مجله علمی پژوهشی «علوم و فناوری‌های پدافند نوین»
سال نهم، شماره ۲، تابستان ۱۳۹۷؛ ص ۲۱۰-۲۰۱

خوشه‌بندی محتوایی- ساختاری گراف و معیاری جدید جهت ارزیابی آن

کبری رحمتی[۱]، حسن نادری[۲*]، سامان کشوری[۳]
۱- دانشجوی کارشناسی‌ارشد ۲- استادیار دانشگاه علم و صنعت ایران، ۳- دانشجوی کارشناسی‌ارشد دانشگاه جامع امام حسین (ع)
(دریافت: ۹۵/۱۰/۲۰، پذیرش: ۹۶/۰۲/۲۴)

چکیده

امروزه با گسترش شبکه‌های اجتماعی در بین مردم، تلاش‌های مخالفین برای بدبین کردن ایشان نسبت به حکومت که از آن به عنوان جنگ نرم یاد می‌شود افزایش یافته است، بنابراین توجه به این شبکه‌ها برای ارگان‌های نظامی و امنیتی بیش از پیش اهمیت دارد. خوشه‌بندی گراف از جمله اولین کارهای تحلیلی یک یا چند شبکه اجتماعی است. متأسفانه اکثر خوشه‌بندی‌های گرافی انجام شده بر روی جنبه‌های ساختاری یا محتوایی گره‌های گراف به صورت مستقل تأکید دارند. هدف از این مقاله (پیاده‌سازی شده در قالب الگوریتم CS-Cluster) رسیدن به خوشه‌هایی با ساختار درونی منسجم و مقادیر ویژگی (محتوایی) همگن در گراف است. از طرفی پس از جستجوهای صورت گرفته در این تحقیق، هیچگونه معیاری جهت ارزیابی الگوریتم‌های خوشه‌بندی که جنبه‌های ساختاری و محتوایی گره‌ها را به صورت هم‌زمان در نظر بگیرد، یافت نشد. به همین دلیل در دومین گام معیاری جدید به نام CS-Measure ارائه شد که قادر است الگوریتم‌های خوشه‌بندی گراف را از هر دو جنبه ساختار و محتوا به صورت هم‌زمان مورد سنجش قرار دهد. مقایسه الگوریتم مطرح شده با دو الگوریتم خوشه‌بندی ساختاری-محتوایی (از سه الگوریتم شناخته شده تاکنون) بر اساس معیارهای میانگین شباهت، خطای یال و معیارِ جدیدِ ساختاری-محتوایی، بیانگر عملکرد بهتر روش ارائه شده است و از نظر معیار تراکم نیز عملکرد نسبتاً خوبی دارد.

کلید واژه‌ها: خوشه‌بندی، گراف محتوایی، خوشه‌بندی ساختاری- محتوایی، ارزیابی ساختاری- محتوایی

Content-Structural Graph Clustering and a New Criterion for its Evaluation

K. Rahmati, H. Naderi*, S. Keshvari
Iran University of Science and Technology
(Received: 09/01/2017; Accepted: 14/05/2017)

Abstract

Today, with the spread of social networks, the opposition's efforts to chill out people from government (known as "soft war") are increased. Therefore, dealing with this type of networks is important for military and security organizations. Graph clustering is one of the first attempts toward analyzing social networks which can appropriately be modeled by a content graph. In contrast, most of the existing graph clustering methods independently focused on one of the content or structural aspects of a graph. The aim of this paper (implemented as CS-Cluster algorithm) is to achieve well connected clusters while their nodes benefits from homogeneous attribute values (content). In the second step of our research, after an intensive search, no measure has found which could simultaneously consider content and structural features of clustering algorithms. So to be able to appropriately evaluate our algorithm, a new content-structural measure (so-called "CS-Measure") is proposed. Our experimentation shows that the proposed clustering algorithm outperforms two other well-known content-structural clustering algorithms, using the new content-structural, average similarity, and Error link measure as well as the previous content and structural measures, And it also performed relatively well in density measure.

Keywords: Clustering, Content Graph, Content-Structural Clustering, Content-Structural Evaluation

.

*Corresponding Author E-mail: Naderi@iust.ac.ir

۱. مقدمه

خوشه‌بندی مسئله‌ای مهم در تحلیل داده‌های گراف و یک روش دسته‌بندی بدون ناظر برای داده‌های گراف است. خوشه‌بندی شامل تقسیم گره‌های گراف به گروه‌هایی با گره‌های مشابه است. شباهت بین گره‌ها معمولاً توسط یک تابع هدف ریاضی تعریف می‌شود [۱]. معمولاً مسئله خوشه‌بندی گراف در رده مسائل NP-Hard قرار می‌گیرد. حل این مسائل به صورت کلی از طریق روش‌های مکاشفه‌ای[1] و تقریبی حاصل می‌شود. این توابع هدف در روش‌های مکاشفه‌ای به دو شیوه محلی و سراسری تعریف می‌شوند [۲ و ۳].

در بسیاری از کاربردهای نظامی، به منظور تحلیل فعالیت‌های دشمن بهتر است ارتباطات برقرار شده به صورت گراف مدل شوند. در این گراف علاوه بر ساختار محتوای گره‌ها نیز مهم هستند. به عنوان مثال در رصد اطلاعاتی، دشمن ارتباطاتی از طریق پیام‌های صوتی و متنی برقرار می‌کند که علاوه بر اینکه این ارتباطات دارای ساختار مشخصی بوده (افراد به عنوان رأس و ارتباطات به صورت یال در گراف تشکیل شده)، دارای محتوا نیز هستند، بنابراین در این مثال توجه به ساختار و محتوا در کنار یکدیگر از اهمیت بالایی برخوردار است.

توجه صرفاً به ساختار گراف تشکیل شده کارایی مناسبی برای خوشه‌بندی ندارد به همین دلیل افرادی که محتوای مطابق با معیارهای مد نظر تولید می‌کند با کسی که محتوای مورد نظر را تولید نمی‌کند در این خوشه‌بندی یکسان در نظر گرفته می‌شود.

توجه صرفاً به محتوا نیز اثربخش نیست زیرا به عنوان مثال اگر بخواهید فرماندهان رده بالای دشمن را در رصد اطلاعاتی در یک خوشه قرار دهید، توجه صرفاً به محتوا ارتباطات را در نظر نگرفته و یک فرد عادی را با توجه به محتوای رد و بدل شده -که ممکن است با معیارهای محتوایی تعریف شده مطابقت داشته باشد- در خوشه فرماندهان قرار دهد. اغلب روش‌های خوشه‌بندی که تاکنون ارائه شده است فقط یک جنبه ساختاری یا محتوایی را در نظر گرفته‌اند و روش‌های خوشه‌بندی که هم ساختار و هم محتوا را در نظر بگیرند، کمتر ارائه شده است. همین امر باعث شده است که معیارهای موجود برای ارزیابی جوامع نیز اغلب بر اساس ویژگی‌های ساختاری گراف باشند.

در حالی که هیچ کدام از این معیارها به تنهایی برای ارزیابی هم‌زمان ویژگی‌های ساختاری و محتوایی کاربرد مناسب و دقیقی ندارند، بنابراین نمی‌توان از هیچ کدام از این معیارها برای ارزیابی الگوریتم‌های خوشه‌بندی ساختاری- محتوایی استفاده کرد؛ در نتیجه به یک معیار ارزیابی دقیق‌تر و کامل‌تر نیاز است که خوشه‌بندی ساختاری- محتوایی را به طور هم‌زمان از نظر ساختاری و محتوایی ارزیابی کند.

هدف کلی، تشخیص خوشه‌هایی از گراف است که گره‌های درونی هر خوشه به طور متراکم به یکدیگر مرتبط بوده و از لحاظ معنایی شبیه به هم باشند و ارتباط ساختاری و معنایی بین خوشه‌های مجزا به حداقل خود برسد. در این مقاله یک روش خوشه‌بندی مبتنی بر انتشار برچسب ارائه شده است که با ایجاد تعادل بین ویژگی‌های ساختاری و محتوایی گراف، عمل خوشه بندی ساختاری- محتوایی را به نحو مطلوبی انجام می‌دهد. برای ارزیابی دقیق‌تر خوشه‌بندی‌های ساختاری- محتوایی برای اولین بار، معیاری به نام CS-Measure[2] پیشنهاد شده است که با ادغام هر دو مفهوم ساختار و محتوا، عمل ارزیابی خوشه‌بندی‌های ساختاری- محتوایی را به طور مناسبی انجام می‌دهد. نتایج آزمایش‌های تجربی نشان می‌دهد که خوشه بند پیشنهاد شده به نحو مطلوبی عمل خوشه‌بندی با ویژگی‌های مورد نظر را انجام می‌دهد.

خوشه‌بندی گراف یکی از راه‌های تحلیل گراف‌های حجیم و پیچیده است و تاکنون روش‌های خوشه‌بندی مختلفی برای آن ارائه و استفاده شده است. بیشتر این روش‌ها برای عمل خوشه بندی تنها جنبه ساختاری گراف را در نظر می‌گیرند [۷-۴]، که از این دسته می‌توان TopGC و Louvain را نام برد [۸ و ۹]. برخی روش‌ها نیز مطرح شده‌اند که عمل خوشه‌بندی را بر اساس محتویات گره‌ها انجام می‌دهند، از جمله روش‌های k-Means و k-Medoids [۸ و ۱۲-۱۰]. در حالی که در بسیاری از کاربردهای دنیای واقعی هر دو جنبه ساختار و محتوا در کنار یکدیگر و به طور هم‌زمان مورد نظر هستند[۱۶-۱۳]، پس بهتر است در خوشه‌بندی شباهت ساختاری و محتوایی گره‌ها در کنار یکدیگر در نظر گرفته شود. تاکنون تعداد کمی الگوریتم خوشه‌بندی که به صورت ساختاری- محتوایی به عمل خوشه‌بندی می‌پردازند، ارائه شده است که برخی از آن‌ها عبارتند از:

SA-Cluster[3]: در این الگوریتم به ازای هر خصوصیت موجود در هر گره، یک گره خصوصیت به گراف اضافه می‌شود و گره‌هایی که دارای آن خصوصیت هستند از طریق یک یال خصوصیت به آن گره وصل می‌شوند. این الگوریتم یک الگوریتم تکرار شونده است که در آن ابتدا مرکزی‌ترین گره‌ها (گره‌ها با تراکم بالا) به عنوان مرکز خوشه انتخاب شده و سپس باقی گره‌ها به نزدیک ترین مراکز، تخصیص داده می‌شوند. سپس در هر خوشه مراکز

[1] Heuristic Methods

[2] Structural-Content-Measure

[3] Structural Attribute Cluster

به‌هنگام رسانی می‌شوند (همچنین وزن یال‌ها که بر اساس ویژگی‌های محتوایی و ساختاری تنظیم شده‌اند و سایر پارامترهای الگوریتم به‌روز می‌شود) و عملیات تا همگرا شدن تکرار می‌شود [۱۳]. از آنجا که این الگوریتم به صورت تکرار شونده است، زمان اجرای بالایی دارد.

SANS[1]: در این الگوریتم، ابتدا در گراف وزن‌دار ورودی شاخص وزن هر گره -که به صورت مجموع وزن یال‌های متصل به گره است- محاسبه شده و گره‌ای که بیشترین شاخص وزن را دارد به عنوان مرکز[2] خوشه انتخاب می‌شود؛ سپس گره‌های همسایه گره مرکزی به خوشه مربوط به آن اضافه می‌شوند و در ادامه گره‌هایی که با گره‌های درون خوشه شباهت (شباهت محتوایی) بیش از حد آستانه دارند نیز به خوشه اضافه می‌شوند. پس از طی این مراحل، مجدد از بین گره‌های باقیمانده گره‌ای که بیشترین شاخص وزن را دارد به عنوان مرکز بعد انتخاب می‌شود و عملیات تا خوشه‌بندی همه گره‌ها تکرار می‌شود [۱۴]. زمان اجرای این الگوریتم نسبت به SA-Cluster بهتر است و نسبت به الگوریتم SA-Cluster خوشه‌های متراکمتری به‌دست می‌آورد.

DCM[3]: این الگوریتم از مجموعه‌ای از جوامع کاندید شروع کرده و دو گام اصلی را به طور متناوب تا رسیدن به همگرایی تکرار می‌کند. در گام اول جوامع با بهترین امتیاز جامعه (از نظر ساختاری) را به‌دست آورده و در گام بعد سعی می‌کند یک توصیف مناسب برای این جوامع به‌دست آورد. برای تطابق بیشتر توصیف با جوامع، ممکن است در صورت لزوم جامعه را مقداری تغییر داده و گام‌های الگوریتم را تا همگرا شدن تکرار کند [۱۵]. این روش لزوماً همه گره‌ها را پوشش نمی‌دهد و برای مجموعه داده‌های کوچک کند و برای مجموعه داده‌های بزرگ سرعت بالا دارد.

۳. روش پیشنهادی

یک گراف ویژگی را به صورت G=(V,E,A) نشان داده می‌شود، به طوری‌که در آن V مجموعه گره‌ها، E مجموعه یال‌ها و A مجموعه خصوصیات هر گره گراف را نشان می‌دهد. به هر گره $v_i \in V$ یک بردار ویژگی $[a_1(v_i), \dots, a_m(v_i)]$ تخصیص داده می‌شود که در آن، $a_j(v_i)$ مقدار خصوصیت a_j در گره v_i است. سعی بر آن است گراف ویژگی G به صورتی خوشه‌بندی شود که هر کدام از خوشه‌ها علاوه بر نزدیکی ساختاری، از لحاظ محتوایی نیز مشابه باشند. در ادامه روش ارائه شده تشریح می‌شود.

[1] Structural Attribute Neighbourhood Similarity
[2] Centroid
[3] Description-Driven Community Detection

۳-۱. روش ارائه شده (CS-Cluster[4])

همان‌طور که گفته شد در گراف G هر گره حاوی تعدادی ویژگی است. در راه‌حل پیشنهادی شباهت جاکارد[5] محتوای هر دو گره محاسبه شده و به عنوان وزن یال بین دو گره در نظر گرفته می‌شود. شباهت جاکارد دو مجموعه C و C' با استفاده از فرمول (۱) محاسبه می‌شود [۱۵]. با این کار گراف G به یک گراف وزن دار G'_w تبدیل خواهد شد. در این مقاله منظور از وزن یال، مقدار شباهت محتوایی دو گره متصل به آن یال است. روش خوشه‌بندی ارائه شده در این مقاله CS-Cluster نامیده شده است. این روشِ خوشه‌بندی مبتنی بر انتشار برچسب است که با ترکیب مفاهیم ساختاری و محتوایی عمل خوشه‌بندی را به نحو مطلوبی انجام می‌دهد.

$$J(C,C') = \frac{|C \cap C'|}{|C \cup C'|} \qquad (۱)$$

در این در این روش ابتدا حد آستانه شباهت و حداکثر فاصله بین مراکز خوشه‌ها تعیین می‌شود، با تعیین این دو معیار، تعداد خوشه‌ها به طور خودکار مشخص خواهد شد. در واقع، حد آستانه شباهت برابر با میانگین شباهت گره‌های گراف است. شباهت جاکاردِ محتوای هر جفت گره در گراف نیز محاسبه شده و در ماتریس شباهت نگهداری می‌شود. به عنوان مثال برای گراف G_w شکل (۱)، مقادیر شباهت محتوای هر جفت گره به صورت زیر است:

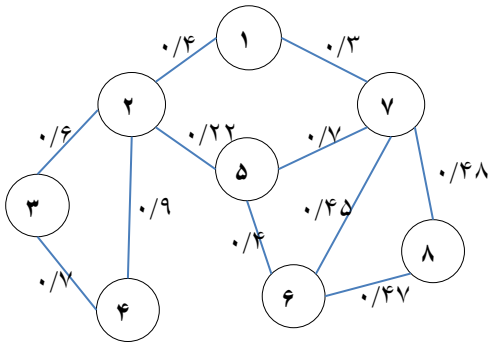
$$sim = \begin{bmatrix} - & 0.40 & 0.52 & 0.60 & 0.32 & 0.30 & 0.30 & 0.15 \\ - & - & 0.60 & 0.90 & 0.22 & 0.48 & 0.19 & 0.36 \\ - & - & - & 0.70 & 0.42 & 0.52 & 0.21 & 0.32 \\ - & - & - & - & 0.30 & 0.23 & 0.09 & 0.16 \\ - & - & - & - & - & 0.40 & 0.70 & 0.50 \\ - & - & - & - & - & - & 0.45 & 0.47 \\ - & - & - & - & - & - & - & 0.48 \\ - & - & - & - & - & - & - & - \end{bmatrix}$$

در اولین گامِ این روش با توجه به گراف G_w (گراف اصلی) گرافی به نام G'_w تشکیل می‌شود که در آن هر یال نشان دهنده وجود شباهت بیش از حد آستانه بین دو گره است و وزن یال بیانگر فاصله دو گره بر اساس تعداد یال‌های بین آن دو گره در گراف G_w است.

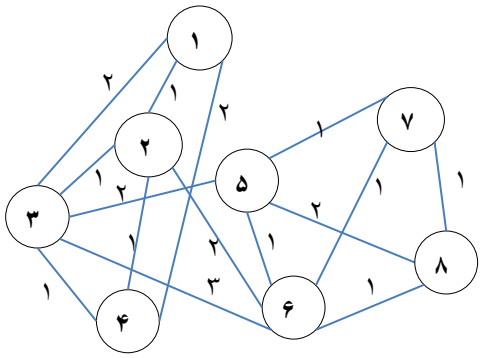
به منظور تشکیل گراف G'_w، در گراف G_w شباهت هر گره با سایر گره‌ها بررسی می‌شود، اگر شباهت بین آن‌ها بیش از حد آستانه شباهت مورد نظر بود بین آن دو گره یک یال رسم کرده و وزن آن یال برابر فاصله بین آن دو گره خواهد بود (این فاصله

[4] Structural Content Clustering
[5] Jaccard Similarity

برابر تعداد یال‌های بین آن دو گره است). شکل (۱) نمونه‌ای از تشکیل گراف G'_w را نشان می‌دهد که حد آستانه شباهت مورد نظر در آن ۰/۴ است. همانطور که گفته شد، این حد آستانه برابر با میانگین شباهت محتوایی (شباهت جاکارد) در ماتریس شباهت است. حداقل فاصله مجاز بین مراکز خوشه‌ها نیز در این مثال برابر با ۲ در نظر گرفته شده است تا در نهایت گراف مورد نظر شامل دو خوشه شود.



(الف)



(ب)

شکل ۱. (الف) گراف $\mathbf{G_w}$ و (ب) گراف $\mathbf{G'_w}$

پس از به‌دست آوردن گراف G'_w، گره‌های مرکزی به روشی که در ادامه شرح داده می‌شود، انتخاب می‌شوند. سپس گره‌های مرکزی به ترتیب با بررسی گره‌های مجاور خود در G_w مشابه ترین گره را به خوشه مربوط به خود اضافه می‌کنند و به همین ترتیب گره‌های مجاور گره‌های داخل خوشه بررسی شده و در صورت داشتن شباهت بیشتر از حد آستانه به خوشه مورد نظر اضافه می‌شوند. برای انتخاب مراکز خوشه‌ها، با توجه به هدف مسئله بهتر است از معیار مرکزیتی استفاده شود که شباهت‌های ساختاری و محتوایی را به طور هم‌زمان در نظر بگیرد. به همین منظور یک معیار مرکزیت ساختاری- محتوایی تعریف شده است که وضعیت ساختاری و محتوایی گره‌ها را به طور هم‌زمان در نظر می‌گیرد. این معیار مرکزیت به وسیله فرمول (۲) برای هر گره در گراف G'_w محاسبه می‌شود:

$$C_v = \frac{(D_v)^2}{\sum_{e \in E_v} w_e} \qquad (۲)$$

که در آن، C_v امتیاز مرکزیت گره v، E_v مجموعه یال‌های متصل به v در گراف G'_w، w_e وزن یال e در گراف G'_w و D_v درجه گره v در گراف G'_w است.

در این روش، گره‌ای که با گره‌های بیشتری در گراف شباهت محتوایی بیش از حد آستانه دارد و در کمترین فاصله ممکن با آن‌ها است، به عنوان گره مرکزی انتخاب می‌شود. به عبارتی، گره مرکزی دارای بیشترین درجه با کمترین حاصل جمع وزن یال های متصل است (یال‌های متصل نشان دهنده وجود شباهت محتوایی بیش از حد آستانه مورد نظر هستند و وزن آن‌ها بیانگر فاصله ساختاری دو گره است)؛ یعنی اگر مجموع وزن یال‌های متصل به گره کمتر باشد گره‌های مشابه در فاصله کمتری از گره مورد نظر قرار گرفته‌اند. پس از محاسبه امتیاز مرکزیت گره‌ها، به ترتیب گره‌های با امتیاز بیشتر را به شرط اینکه نسبت به مراکز انتخاب شده قبلی فاصله‌ای کمتر از حد مجاز نداشته باشند، به عنوان مرکز انتخاب می‌شوند. مقادیر مرکزیت گره‌های موجود در گراف شکل (۱)، در جدول (۱) مشخص شده است.

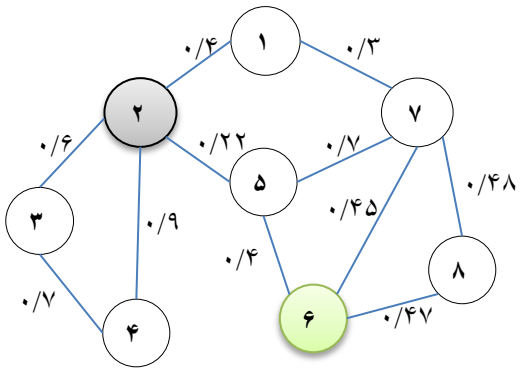
جدول ۱. مقدار مرکزیت گره‌های G'_w

شماره گره	۱	۲	۳	۴	۵	۶	۷	۸
مقدار مرکزیت	۱/۸	۳/۲	۲/۷	۲/۲	۲/۶	۳/۱	۳	۲/۲

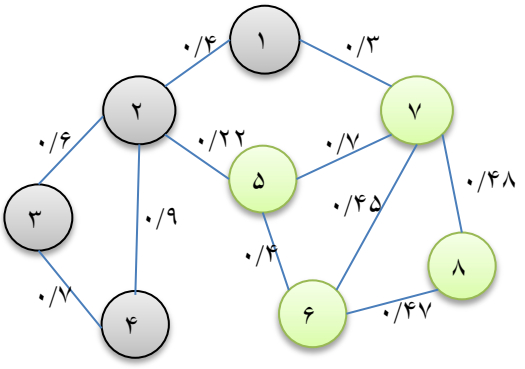
با توجه به اینکه گره ۲ بیشترین مقدار مرکزیت را دارد. به عنوان اولین مرکز خوشه در نظر گرفته می‌شود. پس از آن گره ۶ دارای بیشترین مقدار مرکزیت است؛ با توجه به اینکه گره ۶ و گره ۲ در گراف G_w دارای فاصله ۲ هستند و این مقدار برابر حداقل فاصله مجاز است.

همان‌طور که در شکل (۲) مشاهده می‌شود، گره ۶ به عنوان گره مرکزی بعدی در نظر گرفته می‌شود. گره بعدی که بیشترین مقدار را دارد، گره ۷ است. ولی به علت اینکه فاصله این گره از گره شماره ۶ برابر ۱ بوده و این مقدار از حداقل فاصله مجاز بین مراکز کمتر است؛ گره شماره ۷ نمی‌تواند به عنوان مرکز خوشه بعدی انتخاب شود. سایر گره‌ها نیز به همین ترتیب نمی‌توانند به عنوان گره مرکزی انتخاب شوند. چون در صورت انتخاب هر کدام از این گره‌ها به عنوان گره مرکزی حداقل فاصله مجاز بین مراکز رعایت نخواهد شد. در نتیجه درنهایت دو خوشه به‌دست خواهد آمد. پس از انتخاب گره‌های مرکزی، در هر مرحله گره‌ای که از طریق یالی با بیشترین وزن به خوشه متصل است را به شرط اینکه گره قبلاً در هیچ خوشه‌ای نباشد، به خوشه مربوطه اضافه می‌شود.

به همین منظور، گزینه‌های ممکن برای اضافه شدن به خوشه مربوط به گره مرکزی ۲، شامل گره‌های ۱، ۳، ۴ و ۵ است. از بین این گره‌ها چون گره ۴ از طریق یالی با وزن بیشتر به گره ۲ متصل است (شباهت بیشتری به گره ۴ دارد) به خوشه مربوط به گره ۲ اضافه می‌شود. پس از اضافه کردن گره ۴ به خوشه مربوطه، گزینه‌های ممکن برای اضافه شدن به خوشه مربوط به گره ۶ بررسی می‌شود. این گزینه‌ها عبارتند از گره‌های ۵، ۷ و ۸ که از بین آن‌ها گره ۸ که با یالی با وزن بیشتر به گره ۶ متصل است، به خوشه مربوطه اضافه می‌شود. سپس مجدد گره‌های متصل به گره مرکزی ۲ بررسی می‌شوند؛ در این مرحله گره ۳ به خوشه اضافه خواهد شد چون از طریق یالی با وزن بیشتر نسبت به سایر گره‌ها به گره ۴ متصل است. در ادامه به خوشه با مرکزیت گره ۶، گره ۷ اضافه شده و این مراحل تا زمانی که تمامی گره‌ها عضو خوشه شوند، ادامه خواهد یافت؛ خوشه‌های حاصل در شکل (۳) نشان داده شده است.



شکل ۲. مرکزهای خوشه‌ها در گراف



شکل ۳. خوشه‌های حاصل از روش پیشنهادی CS-Cluster

همان‌طور که شرح داده شد، در روش CS-Cluster در تمام مراحل خوشه‌بندی، هر دو جنبه ساختار و محتوا به صورت هم زمان در نظر گرفته می‌شود که این منجر به تشکیل خوشه‌های با معنا و کیفیت بالا می‌شود.

۴. نتایج و بحث

جهت ارزیابی روش ارائه شده با سایر روش‌های خوشه‌بندی ساختاری- محتوایی آزمایش‌های مختلفی بر اساس معیارهای مختلف انجام شده است. در ادامه، نتایج مقایسه آزمایش‌های انجام شده بر روی سه روش SANS، SA-Cluster و CS-Cluster آمده است. چون روش خوشه‌بندی DCM تمامی گره‌ها را پوشش نمی‌دهد، در مقایسه در نظر گرفته نشده است. در این آزمایش‌ها شباهت جاکارد بین گره‌ها محاسبه شده و به عنوان وزن یال‌ها در نظر گرفته می شود. با توجه به اینکه هر کدام از روش‌های مورد مقایسه دارای پارامترهای متفاوتی هستند. برای مقایسه این روش‌ها محدودیت‌هایی وجود دارد. به عنوان مثال با توجه به اینکه درروش SA-Cluster تعداد خوشه‌ها از ابتدا مشخص است ولی در روش پیشنهادی CS-Cluster مشخص نیست و تنها دو پارامتر حد آستانه شباهت و فاصله مراکز خوشه‌ها در اختیار است؛ همچنین در روش SANS نیز فقط پارامتر حد آستانه شباهت در دسترس است. در انجام آزمایش‌ها تنظیم پارامترهای این روش‌ها به گونه‌ای که هر سه روش تعداد خوشه‌های (k) یکسانی به‌دست آورند به سختی انجام شد. در نتیجه ۱۰ وضعیت که در آن‌ها تعداد خوشه‌های یکسانی حاصل می‌شود را با یکدیگر مقایسه شده است. البته در برخی از این موارد نیز تعداد خوشه های حاصل شده در هر روش اندکی تفاوت دارد. مثلاً در وضعیت ۱۰۸ خوشه، مقادیر مربوط به وضعیت ۱۱۰ خوشه در روش پیشنهادی CS-Cluster با مقادیر مربوط به وضعیت ۱۰۸ خوشه در SA-Cluster مقایسه شده است.

۴ -۱. مجموعه داده[1]: در آزمایش‌های انجام شده از مجموعه داده Delicious استفاده شده است. این مجموعه داده دارای ۱۸۶۱ گره و ۷۶۶۴ یال است. خصوصیت‌های هر گره -که تعداد آن‌ها ۱۳۵۰ عدد است- نیز به وسیله یک آرایه دودویی مشخص شده است. خصوصیات با استفاده از مقادیر ۰ و ۱ در آرایه مربوط به هر گره مشخص شده که برای تعیین وجود یا عدم وجود خصوصیت در یک گره استفاده می‌شوند. در واقع برای هر گره یک آرایه ۱۳۵۰ خانه‌ای وجود دارد که برای هر ویژگی یک خانه در نظر گرفته شده است. اگر گره دارای ویژگی مورد نظر بود، خانه مورد نظر برای آن ویژگی مقدار ۱ خواهد داشت، در غیر این صورت مقدار خانه برابر با ۰ خواهد بود. این کار باعث می‌شود که خصوصیات گره‌ها به صورت دودویی ذخیره شده تا هنگام محاسبه شباهت، بتوان از شباهت جاکارد به راحتی استفاده نمود. لازم به توضیح است که Delicius یک سرویس خدمات

[1] Dataset

Bookmark است که امکان برچسب‌گذاری، ذخیره و به اشتراک گذاری تمام صفحات وب در یک محل را ارائه می‌دهد. در این مجموعه داده گره‌ها کاربران، یال‌ها ارتباطی بین آن‌ها و محتوا شامل برچسب‌های مشخصی است که در آدرس‌های ذخیره شده توسط هر کاربر وجود دارد.

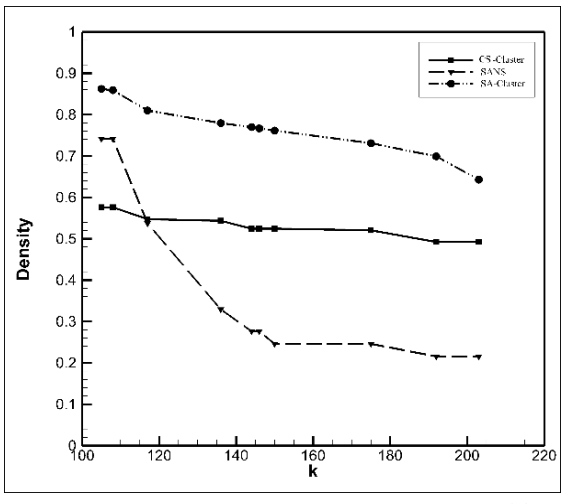
۴-۲. معیارهای بررسی شده

مقایسه‌های انجام شده بر اساس معیارهای مختلفی صورت پذیرفته که در ادامه تشریح می‌شوند.

تراکم: این معیار که خوشه‌بندی را از نظر ساختاری ارزیابی می‌کند، از طریق فرمول (۳) محاسبه می‌شود [۹]:

$$Density = \frac{\sum_{c \in C} E_c}{E} \qquad (۳)$$

که در آن، C مجموعه خوشه‌های حاصل از خوشه‌بندی، E_c یال‌های موجود در خوشه c و E مجموعه کل یال‌های گراف را نشان می‌دهد. شکل (۴) مقایسه تراکم خوشه‌های حاصل از خوشه‌بندی سه روش را نشان می‌دهد. با توجه به اینکه در روش SA-Cluster مراکز خوشه‌ها در نقاط متراکم گراف انتخاب می‌شود و گره‌های همسایه گره مرکزی به هر خوشه تخصیص داده می‌شود؛ درنهایت همان‌طور که در نمودار شکل (۴) ملاحظه می‌شود، خوشه‌هایی که با استفاده از روش SA-Cluster به‌دست می‌آیند دارای بیشترین تراکم و پس از آن روش پیشنهادی CS-Cluster تراکم بالاتری نسبت به روش SANS دارد.



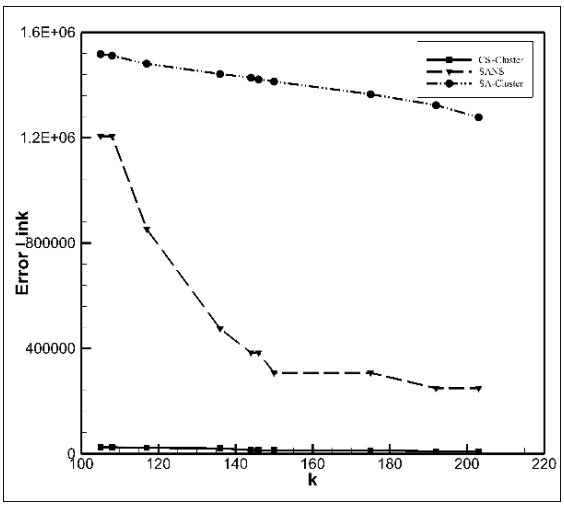
شکل ۴. مقایسه تراکم حاصل از روش‌های خوشه‌بندی

با توجه به این‌که هدف ما خوشه‌بندی گراف بر اساس محتوا و ساختار است و معیار تراکم تنها ساختار گراف را در نظر می‌گیرد، در نتیجه معیار مناسبی برای بررسی کیفیت خوشه‌های نهایی نیست. به عبارتی لزوماً گره‌های موجود در خوشه‌ای که تراکم بالایی دارد، از نظر محتوایی شباهت بالایی نخواهند داشت. در نتیجه بهتر است روش‌های خوشه‌بندی بر اساس معیاری مقایسه شوند که ساختار و محتوا را هم‌زمان در نظر می‌گیرد.

خطای یال: با توجه به اینکه از جنبه ساختاری، خوشه‌ای ایده آل است که تعداد یال‌های داخلی زیاد و تعداد یال‌های خروجی (بین خوشه‌ای) کم دارد و معیار تراکم تنها تعداد یال‌های داخلی را در نظر می‌گیرد، بنابراین حتی از لحاظ ساختاری معیار کاملی نبوده و همه جنبه‌ها را در نظر نمی‌گیرد. معیار خطای یال که در ادامه شرح داده می‌شود از لحاظ ساختاری کامل‌تر است، چون هم یال‌های داخلی و هم یال‌های خروجی را در نظر می‌گیرد. این معیار مجموع خطای یال خوشه‌های حاصل از عمل خوشه‌بندی را محاسبه می‌کند [۱۵]. این معیار نیز یک معیار ساختاری است و خوشه‌بندی را از نظر ساختاری مورد ارزیابی قرار می‌دهد.

در روش پیشنهادی CS-Cluster ، هنگام انتخاب مراکز خوشه‌ها و توسعه هر خوشه وضعیت اتصال گره‌ها در نظر گرفته می‌شود در نتیجه خوشه‌ها دارای کمترین خطای یال هستند. همان‌طور که در نمودار شکل (۵) مشاهده می‌شود، روش پیشنهادی CS-Cluster کمترین خطای یال را دارد و بعد از آن روش SANS خطای یال کمتری نسبت به SA-Cluster دارد.

در معیار خطای یال وزن یال‌ها در نظر گرفته نشده است؛ یعنی تأثیر وجود یک یال خارجی با وزن کم برابر تأثیر یک یال خارجی با وزن بسیار بالاست. در صورتی که تأثیر وجود این دو یال باید متفاوت باشد.



شکل ۵. مقایسه خطای یال روش‌های خوشه‌بندی

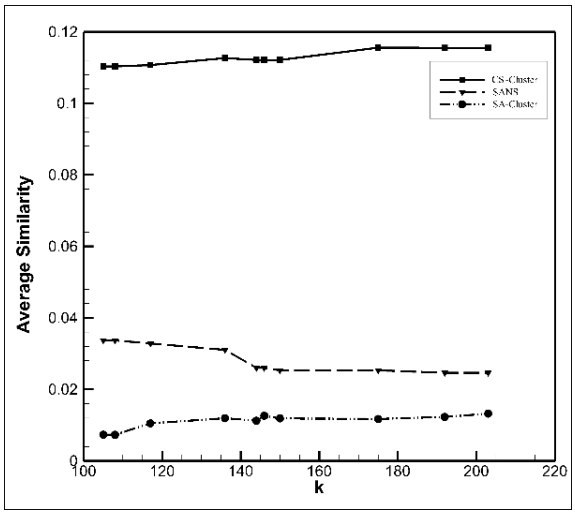
این معیار نیز فقط ساختار گراف را در نظر می‌گیرد، در نتیجه معیار مناسبی برای بررسی کیفیت خوشه‌های نهایی نیست و همان‌طور که گفته شد بهتر است روش‌های خوشه‌بندی را بر اساس معیاری مقایسه شود که هم ساختار و هم محتوا را در نظر می‌گیرد.

میانگین شباهت[1]: این معیار میانگین شباهت خوشه‌های گراف را مطابق با فرمول (۴) محاسبه می‌کند.

$$Avg\_Sim\_Graph = \frac{\sum_{c\in C} c_{avgsim}}{|C|} \qquad (۴)$$

که در آن، C خوشه‌های حاصل از خوشه‌بندی، c_{avgsim} میانگین شباهت درون خوشه‌ای خوشه c و |C| تعداد خوشه‌های حاصل از خوشه‌بندی را نشان می‌دهد. این معیار خوشه‌بندی را از نظر محتوایی مورد ارزیابی قرار می‌دهد.

همان‌طور که شرح داده شد، در روش CS-Cluster ابتدا گره های مرکزی خوشه‌ها با توجه به موقعیت ساختاری و محتوای آن‌ها انتخاب شده و سپس با اضافه کردن مشابه‌ترین گره‌های متصل به آن‌ها به هر خوشه، توسعه داده می‌شوند. در نتیجه تا جایی که ممکن است خوشه‌ها از نظر محتوایی میانگین شباهت بالایی خواهند داشت. با توجه به نمودار مقایسه‌ای شکل (۶) روش پیشنهادی CS-Cluster و پس از آن روش SA-Cluster بالاترین میانگین شباهت را به‌دست می‌آورند و روش SANS کمترین میانگین شباهت را دارد.



شکل ۶. میانگین شباهت حاصل از روش‌های خوشه‌بندی

معیار میانگین شباهت تنها شباهت محتوایی گراف را در نظر می‌گیرد، در نتیجه معیار مناسبی برای بررسی کیفیت خوشه‌های نهایی نیست و بهتر است روش‌های خوشه‌بندی را بر اساس معیاری که هم ساختار و هم محتوا را در نظر می‌گیرد مقایسه شود.

معیار ارزیابی[2]: همان‌طور که گفته شد در اکثر روش‌های خوشه‌بندی، عمل خوشه‌بندی بر اساس ساختار گراف انجام می‌شود و معیارهای ارزیابی جوامع مورد استفاده نیز اکثراً بر اساس ویژگی‌های ساختاری گراف هستند، از جمله این معیارهای ساختاری معیار تراکم[3]، پیمانگی[4]، میانگین درجه[5]، نسبت برش[6]، هدایت[7]، خطای یال[8] را می‌توان نام برد [۴]. برخی معیارهای کارایی نیز وجود دارد که محتوای گره‌ها را مدنظر دارند از جمله فراخوانی[9]، دقت[10] و صحت[11] که البته این دو معیار در مورد خوشه‌بندی‌هایی که برای هر خوشه یک توصیف مشخص و خاص مدنظر است کاربرد دارد [۱۷]. با این وجود، به طور کلی هیچ کدام از این معیارها برای ارزیابی خوشه‌بندی مورد نظر ما که دو هدف ساختار و محتوا را به طور هم‌زمان در نظر دارد، به تنهایی کاربرد مناسب و دقیقی نداشته و نمی‌توان از هیچ کدام از این معیارها برای ارزیابی الگوریتم خوشه‌بندی هدف استفاده کرد. در نتیجه به یک معیار ارزیابی دقیق‌تر و کامل‌تر نیاز است که خوشه‌بندی مورد نظر را از نظر ساختاری و محتوایی به طور هم‌زمان ارزیابی کند؛ بنابراین معیاری ارائه شده است که تعادلی بین ساختار و محتوای خوشه‌ها در نظر گرفته و روش‌های خوشه‌بندی را به طور دقیق‌تری ارزیابی کند.

با توجه به اینکه در این مقاله هدف به‌دست آوردن خوشه هایی است که در آن‌ها علاوه بر اینکه گره‌ها از نظر ساختاری دارای انسجام داخلی بالایی هستند، بالاترین شباهت محتوایی را نیز داشته باشند؛ بنابراین باید میانگین شباهت گره‌های درون هر خوشه حداکثر باشد. به طور کلی، یکی از مناسب‌ترین معیارهای ارزیابی ساختاری معیار خطای یال است، مزیت این معیار نسبت به معیارهای دیگر این است که معیار خطای یال فقط مبتنی بر شمارش خطاها است و همه اطلاعات مورد نیاز برای تعیین کیفیت یک خوشه را به طور محلی فراهم می‌کند. همچنین بر اساس این معیار جوامع بزرگ‌تر امتیاز بیشتری می‌گیرند که این مطابق اهداف سطح بالای روش ما است. خطای یال خوشه به صورت فرمول‌های (۵، ۶ و ۷) تعریف می‌شود [۱۵]:

$$\epsilon_{within}(C,E) = |\{(v,w)|v,w \in C \wedge v \neq w \wedge (v,w) \notin E\}| \qquad (۵)$$

$$\epsilon_{between}(C,E) = |\{(v,w) \in E|v \in C \wedge w \notin C\}| \qquad (۶)$$

$$\epsilon(\mathcal{C},G) = \sum_{C\in\mathcal{C}} \epsilon_{within}(C,E) + \frac{\epsilon_{between}(C,E)}{2} \qquad (۷)$$

[1] Average Similarity
[2] CS-Measure

[3] Density
[4] Modularity
[5] Average Degree
[6] Cut Size
[7] Conductance
[8] Error Link
[9] Recall
[10] Precision
[11] Accuracy

چون خطای یال‌های خارجی برای هر دو خوشه‌ای که شامل گره‌های مبدأ و مقصد یال هستند محاسبه می‌شود، خطای یال خارجی تقسیم بر ۲ می‌شود [۱۵]. معیار فوق فقط ویژگی‌های ساختاری گراف را در محاسبات خود در نظر می‌گیرد. به عبارتی هنگام محاسبه خطای یال خارجی به این نکته که یال مورد نظر بین دو گره با شباهت محتوایی بالا وجود دارد یا بین دو گره با شباهت محتوایی کم، توجهی نمی‌شود و تأثیر وجود این دو یال یکسان است. به طوری که مثلاً اگر بین دو خوشه یک یال با وزن بالا (میزان شباهت هر دو گره به عنوان وزن یال بین دو گره در نظر گرفته شده است) وجود داشته باشد، امتیاز خوشه‌بندی باید کمتر از زمانی باشد که بین دو خوشه یک یال با وزن خیلی کم وجود دارد. در واقع حالت دوم وضعیتی بهتر را نشان می‌دهد. در صورتی که در معیار خطای یال ذکر شده این مورد رعایت نشده است و تأثیر وجود یا عدم وجود دو یال با وزن‌های مختلف بین دو خوشه یکسان است؛ بنابراین با در نظر گرفتن مقدار شباهت گره‌ها و ترکیب این مقادیر شباهت با معیار خطای یال می‌توان به معیاری مناسب جهت ارزیابی خوشه‌بندی‌های ساختاری- محتوایی دست یافت.

سعی بر آن است که امتیاز خوشه بر اساس شباهت درون خوشه‌ای و خطای یال خوشه محاسبه شود. برای این کار وزن یال‌های درون خوشه‌ای و بین خوشه‌ای را متناسب با وضعیتی که دارند به‌هنگام رسانی می‌شود. سپس، میانگین وزن یال‌های هر خوشه محاسبه می‌شود. به همین منظور، چون حالت ایده‌آل به این صورت است که خوشه‌ها یال‌های خروجی نداشته باشند و یا در صورت وجود این یال‌ها وزن کمی داشته باشند، وزن یال‌های خروجی را به صورت منفی در نظر گرفته شده است $(-w)$، این عمل باعث می‌شود امتیاز کلی خوشه‌بندی متناسب با وزنی که یال خروجی دارد کم شود. اگر وزن یال‌های خروجی زیاد باشد این کاهش امتیاز زیاد خواهد بود و اگر وزن یال‌های خروجی کم باشد این کاهش امتیاز کمتر خواهد بود. وزن یال‌های ناموجود را برابر صفر در نظر گرفته شده است. سپس میانگین وزن یال‌ها (از جمله یال‌های موجود در خوشه، یال‌های ناموجود و یال‌های خروجی خوشه) را محاسبه کرده و به عنوان امتیاز خوشه در نظر گرفته می‌شود. پس از محاسبه امتیاز هر خوشه، مطابق با فرمول (۸ و ۹) میانگین امتیاز خوشه‌ها را به عنوان امتیاز کلی خوشه‌بندی محاسبه می‌شود.

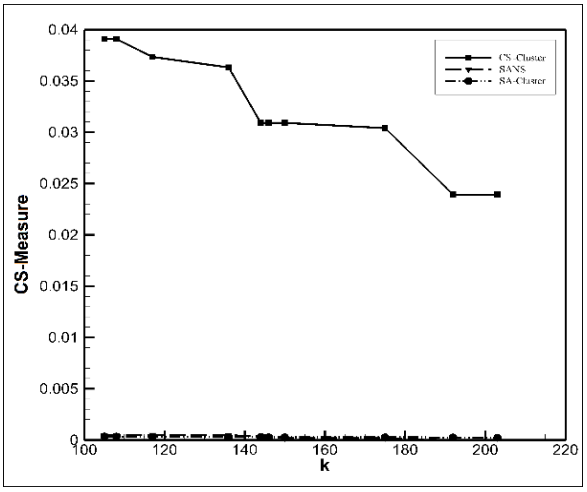
$$EScore(\mathcal{C}) = \frac{\sum_{e\in E_{in_{\mathcal{C}}}} w_e + \frac{\sum_{e\in E_{ex_{\mathcal{C}}}}(-w_e)}{2}}{\left|E_{in_{\mathcal{C}}}\right| + \left|E_{ex_{\mathcal{C}}}\right|} \qquad (۸)$$

$$SC - Similarity = \frac{\sum_{\mathcal{C}\in C} EScore(\mathcal{C})}{|C|} \qquad (۹)$$

$EScore(\mathcal{C})$ امتیاز جامعه $\mathcal{C}$، $E_{in\_\mathcal{C}}$ مجموعه یال‌های داخلی جامعه $\mathcal{C}$، $E_{ex\_\mathcal{C}}$ مجموعه یال‌های خروجی از جامعه $\mathcal{C}$، W_e وزن یال e، |C| تعداد خوشه‌های حاصل از خوشه‌بندی مورد نظر را نشان می‌دهد.

با توجه به اینکه بر خلاف سایر معیارهای موجود که فقط جنبه ساختاری یا محتوایی گراف را در نظر می‌گیرند و نمی‌توانند به طور مطلوبی عملکرد خوشه‌بندی‌های ساختاری- محتوایی را ارزیابی کنند، معیار CS-Measure هر دو جنبه ساختاری و محتوایی را با یکدیگر ترکیب می‌کند، در نتیجه استفاده از این معیار در ارزیابی عملکرد خوشه‌بندی‌های ساختاری- محتوایی دید مطلوب‌تر و روشن‌تری نسبت به عملکرد روش‌های خوشه بندی ساختاری- محتوایی را ارائه می‌دهد.

در روش CS-Cluster نیز هم در هنگام انتخاب مراکز خوشه ها و هم در هنگام توسعه خوشه‌ها ویژگی‌های ساختاری و محتوایی گره‌ها به طور هم‌زمان در نظر گرفته می‌شود. در نتیجه خوشه‌های حاصل از لحاظ ساختاری و محتوایی شباهت بالایی دارند و باعث می‌شود از نظر معیار CS-Measure مقادیر بالایی را به‌دست آورند. همان‌طور که در شکل (۷) مشاهده می‌شود روش پیشنهادی CS-Cluster بالاترین مقدار CS-Measure را به‌دست می‌آورد و روش SANS کمترین مقدار را دارد.



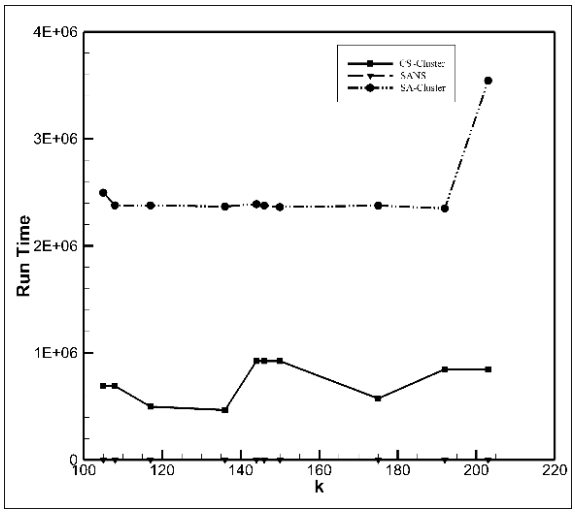
شکل ۷. مقایسه CS-Measure حاصل از روش‌های خوشه‌بندی

همان طور که قبلاً توضیح داده شد در این معیار، وزن یال های خارجی در (۱-) ضرب شده و در محاسبه میانگین شباهت در نظر گرفته می‌شود. در نتیجه تأثیر وجود یک یال خارجی با وزن بالا در کم کردن میانگین شباهت خوشه، خیلی بیشتر از وجود یک یال خارجی با وزن کم است. در نتیجه نسبت به معیار خطای یال دقیق‌تر عمل می‌کند. با توجه به اینکه این معیار، شباهت ساختاری و محتوایی را به طور هم‌زمان در نظر می‌گیرد، معیار مطلوب‌تری برای ارزیابی کیفیت خوشه‌های نهایی است؛ بنابراین با توجه به هدف خوشه بندی، می‌توان گفت روش

CS-Measure نسبت به سایر روش‌ها در مجموع بهتر عمل می‌کند.

زمان اجرا: با استفاده از این معیار مدت‌زمان خوشه‌بندی محاسبه می‌شود. در این مقاله آزمایش‌ها در یک سیستم رایانه‌ای با پردازنده ۳ هسته‌ای GH ۲/۲۶ و حافظه اصلی ۴ گیگابایتی انجام شده است. به منظور پیاده‌سازی الگوریتم‌ها از زبان جاوا در محیط اکلیپس استفاده شده است. با توجه به نمودار حاصل از آزمایش‌های انجام شده که در شکل (۸) مشاهده می‌شود؛ از بین روش‌های مقایسه شده روش SANS نسبت به بقیه روش‌ها سریع‌تر است، پس از آن روش CS-Measure با سرعت بیشتری عمل خوشه‌بندی را انجام می‌دهد. روش SA-Cluster نیز نسبت به بقیه کندتر است.

روش ارائه شده با پیاده‌سازی بر روی مجموعه داده با دو روش دیگر از جهت‌های مختلف ارزیابی و مقایسه شد که نتایج کلی این مقایسه‌ها در جدول (۲) آمده است.



شکل ۸. زمان اجرای روش‌های خوشه‌بندی

۵. نتیجه‌گیری

در بسیاری از کاربردهای دنیای واقعی در خوشه‌بندی گراف هم ساختار گراف و هم محتوای گره‌ها مهم هستند. به عنوان مثال، در یک شبکه اجتماعی، استفاده از محتوای پروفایل کاربران در کنار بهره‌گیری از ساختار شبکه ارتباطی آن‌ها می‌تواند کمک مؤثری در گروه‌بندی افراد شبکه‌های اجتماعی کند. با این وجود، بیشتر روش‌های خوشه‌بندی که تاکنون ارائه شده است فقط یک جنبه ساختاری یا محتوایی را در نظر گرفته‌اند و روش‌های خوشه‌بندی که هم ساختار و هم محتوا را در نظر بگیرند کمتر ارائه شده است؛ به همین دلیل، معیارهای ارزیابی روش‌های خوشه‌بندی نیز معیارهایی هستند که روش‌های خوشه‌بندی را فقط از نظر ساختاری ارزیابی می‌کنند. در این مقاله، با ترکیب جنبه‌های ساختاری و محتوایی گراف، برای خوشه‌بندی ساختاری- محتوایی گراف روش پیشنهادی CS-Cluster پیشنهاد شد. این روش که در واقع یک روش مبتنی بر انتشار برچسب است با ایجاد تعادلی مناسب بین جنبه‌های ساختاری و محتوایی، عمل خوشه‌بندی ساختاری- محتوایی را به خوبی انجام می‌دهد. علاوه بر این، در این روش از یک معیار مرکزیت ساختاری- محتوایی استفاده شده است که در انجام عمل خوشه‌بندی ساختاری- محتوایی گراف کمک بسیار مؤثری می‌کند. با توجه به ضعف معیارهای ارائه شده تاکنون در ارزیابی هم‌زمان محتوایی و ساختاری در این مقاله یک معیار ارزیابی جدید به نام CS-Measure که عمل خوشه‌بندی را با در نظر گرفتن هر دو جنبه ساختاری و محتوایی ارزیابی می‌کند نیز ارائه شد. این معیار خوشه‌ها را با دقت بیشتری و از نظر هر دو جنبه ساختاری و محتوایی ارزیابی می‌کند. همان‌طور که در آزمایش‌های انجام شده مشاهده می‌شود، خوشه‌های حاصل از روش پیشنهادی CS-Cluster نسبت به سایر روش‌ها، از نظر محتوایی و ساختاری کیفیت بهتری دارند و این تعادل به‌خوبی برقرارشده است. لازم به ذکر است استفاده از معیار ذاتاً جدید ساختاری و محتوایی باعث استخراج خصوصیت‌های جالبی از الگوریتم‌های خوشه‌بندی شد که قبل از آن توسط معیارهای منحصراً ساختاری قابل مشاهده نبود. به همین دلیل پیش‌بینی می‌شود در آینده معیارهای ارزیابی با خصوصیت‌های متنوع‌تری با در نظر گرفتن ویژگی‌های متفاوت داده‌های ساختاری و محتوایی به جامعه علمی در زمینه شبکه‌های اجتماعی ارائه شوند.

جدول ۲. مقایسه کلی روش‌های خوشه‌بندی بر اساس معیارهای ارزیابی شده در مقاله

معیار / رتبه	تراکم	میانگین شباهت	خطای یال	زمان اجرا	CS-Measure
۱	SA-Cluster	CS-Cluster	CS-Cluster	SANS	CS-Cluster
۲	CS-Cluster	SANS	SANS	CS-Cluster	SANS
۳	SANS	SA-Cluster	SA-Cluster	SA-Cluster	SA-Cluster

۶. منابع

[1] Aggarwal, C.; Wang, H. "Managing and Mining Graph Data"; Springer US, 2010.

[2] Patkar, S. B.; Narayanan, H. "An Efficient Practical Heuristic for Good Ratio-Cut Partitioning"; 16th Int. Conf. VLSI Design 2003, 1-6.

[3] Feldmann, A. E.; Foschini, L. "Balanced Partitions of Trees and Applications"; Algorithmica 2015, 71, 354-376.

[4] Newman, M. "Community Detection in Networks: Modularity Optimization and Maximum Likelihood are Equivalent"; Social and Information Networks 2016, 94, 1-8.

[5] Yang, Zh.; Algesheimer, R.; Tessone, C. J. "A Comparative Analysis of Community Detection Algorithms on Artificial Networks"; Scientific Reports 6, http://www.nature.com/articles/srep30750#supplementary-information, 2016.

[6] Fortunatoa, S.; Hricb, D. "Community Detection in Networks: A User Quide"; Phys. Rep. 2016, 659, 1-44.

[7] Khatoon, M.; Aisha Banu, W. "A Survey on Community Detection Methods in Social Networks"; Education and Management Engineering 2015, 1, 8-18.

[8] Elhadi, H.; Agam, G. "Structure and Attributes Community Detection: Comparative Analysis of Composite, Ensemble and Selection Methods"; SNAKDD '13 Proceedings of the 7th workshop on Social Network Mining and Analysis 2013, 1-7.

[9] Harenberg, S.; Bello, G.; Gjeltema, L.; Ranshous, S.; Harlalka, J.; Seay, R.; Padmanabhan, K.; Samatova, N. "Community Detection In Large-Scale Networks: A Survey and Empirical Evaluation"; Computational Statistics 2014, 6, 426-439.

[10] Matthew J. R.; Maier, M.; Jensen, D. "Graph Clustering with Network Structure Indices"; Proceedings of the 24th Int. Con. on Machine Learning 2007, 783-790.

[11] Shchukin, V.; Khristich, D.; Galinskaya, I. "Word Clustering Approach to Bilingual Document Alignment"; First Conf. on Machine Translation 2016, 2, 953-994.

[12] Weber, L. M.; Robinson, M. D. "Comparison of Clustering Methods for High-Dimensional Single-Cell Flow and Mass Cytometry Data"; Cold Spring Harbor Labs J. 2016, 047613.

[13] Zhou, Y.; Cheng, H; Xu Yu, J. "Graph Clustering Based on Structural/Attribute Similarities"; Proceeding of the VLDB Endowment 2009, 2, 718-729.

[14] Parimala, M.; Daphne, L. "Graph Clustering based on Structural Attribute Neighborhood Similarity (SANS)"; IEEE Int. Conf. on Electrical, Computer and Communication Technologies 2015, 1-5.

[15] Pool, S.; Bonchi, F.; Leeuwen, M. "Description-Driven Community Detection"; ACM Transactions on Intelligent Systems and Technology 2014, 5, 1-25.

[16] Qiao, M.; Qin, L.; Cheng, H.; Yu, J. X.; Tian, W. "Top-K Nearest Keyword Search on Large Graphs"; Proceding of the VLDB Endowment 2013, 10, 901-912.

[17] Wang, M.; Wang, Ch.; Xu Yu, J.; Zhang, J. "Community Detection in Social Networks: An In-depth Benchmarking Study with a Procedure-Oriented Framework"; Proceding of the VLDB Endowment 2015, 8, 998-1009.