کرده و به عمل تشخیص اجتماع می‌پردازد، لذا به منظور ارزیابی نتایج حاصل از چارچوب پیشنهادیش، به معرفی یک معیار جدید می‌پردازد. بدیهی است که چنین معیاری بایستی به بررسی اجتماعات تشخیص داده شده، از دو جنبه‌ی موضوع و پیوند بپردازد. این معیار که با نام $PurQ_\beta$ معرفی شده است، به صورت زیر تعریف می‌شود:

$$PurQ_\beta = (1+\beta^2).(Purity.Q)/(\beta^2.Purity+Q) \quad (۳)$$

در رابطه‌ی بالا:

- $Purity$ میزان خلوص موضوعات، در اجتماعات تشخیص داده شده را مشخص می‌کند و از رابطه‌ی زیر محاسبه می-شود:

$$Purity = \frac{1}{N_{cm}}.\sum_{i=1}^{i=N_{cm}} \max_{1\le j\le k}\{\frac{n_{ij}}{n_i}\} \quad (۴)$$

در رابطه‌ی بالا، N_{cm} نشان‌دهنده‌ی تعداد اجتماعات حاصله می‌باشد. همچنین n_{ij} بیانگر تعداد گره‌هایی که به موضوع j و اجتماع i تعلق دارند، می‌باشد. n_i نیز به تعداد اعضای اجتماع i اشاره دارد. هرچقدر مقدار $Purity$ بالاتر باشد نشان‌دهنده‌ی این امر خواهد بود که اجتماعات، از جنبه‌ی موضوع بهتر افراز بندی شده‌اند.

- Q نیز نمایانگر پودمانی می‌باشد که در مورد آن در بخش قبل به طور مفصل توضیح داده شد.
- همچنین در رابطه‌ی بالا، β پارامتری برای تنظیم وزن $Purity$ و Q می باشد: $\beta \in [0,\infty]$

اگر $\beta = 1$ باشد، می تواند اینگونه در نظر گرفته شود که میزان خلوص موضوعات و ساختار شبکه اجتماعی به یک اندازه اهمیت دارند. اگر $1 < \beta < \infty$ باشد، این معنی را می دهد که رابطه، در مقایسه با $Purity$ توجه بیشتری به Q کرده است. اگر $0 < \beta < 1$ باشد، این بدان معنی است که رابطه، تاکید بیشتری بر روی $Purity$ دارد.

۳-۲- تشخیص اجتماع بر روی شبکه‌ی معنایی

در صورت وجود اطلاعات متنی مرتبط با گره‌های شبکه، می‌توان پیش از اعمال روش‌های تشخیص اجتماع، شبکه را با درنظر گرفتن این اطلاعات ترسیم کرد. در این حالت، یک شبکه‌ی معنایی با استفاده از اطلاعات معنایی حاصل خواهد شد. اعمال روش‌های تشخیص اجتماع بر روی چنین شبکه‌ی معنایی، منجر به یافتن اجتماعات پرمعناتری خواهد شد.

در مقاله‌ای که در سال ۲۰۱۲ توسط ژیا و بو منتشر شد[۱۲]، با رویکردی متفاوت نسبت به ژائو و همکاران[۸]، به بحث تحلیل محتوا و به‌کارگیری نتایج حاصل از آن در تشخیص اجتماع پرداخته شد. همانطور که در بخش قبل گفته شد، ژائو و همکاران، از تحلیل محتوا به منظور یافتن خوشه‌های موضوعی سود بردند و سپس با اعمال الگوریتم تشخیص اجتماع بر روی این خوشه‌های موضوعی، به یافتن اجتماعات پرداختند.

اما ژیا و بو، از تحلیل محتوا به منظور ساخت شبکه‌ای که محتوا به صورت وزن‌دهی به یال‌ها در آن اثرگذار خواهد بود، استفاده کردند. بدین‌منظور آنها در یک انجمن مربوط به کشور چین، با بررسی برخی نظرات رد و بدل شده بین کاربران مختلف، مجموعه‌ای از کلمات کلیدی ساختند و به هر یک از این کلمات مقداری بین ۰ و ۱ اختصاص دادند. آن‌ها این کلمات را به دو دسته‌ی پشتیبان[۴۱] و مخالف[۴۲] تقسیم‌بندی کردند، سپس به تحلیل کلیه‌ی نظرات رد و بدل شده بین کاربران پرداختند و با توجه به ارزش کل کلمات کلیدی به‌کار رفته شده در این نظرات (اگر در یک نظری تعداد زیادی کلمه‌ی کلیدی به‌کار رود، میانگین ارزش کل این کلمات محاسبه شده و به یال ترسیم شده بین دو کاربر به عنوان ارزش اعتماد[۴۳] اختصاص داده می‌شود)، اگر ارزش اعتماد یال بین دو کاربری از یک حد آستانه‌ای کمتر باشد، آن یال حذف خواهد شد.

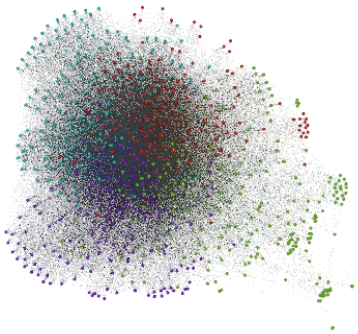
بدین صورت، ژیا و بو به ترسیم یک شبکه‌ی معنایی پرداختند و سپس با استفاده از یکی از الگوریتم‌های تشخیص اجتماع، اجتماعات موجود در این شبکه‌ی معنایی را یافتند. در این مقاله، آن‌ها به طور مستقیم از نتایج حاصل از تحلیل محتوا در اصلاح شبکه‌ی اولیه که تنها با ردوبدل شدن یک نظر بین دو کاربر و بدون توجه به محتوای آن نظر، به رسم یال بین آن دو کاربر می‌پرداخت، بهره جستند.

۴- پیاده‌سازی و تحلیل نتایج

در این بخش، چارچوب‌های ارائه شده در روش‌های تشخیص اجتماع مبتنی بر موضوع و تشخیص اجتماع بر روی شبکه‌ی معنایی که در بخش قبلی به طور کامل توضیح داده شده‌اند، بر روی یکی از مجموعه داده‌های MovieLens اعمال می‌شوند. این مجموعه داده، شامل ۱۰۰۰۰۰ امتیازدهی[۴۴] اعمال شده به ۱۶۸۲ فیلم توسط ۹۴۳ کاربر می‌باشد. تمامی کاربران به حداقل ۲۰ فیلم امتیاز داده‌اند. این امتیازها نیز از ۱ تا ۵ می‌باشند.

به منظور پیاده‌سازی چارچوب‌های ذکرشده، کلیه‌ی فیلم‌هایی که در موضوع مستند[۴۵] یا موضوع وسترن[۴۶] یا در تلفیقی از هر دو موضوع ساخته شده بودند، انتخاب شدند. در واقع در این مجموعه داده، فیلم‌ها به عنوان اشیاء اجتماعی در نظر گرفته شدند. در کل این مجموعه داده، ۷۷ فیلم در یکی از این دو موضوع (وسترن و مستند) یا بصورت تلفیقی از هر دو موضوع، ساخته شده‌اند(برای مثال، امکان دارد فیلم مستندی به سبک وسترن ساخته شده باشد. در این حالت، این فیلم هر دو موضوع را شامل می‌شود). سپس کلیه‌ی امتیازهایی که توسط کاربران به این فیلم‌ها اختصاص داده شده است، بازیابی شدند. لازم به ذکر است که احتمال دارد کاربری به بیش از یک فیلم امتیاز داده باشد.

در مرحله‌ی اول، شبکه‌ای رسم می‌شود که در آن، بین تمامی کاربرانی که به فیلم‌های مشترکی به امتیاز دادند، یال وجود خواهد داشت. وزن این یال‌ها برابر تعداد فیلم‌های مشترکی خواهد بود که دو گره (کاربر) مربوطه به امتیازدهی به آن‌ها پرداختند. در این مرحله، اقدام به اعمال یکی از الگوریتم‌های تشخیص اجتماع بر روی این شبکه می‌شود. مقدار پودمانی محاسبه شده در این حالت، برابر ۰.۱۰۸۶ خواهد شد. در این مرحله هیچ تحلیل محتوایی انجام نگرفته است. شکل زیر نمایی از شبکه‌ی ایجاد شده در این مرحله (که ما به آن شبکه‌ی اصلی می‌گوییم) را نشان می‌دهد:



شکل (۲): نمایی از شبکه‌ی اصلی

در شکل بالا، گره‌ها با رنگ‌های مشابه در یک اجتماع قرار دارند.

در مرحله‌ی بعد، اقدام به اعمال تشخیص اجتماع مبتنی بر عنوان بر روی شبکه‌ی اصلی ایجاد شده در بخش اول می‌شود. در این مرحله، کلیه‌ی گره‌های شبکه که به فیلم‌هایی که در موضوع مستند بودند امتیازدهی کردند، جدا شده و اولین خوشه‌ی موضوعی تشکیل می‌گردد. به همین ترتیب، کلیه‌ی گره‌های شبکه که به فیلم‌هایی که در موضوع وسترن بودند امتیاز دادند، جدا شده و دومین خوشه‌ی موضوعی تشکیل می‌گردد. از آنجایی که امکان دارد گره یا گره‌هایی به ۲ فیلم با موضوعات متفاوت امتیاز داده باشند، بنابراین، این احتمال وجود دارد که گره یا گره‌هایی در هر دو خوشه حضور داشته باشند. با اعمال الگوریتم تشخیص اجتماع بر روی هر یک از این خوشه‌های موضوعی، به مقدار پودمانی ۰.۲۶۸۶ برای خوشه‌ی مستند و ۰.۱۰۳ برای خوشه‌ی وسترن می‌رسیم. پودمانی کل در حالتی که گره‌ها را به خوشه‌های موضوعی تقسیم کردیم، از رابطه‌ی زیر محاسبه می‌شود:

$$\frac{weight_{k1}}{Weight}.Q_{k1}+\frac{weight_{k2}}{Weight}.Q_{k2}+...+\frac{weight_{kn}}{Weight}.Q_{kn} \quad (۵)$$

که در رابطه‌ی بالا، $Weight$ برابر مقدار کل وزن یال‌ها در شبکه‌ی اصلی (بدون جداسازی خوشه‌ها) می‌باشد. همچنین $Weight_{k1}$ برابر وزن مربوط به کلیه‌ی یال‌های موجود در خوشه‌ی موضوعی اول، $Weight_{k2}$ برابر وزن مربوط به کلیه‌ی یال‌های موجود در خوشه‌ی موضوعی دوم و ... می‌باشد. Q_{k1} برابر مقدار پودمانی مربوط به خوشه‌ی موضوعی اول، Q_{k2} برابر مقدار پودمانی مربوط به خوشه‌ی موضوعی دوم و .... می‌باشد. بنابراین، طبق رابطه‌ی فوق، مقدار پودمانی کل در مرحله‌ی دوم برابر ۰.۱۲۴۴۳۷ خواهد بود.

در مرحله‌ی سوم، به منظور تشخیص اجتماع بر روی شبکه‌ی معنایی، یک روش وزن‌دهی جدید به یال‌ها، معرفی شده و اقدام به ساخت یک شبکه‌ی معنایی می‌شود. بدین منظور، با بررسی تک‌تک

امتیازات، به هر یک از یال‌های بین گره‌ها در شبکه‌ی اصلی ایجاد شده در مرحله‌ی اول، وزن جدیدی اختصاص داده می‌شود. در این مرحله، ابتدا با بهره‌گیری از ایده‌ی مطرح شده در مباحث مربوط به تحلیل تمایل[۴۷]، اگر کاربران به فیلمی نظر بالاتر از ۳ (۴ یا ۵) داده باشند، نظر آنها نسبت به فیلم مذکور، مثبت[۴۸] در نظر گرفته می‌شود. همچنین، اگر کاربران به فیلمی امتیاز کمتر از ۳ (۱ یا ۲) داده باشند، نظر آنها نسبت به فیلم مذکور، منفی[۴۹] در نظر گرفته می‌شود و در نهایت، اگر به فیلمی امتیاز برابر ۳ داده شود، نظر کاربر امتیاز دهنده نسبت به فیلم مذکور، بی‌طرفانه[۵۰] درنظر گرفته می‌شود. با توجه به مطالب بالا، اگر نظر کاربری نسبت به فیلمی مثبت باشد، به آن عدد ۱، اگر منفی باشد به آن عدد -۱ و اگر بی‌طرفانه باشد به آن عدد ۰ اختصاص می‌دهیم. برای مثال، اگر کاربری به فیلمی امتیاز ۴ داده باشد، ازآنجایی‌که نظر او نسبت به این فیلم مثبت ارزیابی می‌شود لذا به نظر او عدد ۱ اختصاص داده می‌شود. سپس، با توجه به رابطه‌ی زیر، به ارتباطات موجود در شبکه‌ی اصلی وزن اختصاص داده می‌شود:

$$1-|Value1-Value2| \qquad (۶)$$

در رابطه‌ی بالا، $Value1$ مربوط به عدد اختصاص داده شده به کاربر اول با توجه به مثبت یا منفی یا بی‌طرفانه بودن نظرش می‌باشد و $Value2$ نیز مربوط به عدد اختصاص داده شده به کاربر اول با توجه به مثبت یا منفی یا بی‌طرفانه بودن نظرش می‌باشد. در نهایت، برای هر ارتباط بین دو کاربر، اعداد ۰ یا ۱ یا ۱- اختصاص داده می‌شود. اگر دو کاربر نسبت به یک فیلم هم‌نظر بودند (هر دو نظر مثبت یا منفی یا بی‌طرفانه داشته باشند) عدد ۱ به ارتباطشان اختصاص داده شده و اگر نظراتشان در یک سطح تفاوت از یکدیگر قرار داشته باشد (مثلا یکی منفی و دیگری بی‌طرف یا یکی مثبت و دیگری بی‌طرف باشد) عدد ۰ به ارتباطشان اختصاص داده می‌شود و اگر نظراتشان بسیار متفاوت باشد (مثلا یکی نظر منفی و دیگری نظر مثبت داشته باشد) عدد ۱- به ارتباطشان اختصاص داده می‌شود.

همانطور که قبلا گفته شد، امکان دارد که هر یالی که بین دو کاربر رسم می‌شود، مربوط به ارتباطات زیادی که این دو کاربر با یکدیگر با توجه به امتیازدهی به فیلم‌های مشترک مختلف دارند، باشد. برای مثال، اگر دو کاربر به ۳ فیلم مشترک امتیاز داده باشند و نظر آن‌ها در مورد دو فیلم یکسان بوده ولی در مورد فیلم آخر، یکی نظر مثبت و دیگری نظر منفی داشته باشد، وزن نهایی یال مربوط به این دو کاربر برابر (۱-)+۱+۱ و برابر ۱ خواهد بود. با توجه به مطالب گفته شده، پس از این‌که وزن نهایی مربوط به هر یک از یال‌ها محاسبه شد، برای اجتناب از اختصاص وزن کل منفی به یک یال، کلیه‌ی وزن یال‌هایی که منفی شدند، با وزن صفر جایگزین می‌شوند. این بدان معناست که اگر مجموع نظرات دو کاربر در مورد فیلم‌های مختلف و گوناگون بسیار متفاوت بوده باشد به طوری که میانگین کلی آن صفر یا کمتر باشد، وزن یال رسم شده بین دو کاربر که نمایانگر میزان اشتراک نظر دو کاربر در مورد کل فیلم‌هایی است که به طور مشترک به امتیازدهی به آن‌ها پرداختند، بایستی از کمترین وزن ممکن برخوردار باشد. پس از تصحیح وزن در شبکه‌ی اصلی، اقدام به اعمال الگوریتم تشخیص اجتماع بر روی شبکه‌ی معنایی بدست آمده می‌کنیم. در این حالت، به مقدار پودمانی برابر ۰.۱۶۲۷ خواهیم رسید.

جدول زیر، مشخصات شبکه های ایجاد شده در هر یک از چارچوب‌ها را نشان می‌دهد:

جدول (۱): مشخصات شبکه‌های ایجاد شده

	نام خوشه	تعداد گره‌ها	تعداد یال‌ها
چارچوب۱	-	۵۹۲	۷۶۷۰۵
چارچوب۲	مستند	۳۵۲	۱۵۸۳۳
	وسترن	۴۹۱	۶۹۳۶۹
چارچوب۳		۵۹۲	۷۶۷۰۵

چارچوب اول، بیانگر حالتی است که ما تنها به اعمال الگوریتم تشخیص اجتماع بر روی شبکه‌ی اصلی می‌کنیم و هیچگونه تحلیل محتوایی صورت نمی‌پذیرد. چارچوب دوم همان تشخیص اجتماع مبتنی بر موضوع بوده و چارچوب سوم نیز تشخیص اجتماع بر روی شبکه معنایی می‌باشد. از آنجایی که در چارچوب دوم ما اقدام به یافتن خوشه‌های موضوعی می‌کنیم، بنابراین پس از نسبت دادن گره‌ها به خوشه‌های موضوعی، شبکه‌ی اصلی به دو بخش تقسیم می‌شود که هر یک از این بخش‌ها تنها شامل گره‌هایی خواهند بود که دارای موضوع مشترکند. لذا در جدول بالا در قسمت مربوط به چارچوب دوم، شبکه به دو بخش مستند و وسترن تقسیم شده است.

در نهایت، با توجه به رابطه‌ی (۳)، میزان $PurQ_\beta$ را در هر مرحله محاسبه می‌کنیم. جدول زیر، نشان‌دهنده‌ی نتایج نهایی اعمال چارچوب‌های مختلف بر روی مجموعه داده‌ی MovieLens می‌باشد:

جدول (۲): نتایج نهایی اعمال چارچوب‌های مختلف

	نام خوشه	مقدار پودمانی (Q)	مقدار پودمانی کل	مقدار $Purity$	مقدار $PurQ_\beta$
چارچوب۱	-	۰.۱۰۸۶	۰.۱۰۸۶	۰.۹۷۷۶۵۴	۰.۱۹۱۵۴۵
چارچوب۲	مستند	۰.۲۶۸۶	۰.۱۲۴۴۳۷	۱	۰.۲۲۱۳۳۲
	وسترن	۰.۱۰۳			
چارچوب۳		۰.۱۶۲۷	۰.۱۶۲۷	۰.۸۹۴۷۹۲۸	۰.۲۵۰۴۲۲

از آنجایی که ما به دنبال اجتماعاتی هستیم که این اجتماعات دربرگیرنده‌ی اعضایی باشند که این اعضا، هم دارای ارتباط تنگاتنگ با یکدیگر و هم دارای موضوع مشترک باشند، لذا میزان خلوص

موضوعات و ساختار شبکه برای ما به یک اندازه اهمیت دارند. بنابراین، برای محاسبه‌ی $PurQ_\beta$ مقدار β را برابر ۱ درنظر گرفتیم.

نتیجه ای که می‌توان از جدول بالا بدست آورد، این است که در هر دو چارچوبی که به تحلیل محتوا پرداختند نسبت به حالتی که تحلیل محتوایی انجام نشد، میزان پودمانی افزایش داشته است، با این تفاوت که در چارچوب سوم، با توجه به تحلیل محتوای مربوط به هر گره و اعمال مستقیم آن به صورت وزن‌دهی بر روی یال‌ها، تاثیر بیشتری را بر روی مقدار پودمانی شاهد بودیم. همچنین، واضح است که در چارچوب دوم، از آنجایی که خوشه‌های موضوعی جدا شدند و الگوریتم تشخیص اجتماع بر روی این خوشه‌ها اعمال گردید، میزان خلوص موضوعات ($Purity$) بالاتر از دو چارچوب دیگر می‌باشد. در مجموع، با در نظر گرفتن میزان $PurQ_\beta$ برای ۳ چارچوب پیاده سازی شده، واضح است که چارچوب‌هایی که تنها به ساختار و پیوند توجه نکرده‌اند و به تحلیل محتوا نیز پرداخته اند (چارچوب‌های ۲ و ۳)، به نتایج بهتری منجر شده‌اند. بنابراین می‌توان نتیجه گرفت که استفاده از نتایج تحلیل محتوا، موجب بهبود کارایی عمل تشخیص اجتماع می‌گردد.

۵- نتیجه

در این مقاله، به بررسی روش‌هایی که از نتایج حاصل از تحلیل محتوای یک شبکه به منظور بهبود تشخیص اجتماعات استفاده کردند، پرداخته شد. بدین‌منظور، دو چارچوب تشخیص اجتماع مبتنی بر عنوان و تشخیص اجتماع بر روی شبکه‌ی معنایی پیاده‌سازی شده و بر روی مجموعه داده‌ی MovieLens اعمال شدند. نتایج حاصل از اعمال این دو چارچوب و مقایسه‌ی آن با نتایج بدست آمده از حالتی که تشخیص اجتماع بدون تحلیل محتوا صورت می‌پذیرد، بیانگر این واقعیت بود که تحلیل محتوا در یک شبکه و استفاده از نتایج آن منجر به بهبود عمل تشخیص اجتماع می‌گردد.

از آنجایی که تشخیص اجتماع مبتنی بر عنوان، نسبت به تشخیص اجتماع بر روی شبکه‌ی معنایی و نسبت به روشی که تشخیص اجتماع، بدون تحلیل محتوای شبکه صورت می‌پذیرد، منجر به یافتن اجتماعاتی با میزان خلوص موضوع ($Purity$) بالاتری شد و از طرف دیگر تشخیص اجتماع بر روی شبکه‌ی معنایی، نسبت به تشخیص اجتماع مبتنی بر عنوان و نسبت به روشی که تشخیص اجتماع، بدون تحلیل محتوای شبکه صورت می‌پذیرد، منجر به رسیدن به مقدار پودمانی بالاتری شد، لذا در کارهای آینده بنا داریم تا چارچوبی ارائه دهیم که از مزیت مربوط به تشخیص اجتماع مبتنی بر عنوان و تشخیص اجتماع بر روی شبکه معنایی به طور همزمان استفاده کند. یعنی این چارچوب بتواند به طور همزمان هم میزان پودمانی و هم میزان خلوص موضوعات را افزایش دهد.

مراجع

[۱] حسین‌زاده، رسول، علیزاده، حسین، ناظمی، اسلام، **"تشخیص اجتماعات با رویکرد ترکیبی در شبکه‌های اجتماعی"**، یازدهمین کنفرانس سیستم‌های هوشمند ایران، تهران، اسفند ۱۳۹۱.

[۲] ریحانیان، علی، علیزاده، حسین، مینایی، بهروز، **"تشخیص اجتماع مبتنی بر موضوع، گامی نوین در تحلیل شبکه‌های اجتماعی"**، همایش تخصصی بررسی ابعاد شبکه‌های اجتماعی، سلسله همایش-های فضای سایبر (۳)، تهران، سالن همایش‌های بین‌المللی جمهوری اسلامی ایران، مهر ۱۳۹۲.

[3] Webb, A.R., *Statistical pattern recognition*. 2003: Wiley.

[4] Easley, D. and J. Kleinberg, *Networks, crowds, and markets*, Vol. 8. 2010: Cambridge Univ Press.

[5] Leskovec, J., K.J. Lang, and M. Mahoney. *"Empirical comparison of algorithms for network community detection"*, in Proceedings of the 19th international conference on World Wide Web. 2010. ACM.

[6] Newman, M., Communities, "*modules and large-scale structure in networks"*, Nature Physics, 2011.

[7] Porter, M.A., J.-P. Onnela, and P.J. Mucha, "*Communities in networks"*, Notices of the AMS, 2009. 56(9): p. 1082-1097.

[8] Zhao, Z., et al., *"Topic oriented community detection through social objects and link analysis in social networks"*, Knowledge-Based Systems, 2012. 26: p. 164-173.

[9] Lancichinetti, A. and S. Fortunato, *"Consensus clustering in complex networks"*, Scientific reports, 2012. 2.

[10] Newman, M.E. and M. Girvan, *"Finding and evaluating community structure in networks"*. Physical review E, 2004. 69(2): p. 026113.

[11] Masdarolomoor, Z., et al. *"Distributed Community Detection in Complex Networks"*. in Computational Intelligence, Communication Systems and Networks (CICSyN), 2011 Third International Conference on. 2011. IEEE.

[12] Xia, Z. and Z. Bu, *"Community detection based on a semantic network"*. Knowledge-Based Systems, 2012. 26: p. 30-39.

زیرنویس‌ها

[1] Complex Networks
[2] Entities
[3] World Wide Web
[4] Social Networks
[5] Online Transactions Network
[6] Telephone Call Network
[7] Network of protein-protein interactions

[8] Citation Network
[9] Content Analysis
[10] Topic oriented community detection
[11] Topical Clusters
[12] Community detection based on a semantic network
[13] Adjacency Matrix
[14] Module
[15] Cluster
[16] Subnetwork
[17] Dense
[18] Density
[19] Community structure
[20] Collaborative recommendation
[21] Information spreading
[22] Knowledge sharing
[23] Modularity
[24] Girvan
[25] Newman
[26] Modularity maximization
[27] Social network data modeling
[28] Social objects clustering
[29] Social members partitioning
[30] Link analysis
[31] Formal models
[32] Social objects
[33] Text social objects
[34] Strength
[35] Tightly connected members
[36] Link analysis
[37] Modularity maximization methods
[38] Social object clustering
[39] Topic-based user partitioning
[40] Link-based community detection
[41] Supportive
[42] Opposing
[43] Trust value
[44] Rating
[45] Documentary
[46] Western
[47] Sentiment Analysis
[48] Positive
[49] Negative
[50] Neutral