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

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

مبانی نظری تحقیق:

نحوه نمایش شبکه: می‌توان شبکه‌ها را به دو صورت نمایش داد، به صورت گراف و به صورت ماتریس. در نمایش گرافی شبکه، ساختارهای گرافی، وظیفه متصل کردن گره‌ها و یال‌ها به یکدیگر را بر عهده دارند. به منظور نمایش ماتریسی شبکه، یک ماتریس مربعی ایجاد می‌شود که به آن ماتریس مجاورتی گراف شبکه گفته می‌شود. ابعاد این ماتریس، برابر تعداد گره‌های 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.

تعریف اجتماع: در مقالات و تحقیقات علمی مختلف، تعاریف متفاوتی از اجتماع ارائه شده است. یکی از این تعاریف که در مقاله‌ی ۱ آمده است بیان می‌کند که یک اجتماع شبکه که گاهی از اوقات به آن پیمانه یا خوشه 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.

تابع پودمانی: مشهورترین معیار ارزیابی اجتماعات تشخیص داده شده توسط یک الگوریتم تشخیص اجتماع، پودمانی می‌باشد. 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. 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 پودمانی 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 پودمانی نزدیک به ۱- خواهد بود.

تابع $PurQ_\beta$: در مقاله ی ۲، چارچوبی پیشنهاد شد که این چارچوب مباحث موضوع و پیوند را با یکدیگر ترکیب کرده و به عمل تشخیص اجتماع می‌پردازد. لذا به منظور ارزیابی نتایج حاصل از این چارچوب پیشنهادی، یک معیار جدید معرفی شد. بدیهی است که چنین معیاری، بایستی به بررسی اجتماعات از دو جنبه‌ی موضوع و پیوند بپردازد. این معیار که با نام $PurQ_\beta$ معرفی شده است، به صورت زیر تعریف می‌شود:

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

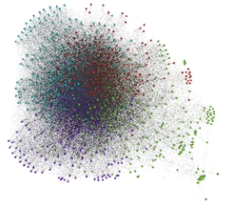
در رابطه‌ی بالا، Purity، میزان خلوص موضوعات، در اجتماعات تشخیص داده شده را مشخص می‌کند. هرچقدر مقدار Purity بالاتر باشد، نشان‌دهنده‌ی این امر خواهد بود که اجتماعات، از جنبه‌ی موضوع، بهتر افراز بندی شده‌اند. Q نیز نمایانگر پودمانی می‌باشد. همچنین در رابطه‌ی بالا، β پارامتری برای تنظیم وزن Purity و Q می‌باشد و می تواند دارای مقادیر از ۰ تا بینهایت باشد. اگر مقدار β برابر با ۱ باشد، می تواند اینگونه در نظر گرفته شود که میزان خلوص موضوعات و ساختار شبکه اجتماعی به یک اندازه اهمیت دارند.

روش تحقیق:

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 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 مطرح شده است، اشاره کرد. در این رویکرد، ابتدا خوشه‌های موضوعی 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 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 شدند.

یافته های تحقیق:

در این تحقیق، چارچوب‌های ارائه شده در روش‌های تشخیص اجتماع مبتنی بر موضوع و تشخیص اجتماع بر روی شبکه‌ی معنایی، بر روی یکی از مجموعه داده‌های MovieLens اعمال شدند. این مجموعه داده، شامل ۱۰۰۰۰۰ امتیازدهی اعمال شده به ۱۶۸۲ فیلم توسط ۹۴۳ کاربر می‌باشد. این امتیازها از ۱ تا ۵ می‌باشند. به منظور پیاده‌سازی چارچوب‌های ذکرشده، کلیه‌ی فیلم‌هایی که در موضوع مستند یا موضوع وسترن 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. سپس 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. شکل زیر نمایی از این شبکه (که ما به آن شبکه ی اصلی می‌گوییم) را نشان می‌دهد:



شکل (۱): نمایی از شبکه‌ی اصلی

جدول زیر نیز مشخصات شبکه‌های 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	۶۹۲۶۹
چارچوب۳		۵۹۲	۷۶۷۰۵

چارچوب اول، بیانگر حالتی است که ما تنها اقدام 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 (۱)، 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	۰.۱۲۴۴۲۷	۱	۰.۲۲۱۳۲۲
	وسترن	۰.۱۰۳			
چارچوب۳		۰.۱۶۲۷	۰.۱۶۲۷	۰.۸۹۴۷۹۲۸	۰.۲۵۰۴۲۲

از آنجایی که ما به دنبال اجتماعاتی هستیم که 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 پرداخته‌اند (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.

نتیجه‌گیری:

در این مقاله، به بررسی روش‌هایی که از نتایج حاصل از تحلیل محتوای یک شبکه به منظور بهبود تشخیص اجتماعات استفاده کردند، پرداخته شد. بدین منظور، 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.

منابع:

1. 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.
2. 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.
3. Xia, Z. and Z. Bu, *"Community detection based on a semantic network"*. Knowledge-Based Systems, 2012. 26: p. 30-39.
4. Easley, D. and J. Kleinberg, *Networks, crowds, and markets*, Vol. 8. 2010: Cambridge Univ Press.
5. Newman, M., Communities, *"modules and large-scale structure in networks"*, Nature Physics, 2011.
6. Porter, M.A., J.-P. Onnela, and P.J. Mucha, *"Communities in networks"*, Notices of the AMS, 2009. 56(9): p. 1082-1097.
7. Lancichinetti, A. and S. Fortunato, *"Consensus clustering in complex networks"*, Scientific reports, 2012. 2.