ارائه روشی سلسله‌مراتبی جهت خوشه‌بندی ساختاری-محتوایی گراف

کبری رحمتی[۱]، کارشناسی ارشد؛ سامان کشوری[۲]، دانشجوی کارشناسی ارشد؛ حسن نادری[۳]، استادیار

۱- دانشکده کامپیوتر - دانشگاه علم و صنعت ایران - تهران - ایران- K_Rahmati@comp.iust.ac.ir
۲- دانشکده فناوری اطلاعات و ارتباطات - دانشگاه جامع امام حسین (ع) - تهران - ایران- SaKeshvari@ihu.ac.ir
۳- دانشکده کامپیوتر- دانشگاه علم و صنعت ایران- تهران- ایران- Naderi@iust.ac.ir

چکیده: موجودیت‌ها در شبکه‌های اجتماعی علاوه بر داشتن ارتباط با یکدیگر، دارای محتوا نیز هستند. این مدل از شبکه‌ها می‌توانند بر روی گراف‌هایی که گره‌های آن شامل متن هستند، مدل شوند. خوشه‌بندی گراف ازجمله مهم‌ترین کارهای تحلیلی شبکه اجتماعی است. باوجوداین دو جنبه، اغلب روش‌های خوشه‌بندی تنها یکی از جنبه‌های ساختاری یا محتوایی گراف را در نظر می‌گیرند. الگوریتم‌های خوشه‌بندی ساختاری-محتوایی، گراف را از هر دو جنبه ساختار و محتوا به‌صورت هم‌زمان در نظر می‌گیرند. هدف این مقاله رسیدن به خوشه‌هایی با ساختار درونی منسجم (ساختاری) و مقادیر ویژگی (محتوایی) همگن در گراف است. الگوریتم ارائه شده در این مقاله RLS-Cluster نام داشته که به‌صورت سلسله مراتبی با حذف یال با کمترین میانگین شباهت میان گره‌های محله آن یال، عمل خوشه‌بندی را انجام می‌دهد. در این روش برای هر یال میانگین شباهت محله محاسبه شده و به‌عنوان وزن آن یال در نظر گرفته می‌شود. یال‌هایی که دارای کم‌ترین وزن هستند حذف می‌شوند. این مرحله تا زمانی که به تعداد خوشه موردنظر کاربر برسد، ادامه میابد. مقایسه الگوریتم مطرح‌شده با سه الگوریتم خوشه‌بندی ساختاری-محتوایی ارائه شده تاکنون، بر اساس معیارهای مختلف سنجش کیفیت خوشه، بیانگر عملکرد مناسب روش ارائه شده است. این معیارها شامل معیارهای ساختاری، محتوایی و ساختاری-محتوایی هستند.

واژه‌های کلیدی: خوشه‌بندی، خوشه‌بندی ساختاری-محتوایی گراف، شبکه اطلاعاتی، شبکه اجتماعی.

A Hierarchical Method For Content-Structured Graph Clustering

K. Rahmati[1],Msc; S. Keshvari[2],Msc Student; H. Naderi[3], Assistant Professor

1- Iran University of Science and Technology, Tehran, Iran, Email: K_Rahmati@comp.iust.ac.ir
2- Emam Hossein comprehensive University, Tehran, Iran, Email: SaKeshvari@ihu.ac.ir
3- Iran University of Science and Technology, Tehran, Iran, Email: Naderi@iust.ac.ir

Abstract: Entities in social networks, in addition to having the relationship with each other, also have content. This type of networks can be modeled by the enriched graph, in which nodes could have text too. Graph clustering is one of the important attempts toward analyzing social networks. Despite these two facts, most of the existing graph clustering methods independently focused on one of the content or structural aspects. Content-Structural graph clustering algorithms simultaneously consider both the structure and the content of the graph. The main aim of this paper is to achieve well connected (structured) clusters while their nodes benefit from homogeneous attribute values (content). The proposed algorithm in this paper so-called RSL-Cluster performs the clustering by hierarchically removing the edge between nodes which has a weight lower that the average similarity of nodes. This stage continues until reaching the user's desired number of clusters. Comparing the proposed algorithm with three well-known content-structural clustering algorithms represents the proper functioning of the proposed method. The used measures to evaluate our method include structural, content and the content-structural measures.

Keywords: Clustering, Graph Content-Structural Clustering, Information Network, Social Network.

تاریخ ارسال مقاله: ۱۳۹۶/۰۳/۲۸
تاریخ اصلاح مقاله: ۱۳۹۶/۰۸/۲۳، ۱۳۹۶/۱۰/۰۹ و ۱۳۹۶/۱۰/۱۸
تاریخ پذیرش مقاله: ۱۳۹۶/۱۱/۲۱
نام نویسنده مسئول: حسن نادری
نشانی نویسنده مسئول: ایران - تهران – نارمک - دانشگاه علم و صنعت ایران - دانشکده مهندسی کامپیوتر

۱ -مقدمه

گراف کاربرد بسیار وسیعی در مباحث علمی دارد زیرا بسیاری از مسائل واقعی علمی جهت تحلیل بر روی گراف مدل می‌شوند [۱] [۲]. شبکه‌های اجتماعی نیز بر روی گراف مدل می‌شوند [۳] و ازآنجایی‌که استفاده از آن‌ها در جامعه در حال افزایش است، به‌دلیل افزایش حجم گراف تشکیل‌شده، پرداختن به خوشه‌بندی آن‌ها بیش‌ازپیش اهمیت دارد. محققان همواره سعی بر بهبود الگوریتم‌های خوشه‌بندی به‌منظور افزایش کیفیت خوشه‌های به‌دست‌آمده دارند [۴]. خوشه‌بندی گراف شامل تقسیم گره‌های گراف به گروه‌هایی با گره‌های مشابه است. شباهت بین گره‌ها معمولاً توسط یک تابع هدف ریاضی تعریف می‌شود [۵] خوشه‌بندی مسئله‌ای مهم در تحلیل داده‌های گراف و یک روش دسته‌بندی بدون ناظر برای داده‌های گراف است. حل مسئله خوشه‌بندی گراف که در رده مسائل NP-Hard قرار می‌گیرد، به‌صورت کلی از طریق روش‌های مکاشفه‌ای[۱] و تقریبی حاصل می‌شود. این توابع هدف در روش‌های مکاشفه‌ای به دو شیوه محلی و سراسری تعریف می‌شوند [۶][۷].

در بسیاری از کاربردها ازجمله خوشه‌بندی شبکه علمی مقالات، به‌منظور تحلیل و ارزیابی مقالات بهتر است ضمن توجه به محتوای موجود در مقالات، ارتباطات برقرارشده بین آن‌ها که از طریق مرجع دهی برقرار می‌شود، به‌صورت گراف مدل شوند. به‌عنوان‌مثال علاوه بر این‌که ارتباطات بین گره‌ها دارای ساختار مشخصی بوده (مقالات به‌عنوان رأس و ارتباطات به‌صورت یال در گراف تشکیل‌شده)، گره‌ها دارای محتوا نیز هستند، لذا 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 مقالاتی که محتوای خاصی را تولید می‌کند، اما در مراجع خود، به موضوعی دیگر نیز ارجاع می‌دهد با مقاله‌ای که محتوای موردنظر را تولید نمی‌کند در این خوشه‌بندی یکسان در نظر گرفته می‌شود. توجه صرفاً به محتوا نیز اثربخش نیست زیرا در این حالت، ارتباطات در نظر گرفته نشده و یک مقاله عادی با توجه به محتوای موجود در آن -که با معیارهای محتوایی تعریف‌شده مطابقت دارد- در خوشه مقالات مهم قرار گیرد.

اغلب روش‌های خوشه‌بندی گراف که تاکنون ارائه‌شده 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 ارائه‌شده است. همین امر باعث شده است که معیارهای موجود برای ارزیابی خوشه‌ها نیز اغلب بر اساس ویژگی‌های ساختاری یا محتوایی گراف باشند. درحالی‌که هیچ‌کدام از این معیارها به‌تنهایی برای ارزیابی هم‌زمان ویژگی‌های ساختاری و محتوایی کاربرد مناسب و دقیقی ندارند؛ بنابراین روش‌های خوشه‌بندی ساختاری-محتوایی باید بتوانند میان ساختار و محتوا به‌گونه‌ای تعادل برقرار کنند که الگوریتم به‌لحاظ معیارهای ساختاری و محتوایی عملکرد مناسبی از خود نشان دهد. ضمن این‌که الگوریتم باید برمبنای تنها معیار ارزیابی ساختاری-محتوایی ارائه شده [۸] نیز به‌خوبی عمل کند.

هدف کلی، تشخیص خوشه‌هایی از گراف است که گره‌های درونی هر خوشه به‌طور متراکم به یکدیگر مرتبط بوده و به‌لحاظ معنایی شبیه به هم باشند و ارتباط ساختاری و معنایی بین خوشه‌های مجزا به حداقل خود برسد. در این مقاله یک روش خوشه‌بندی سلسله‌مراتبی ارائه‌شده است که ضمن در نظر گرفتن تعادل بین ساختار و محتوا، با حذف ارتباط میان دو گره‌ای که گره‌های محله آن کمترین شباهت را به یکدیگر دارند، عمل خوشه‌بندی ساختاری- محتوایی را انجام دهد. نتایج آزمایش‌های تجربی نشان می‌دهد که روش پیشنهادشده به‌نحو مطلوبی عمل خوشه‌بندی با ویژگی‌های موردنظر را انجام می‌دهد.

در ادامه مقاله، در بخش ۲، کارهای انجام‌شده درزمینه خوشه‌بندی گراف به‌صورت مختصر مطرح‌شده است. در بخش ۳، روش تحقیق و روش ارائه شده در این مقاله شرح داده‌شده است. در بخش ۴، با پیاده‌سازی روش‌های خوشه‌بندی بر روی مجموعه داده، عملکرد روش پیشنهادی با سایر روش‌های خوشه‌بندی ساختاری-محتوایی ارزیابی و مقایسه شده است. در پایان و در بخش ۵، نتیجه‌گیری مقاله ارائه می‌شود.

۲ -مرور کارهای انجام‌شده

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 خوشه‌بندی تنها جنبه ساختاری گراف را در نظر می‌گیرند [۹][۱۰][۱۱][۱۲] که از این دسته می‌توان TopGC[۱۳] و Louvain[۱۴] را نام برد. برخی روش‌ها نیز مطرح‌شده‌اند که عمل خوشه‌بندی را بر اساس محتویات گره‌ها انجام می‌دهند [۱۵]، ازجمله این روش‌ها می‌توان k-Means[۱۶] و k-Medoids[۱۷] و Chameleon[۱۸] را نام برد. درحالی‌که در بسیاری از کاربردهای دنیای واقعی هر دو جنبه ساختار و محتوا در کنار یکدیگر و به‌طور هم‌زمان موردنظر هستند [۱۹][۲۰][۲۱]، پس بهتر است در خوشه‌بندی شباهت ساختاری و محتوایی گره‌ها در کنار یکدیگر در نظر گرفته شود. تاکنون تعداد کمی الگوریتم خوشه‌بندی که به‌صورت ساختاری-محتوایی به عمل خوشه‌بندی می‌پردازند ارائه شده است که برخی از آن‌ها عبارت‌اند از:

SA-Cluster[۲]: در این الگوریتم به‌ازای 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 تکرارشونده است که در آن ابتدا مرکزی‌ترین گره‌ها (گره‌ها با تراکم بالا) به‌عنوان مرکز خوشه انتخاب شده و سپس باقی گره‌ها به نزدیک‌ترین مراکز، تخصیص داده می‌شوند. سپس در هر خوشه مراکز به هنگام رسانی می‌شوند (همچنین وزن یال‌ها که بر اساس ویژگی‌های محتوایی و ساختاری

تنظیم‌شده‌اند و سایر پارامترهای الگوریتم به‌روز می‌شود) و عملیات تا همگرا شدن تکرار می‌شود [۱۹]. ازآنجاکه این الگوریتم به‌صورت تکرارشونده است، زمان اجرای بالایی دارد.

SANS[۳]: در این الگوریتم، ابتدا در گراف وزن‌دار ورودی شاخص وزن هر گره -که به‌صورت مجموع وزن یال‌های متصل به گره است- محاسبه‌شده و گره‌ای که بیشترین شاخص وزن را دارد به‌عنوان مرکز[۴] 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) بیش‌ازحد آستانه دارند نیز به خوشه اضافه می‌شوند. پس از طی این مراحل، مجدداً از بین گره‌های باقیمانده گره‌ای که بیشترین شاخص وزن را دارد به‌عنوان مرکز بعد انتخاب می‌شود و عملیات تا خوشه‌بندی همه گره‌ها تکرار می‌شود [۲۰]. زمان اجرای این الگوریتم نسبت به SA-Cluster بهتر است و نسبت به الگوریتم SA-Cluster خوشه‌های متراکم‌تری به‌دست می‌آورد.

DCM[۵]: این الگوریتم از مجموعه‌ای از جوامع کاندید شروع کرده و دو گام اصلی را به‌طور متناوب تا رسیدن به همگرایی تکرار می‌کند. در گام اول جوامع با بهترین امتیاز خوشه (ازنظر 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 خوشه را مقداری تغییر داده و گام‌های الگوریتم را تا همگرا شدن تکرار کند [۲۱]. این روش لزوماً همه گره‌ها را پوشش نمی‌دهد و برای مجموعه داده‌های کوچک کند و برای مجموعه داده‌های بزرگ سرعت‌بالا دارد.

CS-Cluster[۶]: در این الگوریتم، ابتدا گراف وزن‌دار ورودی -که در آن وزن بیانگر شباهت محتوایی است- به یک گراف وزن‌دار که وزن آن بیانگر فاصله ساختاری و یال آن نشان‌دهنده شباهت محتوایی دو گره است تبدیل می‌شود. سپس گره‌های مرکزی به کمک فرمول (۱) به دست می‌آیند.

$$C_v = \frac{(D_v)^2}{\sum_{e \in E_v} w_e} \qquad (۱)$$

که در آن C_v امتیاز مرکزیت گره v، E_v مجموعه یال‌های متصل به v در گراف دوم، w_e وزن یال e در گراف دوم و Dv درجه گره v در گراف دوم است.پس از انتخاب گره‌های مرکزی و در نظر گرفتن آن‌ها به‌عنوان مرکز خوشه، بر اساس فاصله موجود میان این گره‌ها خوشه‌ها تا زمانی که تمامی گره‌ها پوشش داده شوند گسترش میابند [۸]. نتایج حاصل از این پژوهش نشان می‌دهد ازآنجاکه در طول انجام مراحل انتخاب گره‌های مرکزی تا گسترش خوشه در این روش هم‌زمان جنبه ساختار و محتوا دیده می‌شود، این روش به لحاظ ساختاری-محتوایی دارای عملکرد بهتری نسبت به سایر روش‌ها است.

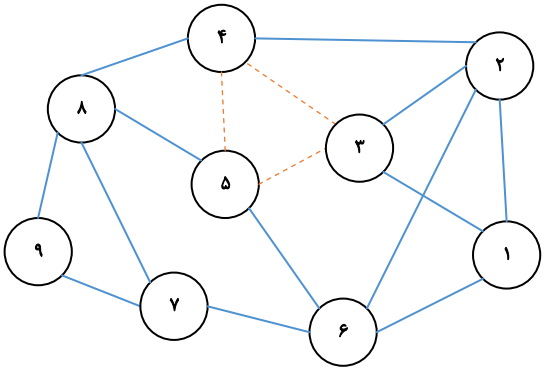
از نقاط ضعف‌های این روش می‌توان به‌سرعت نامناسب، تراکم[۷] و پیمانگی[۸] به نسبت پایین اشاره کرد. به این منظور در ادامه سعی شده روشی ارائه شود که عملکرد مناسبی از خود نشان دهد. در این روش سعی می‌شود با حذف یال‌هایی با میانگین شباهت محتوایی کم، زیر گراف‌هایی جدا از هم تشکیل داد که درنهایت به‌عنوان خوشه‌ها در نظر گرفته می‌شوند. ایده این روش این است که میانگین شباهت کم به‌احتمال‌زیاد در محله‌هایی با تراکم یال کمتر وجود خواهد داشت درنتیجه امتیاز کم به یال‌هایی تعلق خواهد گرفت که در محله‌ای با میانگین شباهت کم (محتوایی) و با تراکم یال کمتر (ساختاری) قرار دارند. درنتیجه با حذف کردن این یال‌ها به‌طور هم‌زمان، هر دو جنبه ساختار و محتوا به‌خوبی در نظر گرفته شده و در رسیدن به نتایج مطلوب کمک خواهد کرد.

۳ -روش پیشنهادی

یک گراف ویژگی را به‌صورت G=(V,E,A) نشان داده می‌شود، به‌طوری‌که 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 هرکدام از خوشه‌ها علاوه بر نزدیکی ساختاری، ازلحاظ محتوایی نیز مشابه باشند. در ادامه روش ارائه‌شده تشریح می‌شود.

۳ ـ۱ -روش ارائه شده (RSL-Cluster[۹])

همان‌طور که گفته شد برای خوشه‌بندی، گراف وزن‌داری در دسترس است که وزن هر یال بیانگر شباهت محتوایی دو رأس و همسایه‌های مشترک آن دو رأس است. در این روش در هر مرحله یالی که ارزش کمتری دارد حذف می‌شود تا زمانی که تعداد خوشه‌ها به تعداد مدنظر برسد. اولویت حذف با یالی است که بر اساس ترکیبی از ویژگی‌های ساختاری و محتوایی در مکانی با میانگین شباهت محله کم قرار دارد. درواقع یالی برای حذف شدن مناسب‌تر است که در زیر گرافی (محله‌ای) با میانگین شباهت محتوایی کم‌تر قرار دارد، زیرا میانگین شباهت کم به‌احتمال‌زیاد در محله‌هایی با تراکم یال کمتر وجود خواهد داشت درنتیجه امتیاز کم به یال‌هایی تعلق خواهد گرفت که در محله‌ای با میانگین شباهت کم و با تراکم یال کمتر قرار دارند. به همین منظور، امتیاز یال بین هر دو گره برابر با میانگین شباهت محتوایی زیر گراف شامل دو گره و همسایه‌های مشترک آن دو گره است؛ که در ادامه امتیاز هر یال به‌عنوان وزن آن یال نشان داده می‌شود. درواقع وزن یال برابر با حاصل جمع شباهت محتوایی گره‌های موجود در زیر گراف موردنظر، تقسیم‌بر کل تعداد یال‌های ممکن در زیر گراف می‌شود.



شکل ۱: نمونه زیر گراف برای محاسبه شباهت دو گره ۳ و ۵

به‌عنوان‌مثال، در گراف شکل (۱) بـرای محاسـبه وزن یـال (۳،۵) مجموع شباهت محتوایی دو گره ۳ و ۵ و همچنـین گـره ۴ (بـه علـت این‌که گره ۴ همسایه مشترک دو گره ۳ و ۵ است) محاسبه شده و بـر کل تعداد یال‌های ممکن بین این سه گره تقسیم می‌شود. به‌طـورکلی، وزن هر یال با استفاده از فرمول (۲) برای کلیه یال‌های گراف محاسـبه می‌شود.

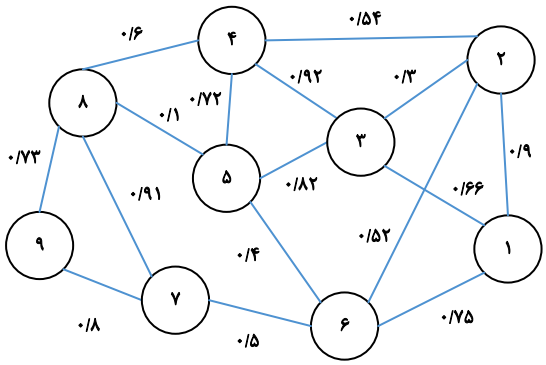
$$W(e_{ij}) = \frac{\sum_{a,b\in N_v} Sim\ (a,b)}{|N_e|} \qquad (۲)$$

که در آن، $W(e_{ij})$ وزن یال بین گره‌های i و j، N_v مجموعه گره‌های شامل دو گره i و j و همسایه‌های مشترک آن‌ها، N_e حداکثر تعداد یال‌های ممکن بین گره‌های موجود در مجموعه N_v و $Sim(a,b)$ شباهت محتوایی دو گره a و b است.

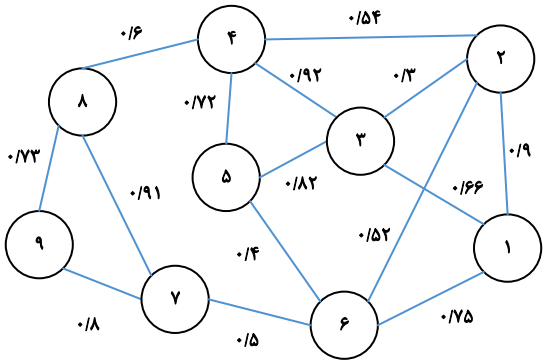
به‌طورکلی، در این روش به ترتیب کم‌وزن‌ترین یال‌ها تـا رسـیدن به تعداد خوشه موردنظر حذف می‌شوند. با توجه به این‌که در این روش ابتدا کم‌وزن‌ترین یال‌ها برای حذف شدن انتخـاب مـی‌شـوند، بنـابراین درنهایت در هر خوشه یال‌هایی که بیشترین وزن را دارند باقی خواهند ماند. درنتیجه خوشه‌هایی که گره‌های درون آن میانگین شباهت محله بالایی دارند، باقی خواهند ماند. به‌عنوان‌مثال، در گراف G ابتدا وزن هر یال با استفاده از رابطه (۲) محاسبه می‌شود. بـا فـرض ایـن‌کـه تعـداد خوشه موردنظر ۳ است، مراحل خوشه‌بندی با مثال گراف ساخته‌شـده در شکل (۲) تشریح می‌شود.

برای مثال ارائه شـده در اولـین مرحلـه، چـون یـال (۵،۸) دارای کم‌ترین وزن است برای حذف شدن انتخاب می‌شود.

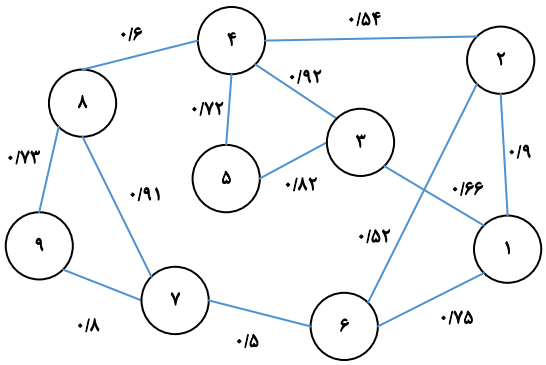
در مرحله بعد، یال (۳،۲) حذف می‌شـود؛ و پـس‌ازآن یـال (۵،۶) حذف خواهد شد لازم به توضیح است که پس از حذف هر یال، تعـداد خوشه‌های تشکیل شده با تعداد موردنظر بررسی می‌شود درصورتی‌کـه تعداد کمتر از آن مقدار باشد، خوشه‌بنـدی ادامـه میابـد، در غیـر ایـن صورت خوشه‌بندی خاتمه میابد. برای مثـال مـوردنظر در ایـن مرحلـه مطابق با شکل (۵) بـه ترتیـب یـال‌هـای (۶،۷)، (۶،۲) و (۲،۴) حـذف می‌شوند، سپس با حذف شدن یال (۴،۸) گراف بـه دو خوشـه تبـدیل خواهد شد زیرا گراف به دو زیر گراف تبدیل می‌شود.



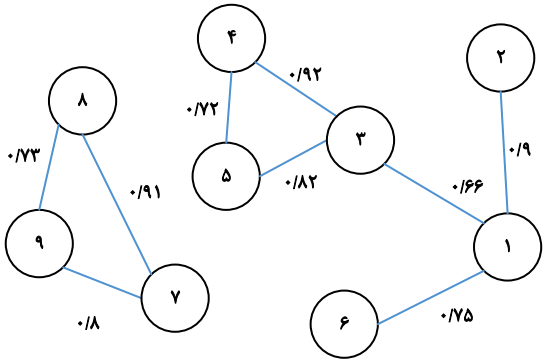
شکل ۲: نمونه گراف اولیه وزن‌دار



شکل ۳: مرحله اول خوشه‌بندی (حذف یال (۵،۸))



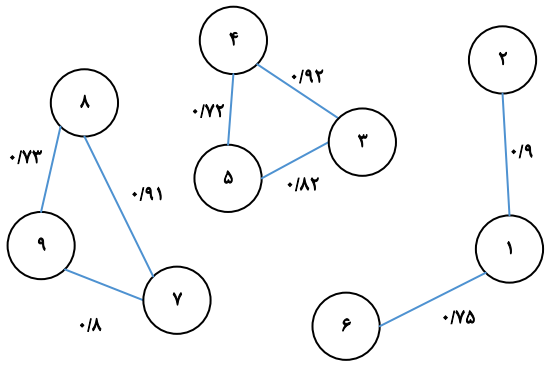
شکل ۴: مرحله دوم و سوم خوشه‌بندی مثال



شکل ۵: تبدیل گراف به دو خوشه

همان‌طور که گفته شد خوشه‌بندی برای مثال موردنظر تا زمـانی که گراف به ۳ خوشه مجزا تبدیل شود، ادامه خواهد یافت. در ادامه بـا

حذف یال (۱،۳) همان‌طور که در شکل (۶) مشاهده می‌شود خوشه‌بندی خاتمه میابد. در ادامه با اجرای روش خوشه‌بندی ارائه شده بر روی مجموعه داده، در شرایط یکسان با سایر روش‌های خوشه‌بندی ساختاری-محتوایی ارائه شده، کارایی آن به‌لحاظ معیارهای مختلف مورد ارزیابی قرارگرفته است.



شکل ۶: تبدیل گراف به سه خوشه و پایان خوشه‌بندی

۴ -ارزیابی

جهت ارزیابی روش ارائه شده و مقایسه آن با سایر روش‌های خوشه‌بندی ساختاری-محتوایی آزمایش‌های مختلفی بر اساس معیارهای مختلف انجام شده است. این آزمایش‌ها در یک سیستم کامپیوتری با پردازنده ۳ هسته‌ای GH۲,۲۶ و حافظه اصلی ۴ گیگابایتی انجام‌شده است. به‌منظور پیاده‌سازی الگوریتم‌ها از زبان جاوا در محیط اکلیپس[۱۰] استفاده شده است. در ادامه، نتایج مقایسه آزمایش‌های انجام‌شده بر روی چهار روش SANS، SA-Cluster، CS-Cluster و RLS-Cluster آمده است. ازآنجاکه خوشه‌بندی DCM غیر همپوشان بوده و تمامی گره‌ها را پوشش نمی‌دهد، در مقایسه در نظر گرفته نشده است. در این آزمایش‌ها شباهت جاکارد[۱۱] بین گره‌ها مطابق با فرمول (۳) محاسبه‌شده و به‌عنوان وزن یال‌ها در نظر گرفته می‌شود [۲۲].

$$J(C,C') = \frac{|C \cap C'|}{|C \cup C'|} \qquad (۳)$$

فرمول (۳) شباهت جاکارد دو مجموعه C و 'C را نشان می‌دهد که برابر با تعداد ویژگی مشترک (شبیه) دو مجموعه به روی کل ویژگی‌های آن است. روش‌های مورد مقایسه دارای پارامترهای متفاوتی هستند. برای مقایسه این روش‌ها محدودیت‌هایی وجود دارد. به‌عنوان‌مثال درروش SA-Cluster و روش RLS-Cluster تعداد خوشه‌ها از ابتدا مشخص می‌شود، ولی در روش 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.

۴ ۱ -مجموعه داده

در آزمایش‌های انجام‌شده از مجموعه داده Delicious استفاده شده است که در کارگاه HetRec 2011 در دسترس است[۱۲]. این مجموعه داده دارای ۱۸۶۱ گره و ۷۶۶۴ یال است. خصوصیت‌های هر گره -که تعداد آن‌ها ۱۳۵۰ عدد است- نیز به‌وسیله یک آرایه دودویی مشخص‌شده است. خصوصیات با استفاده از مقادیر ۰ و ۱ در آرایه مربوط به هر گره مشخص‌شده که برای تعیین وجود یا عدم وجود خصوصیت در یک گره استفاده می‌شوند. درواقع برای هر گره یک آرایه ۱۳۵۰ خانه‌ای وجود دارد که برای هر ویژگی یک خانه در نظر گرفته شده است. این کار باعث می‌شود که خصوصیات گره‌ها به‌صورت دودویی ذخیره‌شده تا هنگام محاسبه شباهت، بتوان از شباهت جاکارد به‌راحتی استفاده نمود. لازم به توضیح است که Delicius یک سرویس خدمات Bookmark است که امکان برچسب‌گذاری، ذخیره و به اشتراک‌گذاری 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 ذخیره‌شده توسط هر کاربر وجود دارد. در این مجموعه داده هر تاپل شامل شماره کاربر، Bookmark و برچسب‌های آن است. این ۱۳۵۰ ویژگی، برچسب‌های مختلفی است که هر شخص دارد، ازجمله آن‌ها نویسنده بودن یا معلم بودن را می‌توان نام برد.

۴ ۲ -نتایج آزمایش

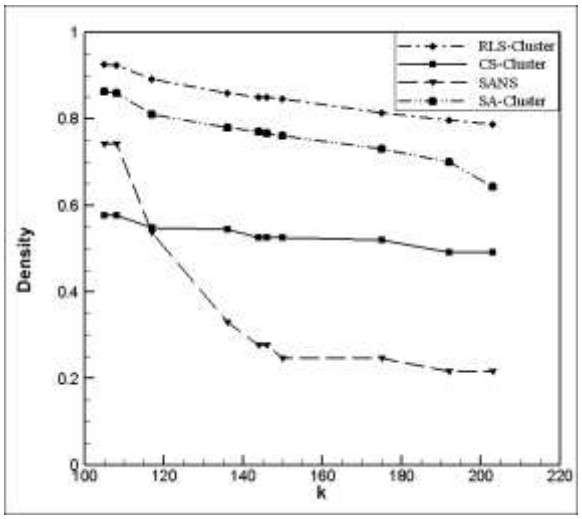
مقایسه‌های انجام شده بر اساس معیارهای مختلفی صورت پذیرفته که در ادامه تشریح می‌شوند.

۴ ۲ ۱ -تراکم

این معیار که خوشه‌بندی را ازنظر ساختاری ارزیابی می‌کند، از طریق فرمول (۴) محاسبه می‌شود [۱۴]:

$$Density = \frac{\sum_{c \in C} E_c}{E} \qquad (۴)$$

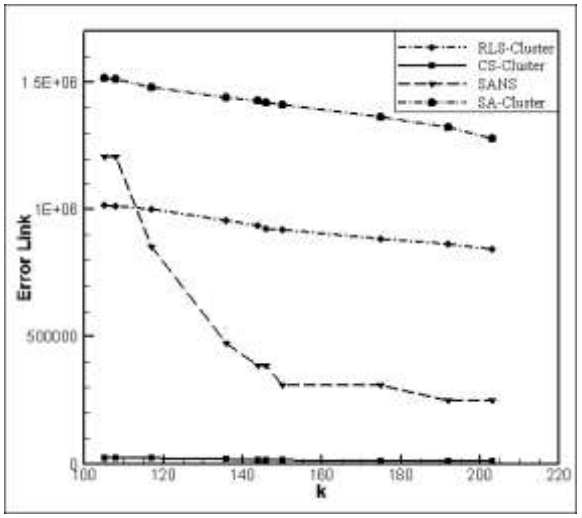
که در آن، C مجموعه خوشه‌های حاصل از خوشه‌بندی، E_c یال‌های موجود در خوشه c و E مجموعه کل یال‌های گراف را نشان می‌دهد. شکل (۷) مقایسه تراکم خوشه‌های حاصل از خوشه‌بندی سه روش را نشان می‌دهد. با توجه به این‌که در راهبرد حذف یال با کمترین میانگین شباهت محله سعی می‌شود یال‌هایی حذف شوند که در محل‌های با میانگین شباهت کم و تراکم کم قرار دارند. درنتیجه همان‌طور که در شکل (۷) مشاهده می‌شود، درنهایت خوشه‌هایی که از این روش به‌دست می‌آید، بیشترین تراکم را دارند. در روش SA-Cluster نیز مراکز خوشه‌ها در نقاط متراکم گراف انتخاب می‌شود و گره‌های همسایه گره مرکزی به هر خوشه تخصیص داده می‌شود، به همین دلیل این روش نیز تراکم بالایی دارد، پس‌ازاین دو روش، روش CS-Cluster تراکم بالاتری نسبت به روش SANS دارد.



شکل ۷: مقایسه تراکم حاصل از روش‌های خوشه‌بندی

۴ ۲ ۴ -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 ازلحاظ ساختاری معیار کاملی نبوده و همه جنبه‌ها را در نظر نمی‌گیرد. معیار خطای یال که در ادامه شرح داده می‌شود ازلحاظ 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 ازنظر ساختاری مورد ارزیابی قرار می‌دهد.



شکل ۸: مقایسه خطای یال روش‌های خوشه‌بندی

در روش CS-Cluster، هنگام انتخاب مراکز خوشه‌ها و توسعه هر خوشه وضعیت اتصال گره‌ها در نظر گرفته می‌شود درنتیجه خوشه‌ها دارای کمترین خطای یال هستند. این در حالی است که در روش RLS-Cluster یال‌های موجود حذف‌شده و درنتیجه در این معیار تأثیرگذار است. همان‌طور که در نمودار شکل (۸) مشاهده می‌شود، روش پیشنهادی CS-Cluster کمترین خطای یال را دارد و بعدازآن روش SANS و RLS-Cluster خطای یال کمتری نسبت به SA-Cluster دارند.

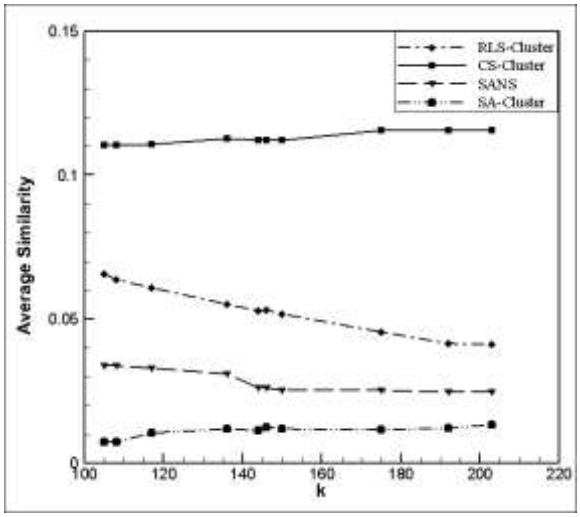
۴ ۲ ۳ -میانگین شباهت[۱۳]

این معیار میانگین شباهت خوشه‌های گراف را مطابق با فرمول (۵) محاسبه می‌کند.

$$Avg\_Sim\_Graph = \frac{\sum_{c \in C} c_{avgsim}}{|C|} \quad (۵)$$

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 ازنظر محتوایی مورد ارزیابی قرار می‌دهد.

درروش CS-Cluster ابتدا گره‌های 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. درنتیجه تا جایی که ممکن است خوشه‌ها ازنظر محتوایی میانگین شباهت بالایی خواهند داشت. درروش RLS-Cluster نیز ازآنجاکه گره‌هایی با کمترین شباهت محلی حذف می‌شوند خوشه‌های باقی‌مانده دارای شباهت قابل قبولی با یکدیگر هستند.



شکل ۹: مقایسه میانگین شباهت روش‌های خوشه‌بندی

با توجه به نمودار مقایسه‌ای شکل (۹) روش CS-Cluster و پس‌ازآن روش پیشنهادی RLS-Cluster بالاترین میانگین شباهت را داشته و روش‌های SA-Cluster و SANS کمترین میانگین شباهت را دارند.

۴ ۲ ۴ -معیار ارزیابی CS-Measure

در اکثر روش‌های خوشه‌بندی، عمل خوشه‌بندی بر اساس ساختار گراف انجام می‌شود و معیارهای ارزیابی جوامع مورداستفاده نیز اکثراً بر اساس ویژگی‌های ساختاری گراف هستند، ازجمله این معیارهای

ساختاری معیار تراکم[۱۴]، پیمانگی[۱۵]، میانگین درجه[۱۶]، نسبت برش[۱۷]، هدایت[۱۸]، خطای یال[۱۹] را می‌توان نام برد [۸]. برخی معیارهای کارایی نیز وجود دارد که محتوای گره‌ها را مدنظر دارند ازجمله فراخوانی[۲۰]، دقت[۲۱] و صحت که البته این دو معیار در مورد خوشه‌بندی‌هایی که برای هر خوشه یک توصیف مشخص و خاص مدنظر است کاربرد دارد [۲۳]. معیار CS-Measure هر دو جنبه ساختاری و محتوایی را با یکدیگر ترکیب می‌کند. خطای یال خوشه به‌صورت فرمول‌های (۶)، (۷) و (۸) تعریف می‌شود [۲۱]:

$$\epsilon_{within}(C,E) = |\{(v,w)|v,w \in C \wedge v \neq w \wedge (v,w) \notin E\}| \quad (۶)$$

$$\epsilon_{between}(C,E) = |\{(v,w) \in E|v \in C \wedge w \notin C\}| \quad (۷)$$

$$\epsilon(\mathcal{C},G) = \sum_{C\in\mathcal{C}} \epsilon_{within}(C,E) + \frac{\epsilon_{between}(C,E)}{2} \quad (۸)$$

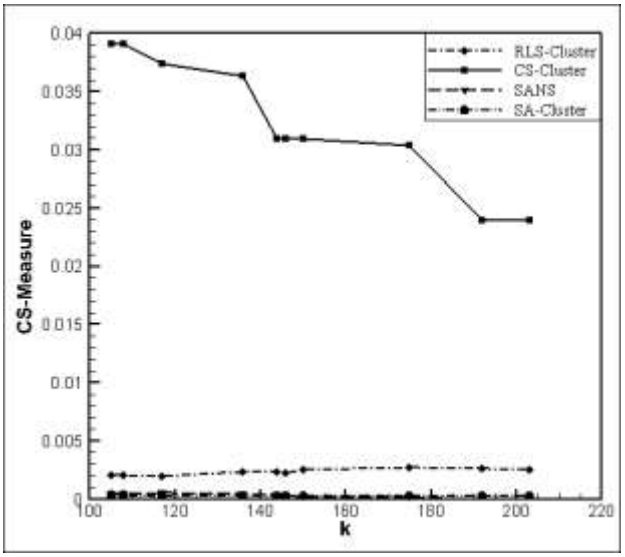
چون خطای یال‌های خارجی برای هر دو خوشه‌ای که شامل گره‌های مبدأ و مقصد یال هستند محاسبه می‌شود، خطای یال خارجی تقسیم‌بر ۲ می‌شود [۲۱]. معیار فوق فقط ویژگی‌های ساختاری گراف را در محاسبات خود در نظر می‌گیرد. به عبارتی هنگام محاسبه خطای یال خارجی به این نکته که یال موردنظر بین دو گره با شباهت محتوایی بالا وجود دارد یا بین دو گره با شباهت محتوایی کم، توجهی نمی‌شود و تأثیر وجود این دو یال یکسان است. به‌طوری‌که مثلاً اگر بین دو خوشه یک یال با وزن بالا (میزان شباهت هر دو گره به‌عنوان وزن یال بین دو گره در نظر گرفته‌شده است) وجود داشته باشد، امتیاز خوشه‌بندی باید کمتر از زمانی باشد که بین دو خوشه یک یال با وزن خیلی کم وجود دارد. درواقع حالت دوم وضعیتی بهتر را نشان می‌دهد. درصورتی‌که در معیار خطای یال ذکرشده 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 دست‌یافت.

در معیار CS-Measure امتیاز خوشه بر اساس شباهت درون خوشه‌ای و خطای یال خوشه محاسبه می‌شود. برای این کار در این معیار وزن یال‌های درون خوشه‌ای و بین خوشه‌ای، متناسب با وضعیتی که دارند به‌روز می‌شود. سپس، میانگین وزن یال‌های هر خوشه محاسبه می‌شود. به همین منظور، چون حالت ایده آل به این صورت است که خوشه‌ها یال‌های خروجی نداشته باشند و یا در صورت وجود این یال‌ها وزن کمی داشته باشند، وزن یال‌های خروجی به‌صورت منفی در نظر گرفته‌شده ($-w$) که این عمل باعث می‌شود در صورت وجود یال خروجی در خوشه، امتیاز کلی خوشه‌بندی متناسب با وزنی که یال خروجی دارد کاهش یابد. 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. همچنین وزن یال‌های ناموجود برابر صفر در نظر گرفته می‌شود. در ادامه میانگین وزن یال‌ها (ازجمله یال‌های موجود در خوشه، یال‌های ناموجود و یال‌های خروجی خوشه) محاسبه‌شده و به‌عنوان امتیاز خوشه در نظر گرفته می‌شود. پس از محاسبه امتیاز هر خوشه، میانگین امتیاز خوشه‌ها به‌عنوان امتیاز کلی خوشه‌بندی مطابق با فرمول‌های (۹) و (۱۰) محاسبه می‌شود [۸].

$$\mathrm{EScore}(\mathcal{C}) = \frac{\sum_{e\in E_{in\_\mathcal{C}}} w_e + \frac{\sum_{e\in E_{ex\_\mathcal{C}}}(-w_e)}{2}}{|E_{in\_\mathcal{C}}|+|E_{ex\_\mathcal{C}}|} \quad (۹)$$

$$\mathrm{CS-Measure} = \frac{\sum_{\mathcal{C}\in C}\mathrm{EScore}(\mathcal{C})}{|C|} \quad (۱۰)$$

که در آن $\mathrm{EScore}(\mathcal{C})$ امتیاز خوشه $\mathcal{C}$، $E_{in\_\mathcal{C}}$ مجموعه یال‌های داخلی خوشه $\mathcal{C}$، $E_{ex\_\mathcal{C}}$ مجموعه یال‌های خروجی از خوشه $\mathcal{C}$، We وزن یال e، |C| تعداد خوشه‌های حاصل از خوشه‌بندی موردنظر را نشان می‌دهد.



شکل ۱۰: مقایسه CS-Measure روش‌های خوشه‌بندی

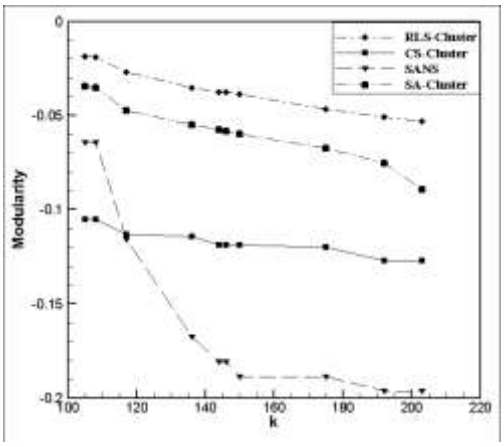
درروش CS-Cluster در هنگام انتخاب مراکز خوشه‌ها و همچنین توسعه خوشه‌ها ویژگی‌های ساختاری و محتوایی گره‌ها به‌طور هم‌زمان در نظر گرفته می‌شود. درنتیجه خوشه‌های حاصل ازلحاظ ساختاری و محتوایی شباهت بالایی دارند و باعث می‌شود ازنظر معیار CS-Measure مقادیر بالایی را به‌دست آورند. همان‌طور که در شکل (۱۰) مشاهده می‌شود پس‌ازاین روش، روش پیشنهادی RSL مقدار مناسبی ارائه داده درحالی‌که دو روش SA-Cluster و SANS کمترین مقدار را دارند.

۴ ۲ ۵ -پیمانگی

این معیار که درواقع خوشه‌بندی را ازنظر ساختاری ارزیابی می‌کند به‌صورت فرمول (۱۱) محاسبه می‌شود [۲۴]:

$$Q = \frac{1}{4m}\sum_s (m_S - E(m_S)) \quad (۱۱)$$

که در آن، m تعداد یال‌های گراف، m_s تعداد یال‌های موجود در خوشه S، $E(m_s)$ تعداد یال‌های مورد انتظار در خوشه S را نشان می‌دهد. همان‌طور که در شکل (۱۱) مشاهده می‌شود، با توجه به این‌که در روش ارائه شده RLS-Cluster سعی می‌شود یال‌هایی حذف شوند که در محله‌ای با میانگین شباهت کم و تراکم کم قرار دارند، درنتیجه درنهایت خوشه‌هایی که به‌دست می‌آیند دارای پیمانگی بالایی هستند.

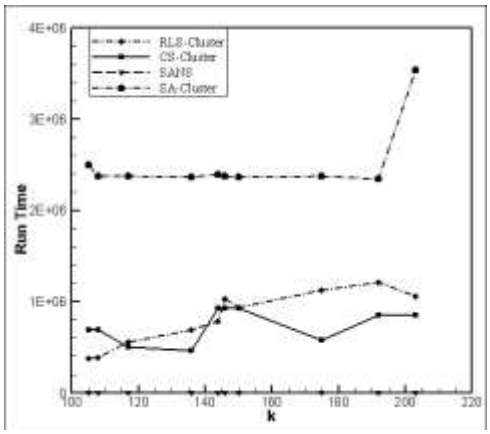


شکل ۱۱: پیمانگی روش‌های خوشه‌بندی

درروش SA-Cluster نیز مراکز خوشه‌ها در نقاط متراکم گراف انتخاب می‌شود و سپس گره‌های همسایه این گره‌های مرکزی به هر خوشه تخصیص داده می‌شود. این عمل درنهایت منجر به بالا بودن پیمانگی خوشه‌ها می‌شود. روش SANS نیز کمترین پیمانگی را دارد. علت این پیمانگی پایین درروش SANS این است که بعد از تعیین مراکز خوشه‌ها، گره‌های مشابه در هر نقطه‌ای از گراف به خوشه اضافه می‌شوند و این باعث می‌شود درنهایت پیمانگی کلی خوشه‌ها پایین باشد. از جدول (۱) تا جدول (۶) نتایج آزمون آماری ارزیابی‌های صورت پذیرفته در اجرای الگوریتم‌های خوشه‌بندی ساختاری-محتوایی گراف ارائه شده است.

۶ ۲ ۴ -زمان اجرا

با استفاده از این معیار مدت‌زمان خوشه‌بندی محاسبه می‌شود. با توجه به نمودار حاصل از آزمایش‌های انجام‌شده که در شکل (۱۲) مشاهده می‌شود؛ از بین روش‌های مقایسه شده روش SANS نسبت به بقیه روش‌ها سریع‌تر است، پس‌ازآن روش‌های RLS-Cluster و CS-Measure با سرعت حدوداً برابر عمل خوشه‌بندی را انجام می‌دهد. روش SA-Cluster نیز نسبت به بقیه کندتر است.



شکل ۱۲: زمان اجرای روش‌های خوشه‌بندی

۵ -نتیجه‌گیری و کارهای آینده

در بسیاری از کاربردهای کنونی دنیای واقعی در خوشه‌بندی گراف هم ساختار گراف و هم محتوای گره‌ها از اهمیت بالایی برخوردار هستند. به‌عنوان‌مثال، 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. بااین‌وجود، بیشتر روش‌های خوشه‌بندی که تاکنون ارائه‌شده 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 ارائه‌شده است.

در این مقاله، با در نظر گرفتن جنبه‌های ساختاری و محتوایی گراف، برای خوشه‌بندی ساختاری-محتوایی گراف روش RLS-Cluster پیشنهاد شد. این روش به‌صورت سلسله مراتبی با حذف یال‌هایی که دو گره اطراف آن دارای کمترین میزان شباهت هستند، گراف را به تعداد خوشه مدنظر تقسیم می‌کند. کارایی این روش با سایر روش‌های خوشه‌بندی ساختاری-محتوایی ارائه شده مورد ارزیابی قرارگرفته است که نتایج کلی آن در جدول (۷) آمده است. همان‌طور که گفته شد خوشه‌بندی که بین ساختار و محتوا تعادل خوبی برقرار کند، به‌لحاظ خوشه‌بندی ساختاری-محتوایی عملکرد خوبی از خود نشان داده است. روش ارائه شده به‌لحاظ پیمانگی و تراکم بهترین عملکرد را از خود نشان داده است. روش RSL-Cluster از نظر معیارهای زمان اجرا، میانگین شباهت و CS-Measure نیز پاسخ قابل قبولی نسبت به سایرین ارائه می‌دهد، این در حالی است که ازلحاظ خطای یال با توجه به این‌که در این روش خوشه‌بندی برخی از یال‌ها حذف می‌شوند، دارای عملکرد مطلوبی نسبت به دو روش CS-Cluster و SANS نیست ولی از SA-Cluster بهتر عمل می‌کند.

اغلب معیارهای ارزیابی روش‌های خوشه‌بندی بر اساس ساختار گراف خوشه‌ها را ارزیابی می‌کنند، بنابراین با توجه به کمبود معیارهای ارزیابی ساختاری-محتوایی 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. در روش RLS-Cluster یک گراف وزن‌دار وجود دارد که وزن هر یال نشان‌دهنده شباهت ساختاری-محتوایی گره‌های اطراف (محله) گره موردنظر است، به‌منظور افزایش کارایی و کاهش زمان الگوریتم می‌توان به‌جای محاسبه وزن یال بر اساس فرمول (۲) از رابطه دیگری استفاده کرد. به‌منظور توسعه خوشه‌ها نیز می‌توان روشی غیر از حذف یال را به‌کار برد.

جدول ۱: مقایسه تراکم روش‌های خوشه‌بندی

تعداد خوشه / روش	۱۰۵	۱۰۸	۱۱۷	۱۳۶	۱۴۴	۱۴۶	۱۵۰	۱۷۵	۱۹۲	۲۰۳
RLS-Cluster	۰/۹۲۵۲۳	۰/۹۲۳۱۴	۰/۸۹۱۳۱	۰/۸۵۹۴	۰/۸۵۰۰۷	۰/۸۴۹۵۵	۰/۸۴۵۳۸	۰/۸۱۲۸۹	۰/۷۹۶۸۴	۰/۷۸۷۵۷
CS-Cluster	۰/۵۷۵۹۳	۰/۵۷۵۹۳	۰/۵۴۷۳۶	۰/۵۴۳۳۱	۰/۵۲۴۳۹	۰/۵۲۴۳۹	۰/۵۲۴۳۹	۰/۵۲۰۴۸	۰/۴۹۲۰۴	۰/۴۹۲۰۴
SANS	۰/۷۴۱۶۴	۰/۷۴۱۶۴	۰/۵۳۷۳۱	۰/۳۲۹۲۰	۰/۲۷۶۰۹	۰/۲۷۶۰۹	۰/۲۴۵۶۹	۰/۲۴۵۶۹	۰/۲۱۵۴۲	۰/۲۱۵۴۲
SA-Cluster	۰/۸۶۲۰۸	۰/۸۵۸۸۲	۰/۸۰۹۸۹	۰/۷۷۹۶۱	۰/۷۶۹۸۳	۰/۷۶۶۴۴	۰/۷۶۱۲۲	۰/۷۳۰۸۱	۰/۶۹۹۱۱	۰/۶۴۲۷۴

جدول ۲: مقایسه خطای یال روش‌های خوشه‌بندی

تعداد خوشه / روش	۱۰۵	۱۰۸	۱۱۷	۱۳۶	۱۴۴	۱۴۶	۱۵۰	۱۷۵	۱۹۲	۲۰۳
RLS-Cluster	۱۰۱۷۹۷۳	۱۰۱۵۱۳۴	۱۰۰۱۱۶۴	۹۵۶۶۵۶	۹۳۴۷۲۷	۹۲۵۱۷۹	۹۲۱۱۳۰	۸۸۵۱۳۰	۸۶۲۷۸۰	۸۴۳۲۵۴
CS-Cluster	۲۴۷۹۳	۲۴۷۹۳	۲۲۱۰۴	۲۰۸۸۹	۱۳۲۵۹	۱۳۲۵۹	۱۳۲۵۹	۱۲۸۱۱	۹۴۳۰	۹۴۳۰
SANS	۱۲۰۵۶۱۶	۱۲۰۵۶۱۶	۸۵۲۳۸۲	۴۷۴۷۲۴	۳۸۳۷۱۸	۳۸۳۷۱۸	۳۰۷۴۷۵	۳۰۷۴۷۵	۲۴۸۳۹۸	۲۴۸۳۹۸
SA-Cluster	۱۵۱۷۸۴۸	۱۵۱۰۹۲۵	۱۴۸۰۵۲۰	۱۴۴۱۶۰۷	۱۴۲۸۱۸۵	۱۴۲۱۴۷۸	۱۴۱۳۱۲۹	۱۳۶۵۲۰۰	۱۳۲۳۰۲۶	۱۲۷۷۱۷۵

جدول ۳: مقایسه میانگین شباهت روش‌های خوشه‌بندی

تعداد خوشه / روش	۱۰۵	۱۰۸	۱۱۷	۱۳۶	۱۴۴	۱۴۶	۱۵۰	۱۷۵	۱۹۲	۲۰۳
RLS-Cluster	۰/۰۶۵۷۸	۰/۰۶۳۸۳	۰/۰۶۰۸۳	۰/۰۵۵۲۴	۰/۰۵۳۰۰	۰/۰۵۳۲۷	۰/۰۵۱۷۲	۰/۰۴۵۴۳	۰/۰۴۱۶۳	۰/۰۴۱۱۱
CS-Cluster	۰/۱۱۰۳۲	۰/۱۱۰۳۲	۰/۱۱۰۷۱	۰/۱۱۲۶۴	۰/۱۱۲۱۲	۰/۱۱۲۱۲	۰/۰۱۱۲۱۲	۰/۱۱۵۵۷	۰/۱۱۵۵۲	۰/۱۱۵۵۲
SANS	۰/۰۳۳۷۰	۰/۰۳۳۷۰	۰/۰۳۲۸۴	۰/۰۳۱۰۴	۰/۰۲۶۰۰	۰/۰۲۶۰۰	۰/۰۲۵۳۱	۰/۰۲۵۳۱	۰/۰۲۴۶۲	۰/۰۲۴۶۴
SA-Cluster	۰/۰۰۷۳۳	۰/۰۰۷۳۲	۰/۰۱۰۴۷	۰/۰۱۱۹۰	۰/۰۱۱۲۴	۰/۰۱۲۵۹	۰/۰۱۱۸۶	۰/۰۱۱۷۱	۰/۰۱۲۲۷	۰/۰۱۳۱۹

جدول ۴: مقایسه CS-Measure روش‌های خوشه‌بندی

تعداد خوشه / روش	۱۰۵	۱۰۸	۱۱۷	۱۳۶	۱۴۴	۱۴۶	۱۵۰	۱۷۵	۱۹۲	۲۰۳
RLS-Cluster	۰/۰۰۲۰۹	۰/۰۰۲۰۵	۰/۰۰۱۹۵	۰/۰۰۲۳۸	۰/۰۰۲۳۰	۰/۰۰۲۲۸	۰/۰۰۲۵۴	۰/۰۰۲۷۶	۰/۰۰۲۵۹	۰/۰۰۲۵۰
CS-Cluster	۰/۰۳۹۰۸	۰/۰۳۹۰۸	۰/۰۳۷۳۴	۰/۰۳۶۳۰	۰/۰۳۰۹۰	۰/۰۳۰۹۰	۰/۰۳۰۹۰	۰/۰۳۰۳۹	۰/۰۲۳۹۱	۰/۰۲۳۹۱
SANS	۰/۰۰۰۴۵	۰/۰۰۰۴۵	۰/۰۰۰۴۹	۰/۰۰۰۴۵	۰/۰۰۰۳۴	۰/۰۰۰۳۴	۰/۰۰۰۱۸	۰/۰۰۰۱۸	-۰/۰۰۰۰۵	-۰/۰۰۰۰۰۵
SA-Cluster	۰/۰۰۰۳۲	۰/۰۰۰۳۲	۰/۰۰۰۲۹	۰/۰۰۰۲۸	۰/۰۰۰۲۷	۰/۰۰۰۲۷	۰/۰۰۰۲۷	۰/۰۰۰۲۶	۰/۰۰۰۲۴	۰/۰۰۰۲۱

جدول ۵: مقایسه ماژولاریتی اجرای روش‌های خوشه‌بندی

تعداد خوشه / روش	۱۰۵	۱۰۸	۱۱۷	۱۳۶	۱۴۴	۱۴۶	۱۵۰	۱۷۵	۱۹۲	۲۰۳
RLS-Cluster	-۰/۰۱۸۶۹	-۰/۰۱۹۲۱	-۰/۰۲۷۱۷	-۰/۰۳۵۱۳	-۰/۰۳۷۴۸	-۰/۰۳۷۶۱	-۰/۰۳۸۶۵	-۰/۰۴۶۷۷	-۰/۰۵۰۷۸	-۰/۰۵۳۱۰
CS-Cluster	-۰/۱۰۵۰۳	-۰/۱۰۵۰۳	-۰/۱۱۳۱۵	-۰/۱۱۴۱۷	-۰/۱۱۸۶۷	-۰/۱۱۸۶۷	-۰/۱۱۸۶۷	-۰/۱۱۹۸۷	-۰/۱۲۶۹۹	-۰/۱۲۶۹۹
SANS	-۰/۰۶۴۵۸	-۰/۰۶۴۵۸	-۰/۱۱۵۶۷	-۰/۱۶۷۷	-۰/۱۸۰۹۷	-۰/۱۸۰۹۷	-۰/۱۸۸۵۷	-۰/۱۸۸۵۷	-۰/۱۹۶۱۴	-۰/۱۹۶۱۴
SA-Cluster	-۰/۰۳۴۴۷	-۰/۰۳۵۲۹	-۰/۰۴۷۵۲	-۰/۰۵۵۰۹	-۰/۰۵۷۵۴	-۰/۰۵۸۳۹	-۰/۰۵۹۶۹	۰/۰۶۷۲۹	-۰/۰۷۵۲۲	-۰/۰۸۹۳۱

جدول ۶: مقایسه زمان اجرای روش‌های خوشه‌بندی

تعداد خوشه / روش	۱۰۵	۱۰۸	۱۱۷	۱۳۶	۱۴۴	۱۴۶	۱۵۰	۱۷۵	۱۹۲	۲۰۳
RLS-Cluster	۳۷۲۴۹۱	۳۸۲۰۱۲	۵۵۳۲۴۷	۶۹۲۴۳۷	۷۷۹۱۱۳	۱۰۲۹۲۳۴	۹۳۸۴۲۵	۱۱۲۸۱۹۲	۱۲۰۶۲۴۰	۱۰۶۰۸۶۸
CS-Cluster	۶۸۹۵۶۱	۶۸۹۵۶۱	۴۹۷۵۴۹	۴۶۵۳۷۳	۹۲۳۱۶۷	۹۲۳۱۶۷	۹۲۳۱۶۷	۵۷۳۶۵۰	۸۴۷۰۵۵	۸۴۷۰۵۵
SANS	۹۲	۹۲	۹۳	۹۴	۶۷	۶۷	۷۶	۷۶	۱۱۸	۱۱۸
SA-Cluster	۲۴۹۶۳۰۲	۲۳۷۷۲۵۶	۲۳۷۸۵۲۶	۲۳۶۷۴۳۳	۲۳۹۰۲۶۵	۲۳۷۶۷۵۶	۲۳۶۱۵۶۶	۲۳۷۶۰۲۹	۲۳۴۹۴۹۷	۳۵۴۲۸۸۲

جدول ۷: مقایسه کلی روش‌های خوشه‌بندی بر اساس معیارهای ارزیابی‌شده در مقاله

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

مراجع

[۱] مریم مرادی، رزا یوسفیان و وحید رافع، «ارائه راهکاری جهت مقابله با مشکل انفجار فضای حالت در سیستم‌های تبدیل گراف با استفاده از الگوریتم‌های پرندگان و جستجوی گرانشی»، *مجله مهندسی برق دانشگاه تبریز*، دوره ۴۵، شماره ۴، صفحه ۱۶۳-۱۷۷، زمستان ۱۳۹۴.

[۲] مرتضی فرهید، موسی شمسی، محمدحسین صداقی، «تأثیر توپولوژی شبکه‌های پیچیده بر روی عملکرد تخمین تطبیقی توزیع شده با مشارکت نفوذی»، *مجله مهندسی برق دانشگاه تبریز*، دوره ۴۶، شماره ۴، صفحه ۲۰۷-۲۱۶، زمستان ۱۳۹۵.

[۳] سمیه توکلی، افسانه فاطمی، «تشکیل تیم دوهدفه در شبکه‌های اجتماعی»، *مجله مهندسی برق دانشگاه تبریز*، دوره ۴۷، شماره ۲، صفحه ۴۲۳-۴۳۳، تابستان ۱۳۹۶.

[۴] سمیرا رفیعی، پرهام مرادی، «بهبود عملکرد الگوریتم خوشه‌بندی فازی سی- مینز با وزن‌دهی اتوماتیک و محلی ویژگی‌ها»، *مجله مهندسی برق دانشگاه تبریز*، دوره ۴۶، شماره ۲، صفحه ۷۵-۸۶، تابستان ۱۳۹۵.

[5] C. Aggarwal, H. Wang, *Managing and Mining Graph Data*, Springer US, 2010.

[6] S. B. Patkar, H. Narayanan "An Efficient Practical Heuristic for Good Ratio-Cut Partitioning", *16th International Conference on VLSI Design (VLSI'03)*, pp. 1-6, 2003.

[7] A. E. Feldmann, L. Foschini, "Balanced Partitions of Trees and Applications"; *ALGORITHMICA*. Vol. 71, pp. 354-376, 2015.

[۸] کبری رحمتی، حسن نادری، سامان کشوری، «خوشه‌بندی محتوایی-ساختاری گراف و معیاری جدید جهت ارزیابی آن»، *مجله علوم و فناوری‌های پدافند نوین*، دوره ۹، شماره ۲، تابستان ۱۳۹۷ (در نوبت چاپ).

[9] M. Newman, "Community Detection in Networks: Modularity Optimization and Maximum Likelihood are Equivalent", *Social and Information Networks* (cs.SI). vol. 94, pp. 1-8, 2016.

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

[11] S. Fortunatoa, D. Hricb, "Community Detection in Networks: A User quide", PHYS REP. vol. 659, pp. 1-44, 2016.

[12] M. Khatoon, W. Aisha Banu, "A Survey on Community Detection Methods in Social Networks", *Education and Management Engineering (IJEME)*. vol. 1, pp. 8-18, 2015.

[13] H. Elhadi, G. Agam, "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*, pp. 1-7, 2013.

[14] S. Harenberg, G. Bello, L. Gjeltema, S. Ranshous, J. Harlalka, R. Seay, K. Padmanabhan, N. Samatova, "Community detection in large-scale networks: a survey and empirical evaluation", *Computational Statistics*, vol. 6, pp. 426-439, 2014.

[15] J. R. Matthew, M. Maier, D. Jensen, "Graph Clustering with Network Structure Indices", *ICML '07 Proceedings of the 24th Int. Con. on Machine learning*, pp. 783-790, 2007.

[16] V. Shchukin, D. Khristich, I. Galinskaya, "Word Clustering Approach to Bilingual Document Alignment", *First Con. on Machine Translation*, vol 2, pp. 953-994, 2016.

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

[18] J. Han, M. Kamber, J. Pei. "Data Mining: Concepts and Techniques", 3rd ed, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers, 2011.

[19] Y. Zhou, H. Cheng, J. Xu Yu, "Graph Clustering Based on Structural/Attribute Similarities"; VLDB. vol. 2, pp. 718-729, 2009.

[20] M. Parimala, L. Daphne, "Graph Clustering based on Structural Attribute Neighborhood Similarity (SANS)"; *IEEE international Conference on Electrical, Computer and Communication Technologies (ICECCT)*. pp. 1-5, 2015.

[21] S. Pool, F. Bonchi, M. Leeuwen, "Description-Driven Community Detection" *ACM Transactions on Intelligent Systems and Technology (TIST)*. Vol. 5, pp.1-25, 2014.

[22] M. Qiao, L. Qin, H. Cheng, J. X. Yu, W. Tian, "Top-K Nearest Keyword Search on Large Graphs", *VLDB*, vol. 10, pp. 901-912, 2013.

[23] M. Wang, Ch. Wang, J. Xu Yu, J. Zhang, "Community Detection in Social Networks: An In-depth Benchmarking Study with a Procedure-Oriented Framework"; *VLDB*. Vol. 8, pp. 998-1009, 2015.

[24] J. Yang, J. Leskovec, "Defining and evaluating network communities based on ground-truth", *Knowledge and Information Systems (KAIS)*, vol. 42, pp. 181-213, 2015.

زیرنویس‌ها

[1] Heuristic methods
[2] Structural Attribute Cluster
[3] Structural Attribute Neighbourhood Similarity
[4] Centroid
[5] Description-Driven Community Detection
[6] Structural Attribute Neighbourhood Similarity
[7] Density
[8] Modularity
[9] Remove Lowest Similarity
[10] Eclipse
[11] Jaccard similarity
[12] http://ir.ii.uam.es/hetrec2011/.
[13] average similarity
[14] density
[15] modularity
[16] average degree
[17] cut size
[18] conductance
[19] error link
[20] recall
[21] precision