نشریه علمی «علوم و فناوری‌های پدافند نوین»
سال یازدهم، شماره ۱، بهار ۱۳۹۹؛ ص ۱۲۱-۱۰۷

یک راه‌حل افزایشی جهت خوشه‌بندی محتوایی- ساختاری یک گراف

سامان کشوری[۱]، محمدعلی جوادزاده[۲*]، حسن نادری[۳]

۱- دانشجوی کارشناسی ارشد، ۲- استادیار، دانشگاه جامع امام حسین (ع) ۳- استادیار، دانشگاه علم و صنعت ایران

(دریافت: ۹۸/۰۱/۲۱، پذیرش: ۹۸/۰۵/۱۵)

چکیده

خوشه‌بندی گره‌های گراف از جنبه ساختاری یا محتوایی، همواره موردتوجه پژوهشگران حوزه داده‌کاوی بوده است؛ اما به خوشه‌بندی گراف بر مبنای ساختار و محتوا به‌طور ترکیبی کمتر توجه شده است. با توجه به نیاز خوشه‌بندی ساختاری-محتوایی در شبکه‌های اطلاعاتی که شبکه‌های اجتماعی نمونه‌ای از آن‌هاست، در این مقاله الگوریتم خوشه‌بندی ICS-Cluster ارائه‌شده که هر دو جنبه ساختار و محتوا را به‌صورت هم‌زمان در نظر می‌گیرد. هدف این روش، رسیدن به خوشه‌هایی با ساختار درونی منسجم (ساختاری) و مقادیر ویژگی (محتوایی) همگن در گراف است. در این روش ابتدا گراف اولیه به یک گراف ساختاری-محتوایی تبدیل می‌شود که در آن وزن هر یال (ارتباط) بیانگر شباهت ساختاری-محتوایی دو گره (موجودیت) است. خوشه‌بندی با توجه به وزن یال‌ها به‌صورت افزایشی انجام می‌شود بدین معنا که گره‌های یالِ با وزن بالا به‌عنوان خوشه در نظر گرفته می‌شوند و وزن یال‌های متصل به خوشه با یکدیگر ادغام‌شده و به‌صورت یک یال متصل به خوشه در نظر گرفته می‌شوند، این مراحل تا زمانی که الگوریتم به تعداد خوشه موردنظر کاربر برسد، ادامه خواهد یافت. الگوریتم ICS-Cluster به هر تعداد خوشه که مدنظر کاربر است، گراف را خوشه‌بندی می‌کند. مقایسه الگوریتم مطرح‌شده با سه الگوریتم خوشه‌بندی ساختاری- محتوایی ارائه‌شده، بر اساس معیارهای شش‌گانه سنجش کیفیت خوشه، بیانگر عملکرد مناسب روش ICS-Cluster است. این معیارها معیارهای ساختاریِ تراکم خوشه، خطای یال و پیمانگی، معیار محتواییِ میانگین شباهت، معیار ساختاری-محتوایی CS-Measure و زمان اجرای روش‌ها است.

کلیدواژه‌ها: خوشه‌بندی گراف، خوشه‌بندی ساختاری- محتوایی، گراف خصوصیت، استخراج خوشه

An Incremental Solution for Content-Structural Graph Clustering

S. Keshvari, M. A. Javadzadeh*, H. Naderi
Imam Hossein University
(Received: 10/04/2019; Accepted: 06/08/2019)

Abstract

Researchers have always been interested in graph nodes clustering based on content or structure. But less attention has been paid to clustering based on both structure and content. But a content-structural clustering is needed in information networks like social networks. In this paper, the ICS-Cluster algorithm is proposed which takes into consideration both the structure and content aspects of the nodes. The purpose of this approach is to gain a coherent internal structure (structural aspect) and homogeneous attribute values (content aspect) in the graph. In this approach firstly the graph is converted into a content-structural graph which edges' weight show similarity between the connected nodes. Incremental clustering is done based on edges' weight in this process the edges with the most weight is considered as clusters then the weight of connected edge to the cluster is aggregated and they'll be one edge, the process is repeated until the algorithm reaches the number of clusters that indicated by the user. ICS-Cluster algorithm number of cluster is indicated by the user. Comparing ICS-Cluster with other content structural algorithm based on six criteria for measuring cluster quality shows that ICS-Cluster has good performance. These criteria contain structural criteria (Modularity, Error Link, and Density), content criterion (Average Similarity), content-structural criterion (CS-Measure) and the run time.

Keywords: Graph Clustering, Content-Structural Clustering, Attributed Graph, Cluster Extraction

*Corresponding Author E-mail: javadzadeh@ihu.ac.ir

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 و ۳]. از گراف به‌منظور تحلیل مسائل مختلف استفاده می‌شود [۵-۴]، شبکه‌های اجتماعی نیز بر روی گراف مدل می‌شوند و ازآنجایی‌که استفاده از آن‌ها در جامعه در حال افزایش است حجم گراف حاصل از آن‌ها افزایش میابد بنابراین، پرداختن به خوشه‌بندی آن‌ها بیش‌ازپیش اهمیت دارد [۷-۶]. محققان همواره سعی بر بهبود الگوریتم‌های خوشه‌بندی به‌منظور افزایش کیفیت خوشه‌های به‌دست‌آمده دارند [۹-۸].

خوشه‌بندی مسئلها ی مهم در تحلیل داده‌های گراف و یک روش دسته‌بندی بدون ناظر برای داده‌های گراف است. خوشه-بندی شامل تقسیم گره‌های گراف به گروه‌هایی با گره‌های مشابه است. شباهت بین گره‌ها معمولاً توسط یک تابع هدف ریاضی تعریف می‌شود [۱۰]. 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 [۱۲-۱۱]. اغلب روش‌های خوشه‌بندی که تاکنون ارائه‌شده است فقط یک جنبه ساختاری یا محتوای گراف را در نظر گرفته‌اند و روش‌های خوشه‌بندی که هم ساختار و هم محتوا را در نظر بگیرند کمتر ارائه‌شده است.

به‌عنوان‌مثال خوشه‌بندی‌های مختلف افراد در شبکه‌های اجتماعی که در شکل (۱) نشان داده شده است. در این خوشه‌بندی اگر نیاز باشد افرادی که با یکدیگر به‌واسطه ویژگی‌های شخصی (در این مثال سن) در ارتباط بوده را خوشه‌بندی کرد، نمی‌توان تنها ویژگی‌های محتوایی یا ساختاری را در نظر گرفت. همان‌طور که در شکل نشان داده شده، اگر خوشه‌بندی بر اساس ساختار گراف تشکیل‌شده انجام شود، ویژگی‌های افراد نادیده گرفته خواهد شد، اگر ویژگی افراد مدنظر قرار گیرد، ارتباط بین آن‌ها نادیده گرفته خواهد شد؛ بنابراین، بهتر است ویژگی‌های ساختاری و محتوایی گراف به‌صورت هم‌زمان در نظر گرفته شود. مطابق با شکل (۱) برای فرد شماره ۵ بااینکه ازلحاظ محتوایی جوان است ولی چون به لحاظ ساختاری با چهار فرد میان‌سال در ارتباط است، در خوشه مربوط به میان‌سال‌ها قرار خواهد گرفت.

هدف کلی، تشخیص خوشه‌هایی از گراف است که گره‌های درونی هر خوشه به‌طور متراکم به یکدیگر مرتبط بوده و ازلحاظ معنایی شبیه به هم هستند. در این حالت باید ارتباط ساختاری و معنایی بین خوشه‌های مجزا به حداقل خود برسد. در این مقاله یک روش خوشه‌بندی سلسله مراتبی ارائه‌شده است که ابتدا گراف اولیه -که در آن یال نشان‌دهنده ارتباط ساختاری و وزن آن بیانگر شباهت محتوایی است- به گرافی ساختاری-محتوایی تبدیل می‌شود. در این گراف دو گره که شباهت ساختاری-محتوایی آن‌ها بیش‌ازحد آستانه است به یکدیگر متصل شده و وزن آن میزان شباهت ساختاری-محتوایی خواهد بود.

روش خوشه‌بندی ICS-Cluster[2] بر روی گراف مذکور انجام می‌شود. در این روش ابتدا گره‌های اطراف یال با بیش‌ترین وزن به‌عنوان خوشه در نظر گرفته می‌شود و گره‌های دیگر به‌صورت مرحله‌به‌مرحله به خوشه‌های تشکیل‌شده اضافه می‌شود.

در ادامه مقاله، در بخش ۲، کارهای انجام‌شده درزمینه خوشه‌بندی گراف به‌صورت مختصر مطرح شده است. در بخش ۳، روش تحقیق و روش ارائه‌شده 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 کلی مقاله مطرح می‌شود.

۲. پیشینه تحقیق

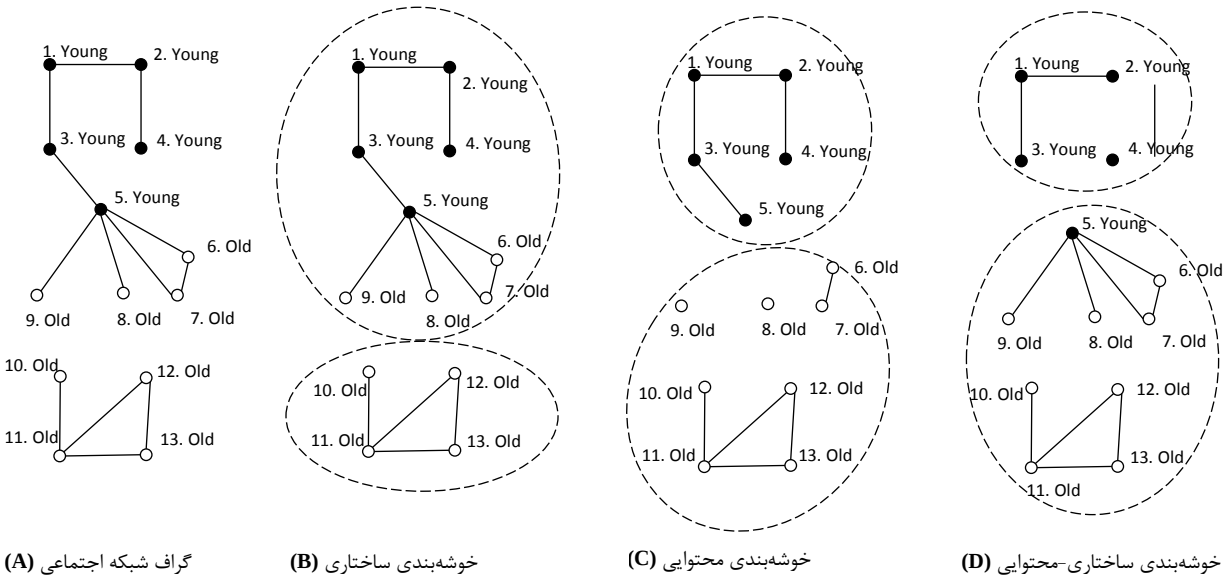
خوشه‌بندی گراف یکی از راه‌های تحلیل گراف‌های حجیم و پیچیده است که تاکنون روش‌های خوشه‌بندی مختلفی برای آن ارائه و استفاده شده است [۱۴-۱۳]. بیشتر این روش‌ها برای عمل خوشه‌بندی تنها جنبه ساختاری گراف را در نظر می‌گیرند [۱۶-۱۵]، که از این دسته می‌توان TopGC و Louvain را نام برد [۱۸-۱۷].

برخی روش‌ها نیز مطرح شده‌اند که عمل خوشه‌بندی را بر اساس محتویات گره‌ها انجام می‌دهند، ازجمله این دست روش‌ها می‌توان k-Means CAMAS[۱۹] و k-Medoids [۲۱-۲۰] را نام برد که تحقیقات برای ارائه روش خوشه‌بندی محتوایی همچنان در حال انجام است [۲۳-۲۲]. در بسیاری از کاربردهای دنیای واقعی هر دو جنبه ساختار و محتوا در کنار یکدیگر و به طور هم‌زمان موردنظر هستند [۲۵-۲۴].

به‌عنوان‌مثال این مسئله در کشف جوامع در شبکه‌های اجتماعی از اهمیت بالایی برخوردار است [۲۸- ۲۶]، بنابراین، بهتر است در خوشه‌بندی شباهت ساختاری و محتوایی گره‌ها در کنار یکدیگر در نظر گرفته شود. در ادامه برخی از روش های خوشه‌بندی ساختاری-محتوایی بیان شده است.

[1] Heuristic Methods

[2] Incremental Content-Structural Clustering



گراف شبکه اجتماعی **(A)** خوشه‌بندی ساختاری **(B)** خوشه‌بندی محتوایی **(C)** خوشه‌بندی ساختاری-محتوایی **(D)**

شکل ۱. نمونه شبکه اجتماعی با محتوای "سن"

SA-Cluster[1]: 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 تکرارشونده است که در آن ابتدا مرکزی‌ترین گره‌ها (گره‌ها با تراکم بالا) به عنوان 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 به روز می‌شود) و عملیات تا همگرا شدن تکرار می‌شود [۲۹]. ازآنجاکه این الگوریتم به‌صورت تکرارشونده است، زمان اجرای بالایی دارد.

SANS[2]: 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[3] 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 گره ها تکرار می‌شود [۳۰]. زمان اجرای این الگوریتم نسبت به SA-Cluster بهتر است و همچنین خوشه‌های متراکم‌تری به‌دست می‌آورد.

DCM[4]: 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 جامعه (ازنظر ساختاری) را به دست آورده و در گام بعد سعی می‌کند یک توصیف مناسب برای این جوامع به دست 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 [۳۱]. این روش لزوماً همه گره‌ها را پوشش نمی‌دهد و برای مجموعه داده‌های کوچک کند و برای مجموعه داده‌های بزرگ سرعت‌بالا دارد.

CS-Cluster[5]: 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 رابطه (۱) به‌دست 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

[1] Structural Attribute Cluster
[2] Structural Attribute Neighbourhood Similarity
[3] Centroid
[4] Description-Driven Community Detection
[5] Content-Structural Clustering

خوشه‌ها تا زمانی که تمامی گره‌ها پوشش داده شوند گسترش میابند. 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 [۳۲]. از نقاط ضعف‌های این روش می‌توان سرعت نامناسب، تراکم[1] و پیمانگی[2] به نسبت پایین اشاره کرد.

در روش‌های ارائه‌شده تاکنون میزان اهمیت شباهت محتوایی با ساختاری یکسان در نظر گرفته شده است، ولی در برخی از خوشه‌بندی‌ها میزان اهمیت شباهت ساختاری نسبت به شباهت محتوایی بیشتر بوده و بالعکس. به‌عنوان‌مثال در شبکه اجتماعی که در آن اطلاعات پروفایل کاربری زیادی وجود ندارد، اهمیت ساختاری گراف بیش از اهمیت محتوایی است، درحالی‌که در خوشه‌بندی صفحات وب، محتوا از اهمیت بالاتری نسبت به ساختار برخوردار است. در ادامه روشی ارائه شده که ضمن پارامتریک کردن اهمیت هرکدام، عملکرد مناسبی به لحاظ معیارهای مختلف از خود نشان می‌دهد.

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 به‌صورت خوشه‌بندی شود که هرکدام از خوشه‌ها علاوه بر نزدیکی ساختاری، ازلحاظ محتوایی نیز مشابه باشند. در ادامه روش ارائه‌شده تشریح می‌شود.

۳-۱. روش ارائه‌شده (ICS-Cluster[3])

در این روش ابتدا با ترکیب ساختار و محتوای گراف اولیه گرافی ساختاری-محتوایی تشکیل می‌شود. همان‌طور که اشاره شد، در روش‌های قبلی میزان اهمیت شباهت محتوایی با ساختاری یکسان در نظر گرفته شده است ولی در روش ارائه شده کاربر می‌تواند با دادن ضریب (وزن)، اهمیت هرکدام را نسبت به دیگری مشخص نماید. اگر شباهت محتوایی بین دو گره i و j را sim_{ij} و فاصله ساختاری را D در نظر گرفته شود شباهت ساختاری-محتوایی طبق رابطه (۲) محاسبه می‌شود، در گراف ساختاری-محتوایی که ایجاد می‌شود این مقدار به‌عنوان وزن یال بین دو گره i و j در نظر گرفته خواهد شد.

$$W_{(i,j)} = a\,(sim_{(i,j)}) + b\left(\frac{1}{D_{(i,j)}}\right) \qquad (۲)$$

که در آن $i \neq j$ بوده و a میزان اهمیت (وزن) شباهت محتوایی و b اهمیت (وزن) شباهت ساختاری میان دو گره i و j است. شباهت ساختاری بین دو گره، عکس فاصله آن دو از یکدیگر در نظر گرفته شده است. در این رابطه $D(i,j)$ مطابق با رابطه (۳) برابر با تعداد یال (فاصله) بین دو گره i و j است. درواقع هرچه این مقدار بیشتر باشد، شباهت ساختاری کمتر خواهد بود. این فاصله از طریق نسخه Modified الگوریتم دایجسترا[4] محاسبه می‌شود که pseudocode آن در شکل (۲) مشاهده می‌شود[۳۳].

$$D_{(i,j)} = \sum_{e=i}^{j} E_e \qquad (۳)$$

Input: A weighted graph $G(V, E, c)$, where V is a set of vertices (represent nodes), E is a set of edges connecting the vertices in V (they represent the connectivity between nodes), and c is a set of cost values (measured in ETX) associated with the edges in E. The SN id is also given as an input.

Output: A shortest path tree rooted at the SN (s) and that reaches each vertex in V through a shortest path.

BEGIN

1: $N \leftarrow s$ // N contains only the source node
2: $d(s) \leftarrow 0$
3: $r \leftarrow s$
4: **for each** node n in V (except the SN s) not in N **do**
5: $d(n) \leftarrow \infty$
6: **end for**
7: **repeat**
8: **for each** v directly connected to r, and not in N **do**
9: **if** $p(r)$ not NULL **then**
10: $c(r, v) = c(r, v) - prob[p(r)][v] \times c(p(r), r)$.
11: **end if**
12: $d(v) = \min\{d(v), d(r) + c(r, v)\}$
13: **end for**
14: find node x not in N and $d(x)$ is minimum
15: insert x into N
16: $p(x) = r$
17: $r = x$
18: **until** $N = V$

END

شکل ۲. الگوریتم دایجسترا استفاده شده [۳۳]

مقدار $sim_{(i,j)}$ نیز از طریق شباهت جاکارد طبق رابطه (۴) محاسبه می‌شود [۳۴].

$$sim_{(i,j)} = J(C, C') = \frac{|C \cap C'|}{|C \cup C'|} \qquad (۴)$$

که در آن J شباهت جاکارد است که برابر است با میزان اشتراک موجود بین دو گره بر روی کل معیارهای موجود در گره، که در بهترین حالت، برابر با یک است.

به‌منظور ساخت گراف ساختاری-محتوایی، اگر $W_{(i,j)}$ از مقداری که توسط کاربر مشخص می‌شود بیشتر بود، بین آن دو یالی را رسم کرده و وزن آن یال برابر با مقدار $W_{(i,j)}$ خواهد بود. در غیر این صورت یالی رسم نمی‌شود. لازم به ذکر است که حد آستانه موردنظر به‌صورت پیش‌فرض برابر با میانگین وزن یال‌ها

[1] Density
[2] Modularity
[3] Incremental Structural Content Clustering
[4] Dijkstra

در نظر گرفته شده است. بنابراین، گراف تشکیل‌شده یک گراف وزن‌دار است که وزن هر یال نشان‌دهنده شباهت ساختاری- محتوایی دو گره است. حین تشکیل گراف ساختاری- محتوایی از روی گراف قبلی، یال‌های (ارتباط‌های) حساس به‌دست می‌آیند. این یال‌ها شامل یال‌هایی است که بیش‌ترین وزن را دارند. در این شرایط یال‌های حساس طبق رابطه (۵) به‌دست می‌آیند.

$$\text{K-Sensitive-edge} = \text{K-MAX}\ (w_i) \qquad (۵)$$

که منظور از K-Sensitive-edge ترتیب یال‌هایی است که دارای بیش‌ترین مقدار وزن هستند و به‌عنوان یال‌های حساس محسوب می‌شوند. در این مرحله نه‌تنها یال‌های حساس به‌دست‌آمده بلکه گره‌های حساس نیز از طریق رابطه (۶) به‌دست می‌آید.

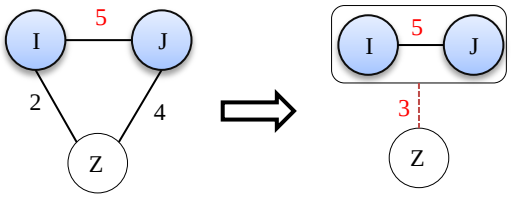
$$\text{K-Sensitive-node} = \text{K-MAX}\ (\sum_{i\in n} w_i) \qquad (۶)$$

که در آن، منظور از K-Sensitive-node تعداد K گره حساس است که برابر است با گره‌هایی که در آن جمع وزن یال‌های متصل به آن حداکثر هستند. در این رابطه i رئوس موردنظر، n تمامی رئوس و w وزن یال‌های متصل به گره است.

تا این مرحله گراف وزن‌داری وجود دارد که وزن هر یال میزان شباهت ساختاری و محتوایی بین دو گره آن را نشان می‌دهد؛ در ادامه و به‌منظور خوشه‌بندی ابتدا هر گره در گراف به‌عنوان یک خوشه در نظر گرفته می‌شود، سپس به ترتیب گره‌های اطراف یال‌های با وزن بالا با یکدیگر ادغام می‌شوند. با این کار این دو گره به‌عنوان یک مجموعه در نظر گرفته می‌شود. سپس این مجموعه به‌عنوان یک خوشه در نظر گرفته می‌شود و اگر گره دیگری به هر دو گره ادغام‌شده در ارتباط بوده، تنها با یک یال به این مجموعه متصل شده و وزن آن مطابق با رابطه (۷) برابر با میانگین حسابی وزن دو یال قبلی می‌شود.

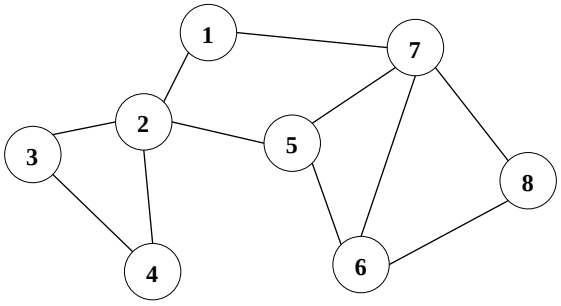
$$W_{new} = \frac{1}{n}\sum_{i=1}^{n} W_i \qquad (۷)$$

که در آن، n تعداد یال‌های موجود در مجموعه (خوشه) بوده که به گره خارج از خوشه متصل هستند. به همین ترتیب به‌صورت سلسله مراتبی تا جایی که تعداد خوشه‌ها برابر با تعدادی باشد که مدنظر کاربر است. یعنی با توجه به شکل (۳) اگر فرض شود یال بین دو گره i و j دارای بیش‌ترین وزن باشد، پس از ادغام آن‌ها گره z که قبلاً با وزن ۴ و ۲ به این گره‌ها وصل بود، با یک یال از طریق رابطه (۷) محاسبه و برابر با ۳ شده به‌روز می‌شود.



شکل ۳. یک نمونه ادغام گره در روش ICS-Cluster

برای ادامه خوشه‌بندی مجدداً بیش‌ترین مقدار وزن گره را در نظر گرفته و تا زمانی که تعداد خوشه‌ها به خوشه‌های مدنظر کاربر برسد، عمل خوشه‌بندی انجام خواهد شد. در ادامه مثالی از نحوه خوشه‌بندی ارائه می‌شود که در آن از مثال ارائه‌شده در [۳۲] استفاده شده است تا تفاوت‌های روش ارائه‌شده با آن روش به‌خوبی مشخص شود. در این مثال تعداد خوشه موردنظر کاربر ۲ در نظر گرفته‌شده است. در ابتدا گراف ساختاری را که در شکل (۴) داده شده است برای شروع کار در نظر گرفته می‌شود.

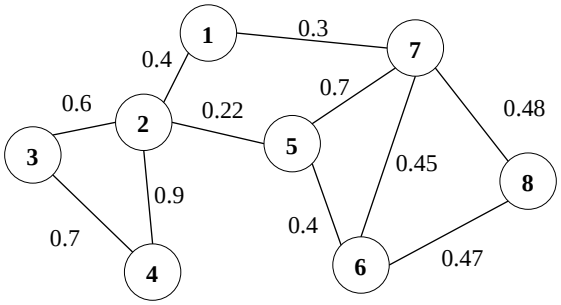


شکل ۴. ساختار گراف نمونه

ماتریس ارائه‌شده در زیر بیانگر میزان شباهت دو گره ازلحاظ محتوایی به یکدیگر هستند.

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

درواقع با مشخص شدن شباهت بین دو گره در هر یال، گراف به‌صورت شکل (۵) در خواهد آمد که در آن وزن هر یال بیانگر شباهت محتوایی دو گره که ارتباط ساختاری دارند است.

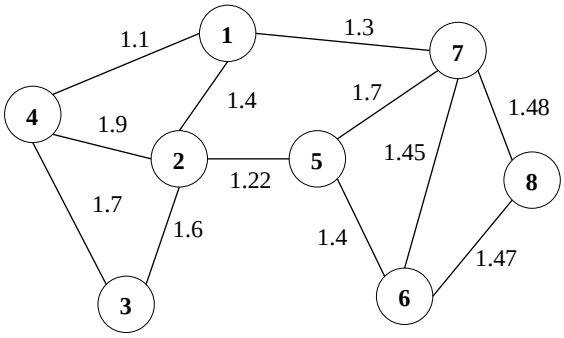


شکل ۵. گراف وزن‌دار نمونه

به‌منظور ساخت گراف جدید با توجه به راه‌حل ارائه‌شده برای هر دو گره موجود در گراف طبق رابطه (۲) وزن یال‌ها محاسبه می‌شود. به این منظور در این مثال میزان اهمیت شباهت ساختاری و محتوایی (یعنی مقدار a و b در این رابطه) هردو برابر با ۱ در نظر گرفته شده است. لازم به ذکر است که در این روش مقادیر a و b عددی بین ۰ تا ۱ باید در نظر گرفته شود تا از بروز پاسخ‌های غیرواقعی جلوگیری شود. ماتریس زیر نشان‌دهنده شباهت ساختاری-محتوایی دو گره است.

$$sim = \begin{bmatrix} - & 1.40 & 1.02 & 1.10 & 0.82 & 0.80 & 1.30 & 0.65 \\ - & - & 1.60 & 1.90 & 1.22 & 0.98 & 0.79 & 0.69 \\ - & - & - & 1.70 & 0.92 & 0.85 & 0.54 & 0.57 \\ - & - & - & - & 0.80 & 0.56 & 0.42 & 0.41 \\ - & - & - & - & - & 1.40 & 1.70 & 1.00 \\ - & - & - & - & - & - & 1.45 & 1.47 \\ - & - & - & - & - & - & - & 1.48 \\ - & - & - & - & - & - & - & - \end{bmatrix}$$

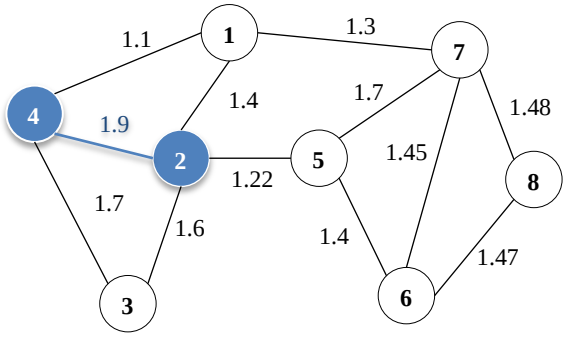
در ادامه باید میزان حد آستانه شباهت ساختاری- محتوایی دو گره تعیین شود، همان‌طور که گفته شد این شباهت به‌صورت پیش‌فرض برابر با میانگین وزن یال‌ها خواهد بود. میانگین شباهت دو گره به یکدیگر در این مثال برابر با ۱/۰۵ است که یال‌هایی که کمتر از این مقدار شباهت دارند رسم نمی‌شوند؛ درنتیجه گراف شکل (۶) نشان‌دهنده گراف ساختاری-محتوایی است که باید خوشه‌بندی شود.



شکل ۶. گراف ساختاری محتوایی نمونه

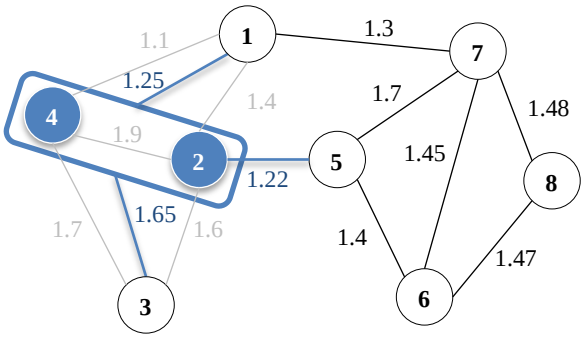
همان‌طور که مشاهده می‌شود، در گراف ساختاری یالی بین گره‌های ۱ و ۴ وجود نداشت که با اعمال اثر محتوا و ساختار در کنار یکدیگر، به دلیل شباهت محتوایی بالای این دو گره، بین آن‌ها یالی رسم شده است. در مرحله بعدی بیش‌ترین شباهت بین دو گره (حساس‌ترین یال‌ها) نقش مهمی در ایجاد خوشه‌ها ایفا می‌کنند، با توجه به اینکه حین ساخت گراف ساختاری-محتوایی، این مقادیر به‌صورت نزولی ذخیره می‌شوند، دسترسی به آن‌ها ساده است.

مطابق با شکل (۷) بیش‌ترین مقدار در گراف فرض شده ۱/۹ بوده که مربوط به یال بین ۲ و ۴ است، در این مرحله این دو گره به‌عنوان یک خوشه در نظر گرفته می‌شوند.



شکل ۷. مرحله اول خوشه‌بندی روش ارائه‌شده

ازآنجاکه دو گره ۲ و ۴ به‌عنوان یک خوشه در نظر گرفته‌شده‌اند، وزن یال گره‌های متصل به این مجموعه با توجه به رابطه (۷) به‌روز می‌شوند. نمایی از به‌روز شدن ارتباط در شکل (۸) نشان داده شده است.



شکل ۸. در نظر گرفتن گره‌های ۲ و ۴ به‌عنوان یک گره و ادغام گره‌های متصل به آن

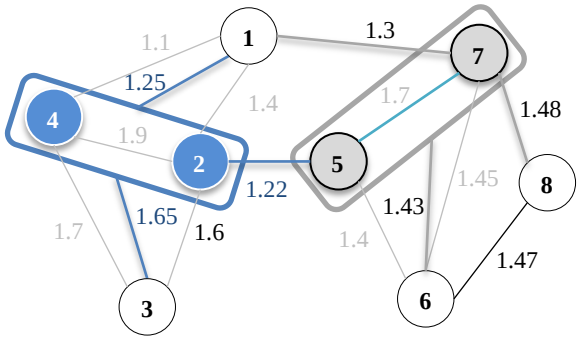
به‌عنوان‌مثال با توجه به این‌که گره ۱ با دو گره ۲ و ۴ در ارتباط بوده وزن ارتباط این گره با خوشه موردنظر به‌صورت زیر محاسبه می‌شود.

$$w_{new(1C_1)} = \frac{1}{2}(1.1 + 1.4) = 1.25$$

که در آن، C_1 نشان‌دهنده خوشه اول (متشکل از گره‌های ۲ و ۴) است. این مقدار برای گره ۳ به‌صورت زیر محاسبه می‌شود.

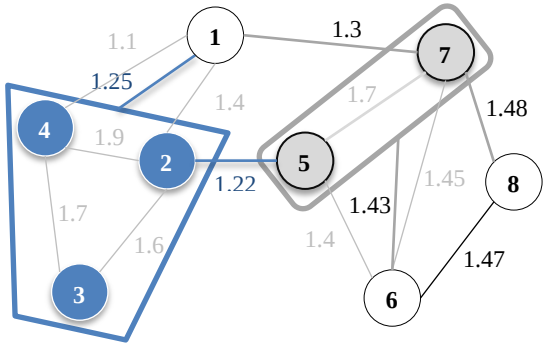
$$w_{new(3C_1)} = \frac{1}{2}(1.7 + 1.6) = 1.65$$

با توجه به این‌که گره ۵ تنها با گره ۲ از خوشه تشکیل‌شده در ارتباط است، مقدار موردنظر برای آن همان مقدار قبلی خواهد بود. در ادامه و پس از به‌روز شدن تمام ارتباط‌ها با خوشه موردنظر، به مقدار بعدیِ بزرگ‌ترین w مراجعه می‌شود. این مقدار با توجه به شکل (۹)، ۱/۷ بوده که بین دو گره ۵ و ۷ وجود دارد. با توجه به اینکه این دو گره عضو هیچ خوشه‌ای نیستند، به‌عنوان خوشه جدید در نظر گرفته می‌شوند. برای ادامه کار مشابه قبل وزن یال مربوط به گره‌هایی که به این خوشه متصل هستند، با توجه به رابطه (۷) به‌روز می‌شوند. نمایی از آن در شکل (۹) مشاهده می‌شود.



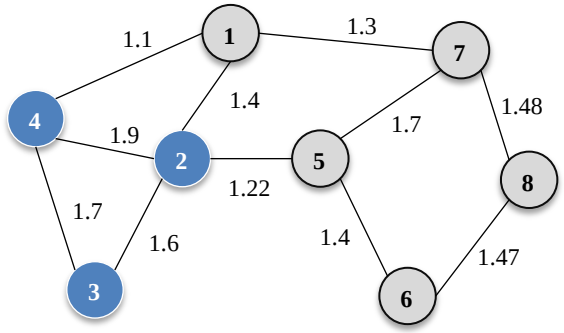
شکل ۹. در نظر گرفتن گره‌های ۵ و ۷ به‌عنوان یک گره و ادغام گره‌های متصل به آن

در ادامه بیش‌ترین مقدار، ۱/۶۵ بوده که مربوط به ارتباط بین گره ۳ با خوشه اول است. بنابراین، مطابق با شکل (۱۰) گره ۳ به خوشه یک اضافه می‌شود.



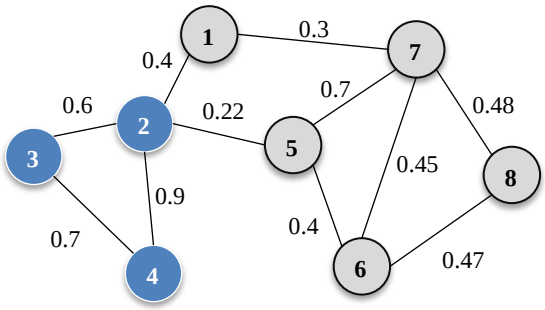
شکل ۱۰. اضافه شدن گره ۳۳ به خوشه اول (متشکل از گره‌های ۲ و ۴)

در ادامه نیز به همین ترتیب به‌صورت سلسله مراتبی یال‌های با وزن بالا محاسبه‌شده و به خوشه‌ها اضافه می‌شوند تا زمانی که تعداد خوشه‌ها به تعدادی مدنظر کاربر برسد. ازآنجاکه گره ۱ بین دو خوشه قرار دارد، به خوشه‌ای تعلق خواهد داشت که با یال با وزن بیشتر به آن متصل است. خوشه‌بندی نهایی گراف ساختاری-محتوایی تشکیل‌شده از گراف مثال موردنظر را شکل (۱۱) نشان می‌دهد.



شکل ۱۱. خوشه‌های نهایی گراف ساختاری-محتوایی

درنهایت خوشه‌بندی ساختاری-محتوایی گراف اولیه به‌صورت شکل (۱۲) انجام شده است. به‌منظور ارزیابی روش ارائه‌شده معیارهای موجود بر روی خوشه‌های گراف اولیه محاسبه می‌شود تا بتوان روش ارائه‌شده را با سایر روش‌های خوشه‌بندی ساختاری-محتوایی مقایسه نمود.



شکل ۱۲. خوشه‌های نهایی در گراف وزن‌دار اولیه

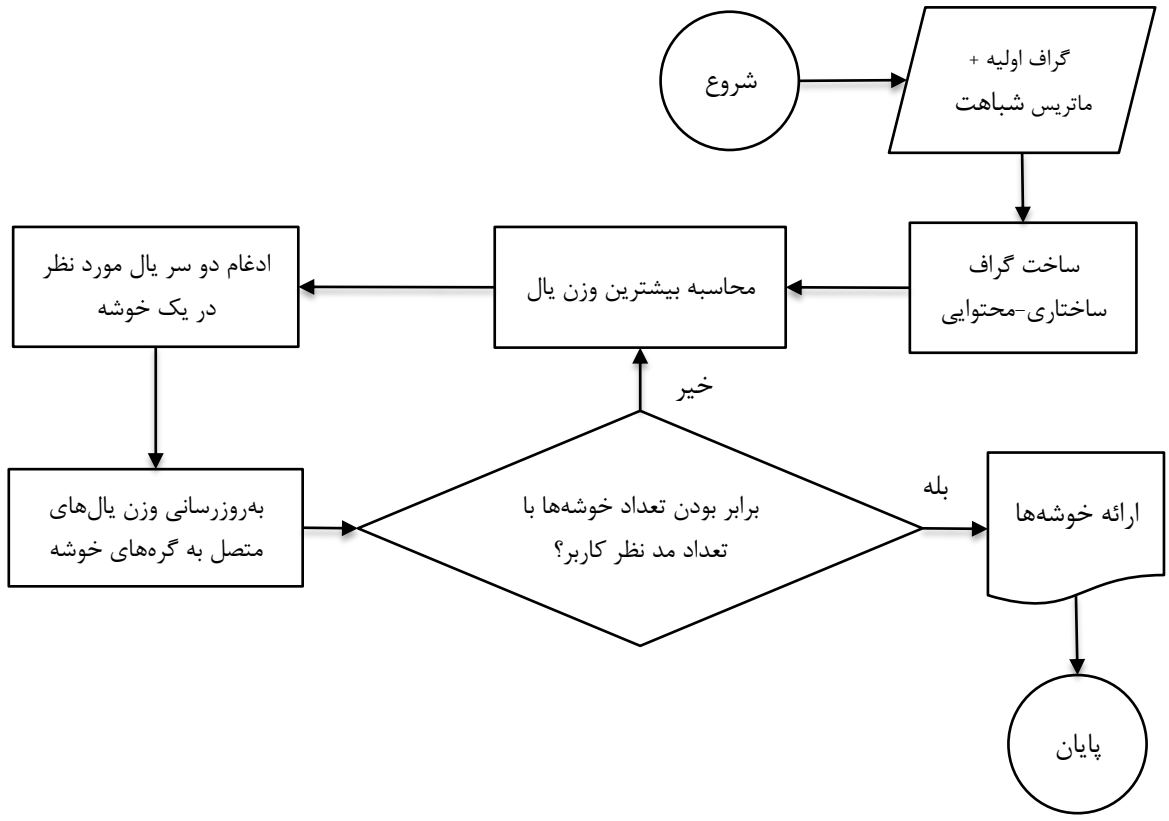
در شکل (۱۳) فلوچارت الگوریتم ارائه شده نشان داده شده است. ورودی الگوریتم ICS-Cluster شامل گراف ساختاری و ماتریس شباهت محتوایی یال‌های آن است. در ادامه گراف ساختاری-محتوایی بر اساس رابطه که درروش ارائه‌شده تولید می‌شود، که در آن وزن گره‌ها، بیانگر شباهت ساختاری-محتوایی بیش‌ازحد آستانه مشخصی است.

در ادامه بیش‌ترین وزن یال که نشان‌دهنده حداکثر شباهت ساختاری-محتوایی بین دو گره است را یافته و آن دو به‌عنوان یک خوشه در نظر گرفته می‌شوند. پس‌ازآن وزن‌های متصل به گره‌های موجود در خوشه به‌روزرسانی می‌شوند، بدین معنا که

تمام گره‌هایی که از یک یال به خوشه‌ای متصل هستند، به یک یال تبدیل‌شده و از وزن آن‌ها میانگین گرفته می‌شود و به‌عنوان وزن ارتباط آن گره با خوشه در نظر گرفته می‌شود.

پس از طی مراحل فوق، بررسی می‌شود که تعداد خوشه موجود به تعداد مدنظر کاربر رسیده یا خیر، درصورتی‌که تعداد برابر نبود، مجدداً حداکثر وزن یال محاسبه‌شده و خوشه‌بندی تا زمانی که تعداد خوشه‌ها به تعداد مدنظر کاربر برسد، ادامه میابد.

در ادامه کیفیت خوشه‌های تولیدشده در روش ارائه‌شده (ICS-Cluster) با روش‌های خوشه‌بندی ساختاری-محتوایی با پیاده‌سازی این روش‌ها بر روی یک مجموعه داده، مورد ارزیابی قرار گرفته است.



شکل ۱۳. فلوچارت روش ICS-Cluster

۴. ارزیابی

جهت ارزیابی روش ارائه‌شده 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 ۴ هسته‌ای GH۲,۲۰ و حافظه اصلی ۸ گیگابایتی انجام‌شده است. به‌منظور پیاده‌سازی الگوریتم‌ها از زبان جاوا در محیط اکلیپس[1] 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.

روش‌های مورد مقایسه دارای پارامترهای متفاوتی هستند. برای مقایسه این روش‌ها محدودیت‌هایی وجود دارد. به‌عنوان‌مثال درروش SA-Cluster و روش ICS-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[2]

در آزمایش‌های انجام‌شده از مجموعه داده Delicious استفاده شده است[3]. این مجموعه داده دارای ۱۸۶۱ گره و ۷۶۶۴ یال است. خصوصیت‌های هر گره (که تعداد آن‌ها ۱۳۵۰ عدد است)

[1] Eclipse

[2] Dataset

[3] https://rdrr.io/cran/mldr.datasets/man/delicious.html

نیز به‌وسیله یک آرایه دودویی مشخص‌شده 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. درواقع برای هر گره یک آرایه ۱۳۵۰ خانه‌ای وجود دارد که برای هر ویژگی یک خانه در نظر گرفته شده است. اگر گره دارای ویژگی موردنظر بود، خانه مربوطه 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 به‌صورت دودویی ذخیره‌شده تا هنگام محاسبه شباهت، بتوان از شباهت جاکارد به‌راحتی استفاده نمود. لازم به ذکر است که در روش ICS-Cluster پس از تبدیل گراف اولیه به گراف ساختاری-محتوایی تعداد یال‌های گراف جدید از ۷۶۶۴ به ۸۷۳۹۹۴ یال می‌رسد که دلیل آن تأثیر شباهت محتوایی در کنار شباهت ساختاری است.

۴-۲. معیارهای بررسی‌شده

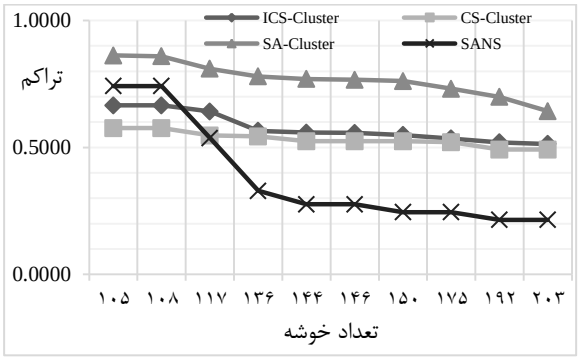
مقایسه های انجام شده بر اساس معیارهای مختلفی صورت پذیرفته که در ادامه تشریح می‌شوند.

۴-۲-۱. تراکم

این معیار که خوشه‌بندی را ازنظر ساختاری ارزیابی می‌کند، از طریق رابطه (۸) محاسبه می‌شود [۱۸]:

$$Density = \frac{\sum_{\mathcal{C}\in C} E_{\mathcal{C}}}{E} \qquad (۸)$$

که در آن، C مجموعه خوشه‌های حاصل از خوشه‌بندی، $E_{\mathcal{C}}$ یال‌های موجود در خوشه $\mathcal{C}$ و E مجموعه کل یال‌های گراف را نشان می‌دهد. شکل (۱۴) مقایسه تراکم خوشه‌های حاصل از خوشه‌بندی چهار روش را نشان می‌دهد. در روش SA-Cluster مراکز خوشه‌ها در نقاط متراکم گراف انتخاب می‌شود و سپس گره‌های همسایه این گره‌های مرکزی به هر خوشه تخصیص داده می‌شود. این عمل باعث می‌شود درنهایت خوشه‌ها تراکم بالایی داشته باشند. در روش ICS-Cluster نیز با توجه به اینکه ابتدا گراف اولیه به گرافی با تعداد یال بیشتر که بیانگر شباهت محتوایی و ساختاری است تبدیل شده و همچنین خوشه‌ها به‌صورت سلسله مراتبی از پایین به بالا تشکیل شده و درنهایت ارزیابی بر روی خوشه نهایی بر روی گراف اولیه انجام می‌شود، این معیار ساختاری برای این روش دارای مقدار مناسبی است. روش SANS تراکم کمی را به‌دست می‌آورد. علت این تراکم پایین در این روش این است که بعد از تعیین مراکز خوشه‌ها، گره های مشابه در هر نقطه‌ای از گراف به خوشه اضافه می‌شوند و این باعث می‌شود درنهایت تراکم کلی خوشه‌ها پایین باشد.



شکل ۱۴. مقایسه تراکم روش ICS-Cluster با سایر روش‌ها

۴-۲-۲. خطای یال

با توجه به اینکه از جنبه ساختاری، خوشه‌ای ایدئال 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 [۳۱]. این معیار نیز یک معیار ساختاری است و خوشه‌بندی را ازنظر ساختاری مورد ارزیابی قرار می‌دهد. خطای یال شامل تعداد یال‌هایی است که به‌اشتباه حذف شده‌اند، به تعبیری دیگر تعداد یال‌هایی که وجود دارند درصورتی‌که بهتر بود حذف شوند. برای ارزیابی این معیار دو نوع خطا تعریف می‌شود، اولین نوع خطا که از طریق رابطه (۹) محاسبه می‌شود، تعداد یال‌های ناموجود بین گره‌ها در یک جامعه c را تعیین می‌کند [۳۱].

$$\epsilon_{\text{within}}(C, E) = \{(v, w) | v, w \in C \wedge v \neq w \wedge (v, w) \notin E\} \qquad (۹)$$

خطای نوع دوم تعداد یال‌های بین جامعه C و جوامع دیگر را بر اساس رابطه (۱۰) تعیین می‌کند.

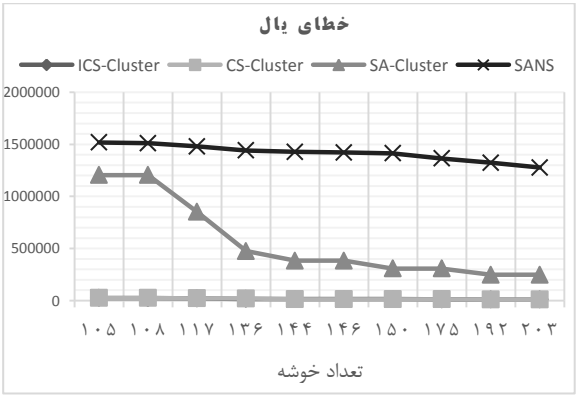
$$\epsilon_{\text{between}}(C, E) = |\{(v, w) \in E | v \in C \wedge w \notin C\}| \qquad (۱۰)$$

این دو خطا روی‌هم‌رفته یک معیار خطای کلی را ارائه می‌دهند که به‌صورت رابطه (۱۱) تعریف می‌شود:

$$\epsilon(\mathcal{C}, G) = \sum_{C\in\gamma\mathcal{C}} \epsilon_{\text{within}}(C, E) + \frac{\epsilon_{\text{between}}(C,E)}{2} \qquad (۱۱)$$

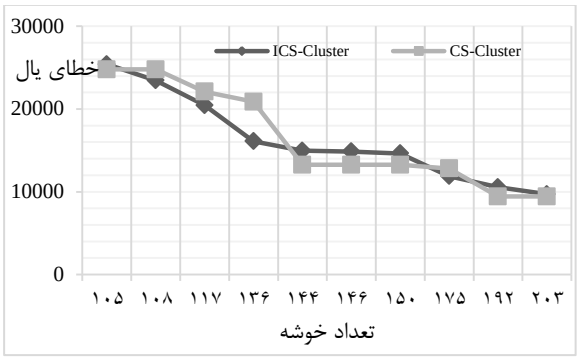
افراز پایه گراف حالتی است که در آن هر گره به‌صورت جداگانه به‌عنوان یک خوشه در نظر گرفته می‌شود. درروش ICS-Cluster، ابتدا یال با وزن بالا (مهم) انتخاب شده، سپس دوسر آن به‌عنوان یک خوشه در نظر گرفته می‌شود، در توسعه آن نیز گره‌هایی که با وزن بالا به خوشه متصل هستند نیز به آن اضافه می‌شوند، بنابراین، در این روش، خوشه‌ها دارای تعداد اتصالات داخلی بالا و تعداد اتصال خارجی کم هستند. درروش CS-Cluster نیز در هنگام انتخاب مراکز خوشه‌ها و هنگام توسعه هر خوشه وضعیت اتصال گره‌ها در نظر گرفته می‌شود. درنتیجه درنهایت خوشه‌ها در این دو حالت دارای کمترین خطای یال هستند. همان‌طور که در شکل (۱۵) مشاهده می‌شود، بعد از آن‌ها روش SANS، به نسبت خطای یال مناسبی دارد، ولی SA-Cluster بیش‌ترین خطای یال را دارد. در روش SA-Cluster چون بعد از انتخاب مراکز خوشه‌ها، گره‌های مشابه مرکز خوشه در هر قسمت از گراف به خوشه مربوط به آن مرکز خوشه تخصیص داده می‌شوند و این گره‌ها گاهی در فاصله خیلی دور و دارای اتصال کمی است، درنتیجه درنهایت میزان خطای یال بسیار زیادی خواهد داشت.

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. درصورتی که تأثیر وجود این دو یال باید متفاوت باشد.



شکل ۱۵. مقایسه خطای یال روش ICS-Cluster با سایر روش‌ها

در شکل (۱۵)، به دلیل بالا بودن خطای یال روش‌های SANS و SA-Cluster، خطای یال روش‌های CS-Cluster و ICS-Cluster به‌خوبی نشان داده نشده است، به همین دلیل در شکل (۱۶) تنها خطای یال این دو روش آمده است تا اختلاف آن‌ها به‌خوبی نشان داده شود.



شکل ۱۶. مقایسه خطای یال روش‌های ICS-Cluster و CS-Cluster

۴-۲-۳. میانگین شباهت[1]

این معیار خوشه‌بندی را ازنظر محتوایی مورد ارزیابی قرار می‌دهد که میانگین شباهت خوشه‌های گراف را مطابق با رابطه (۱۲) محاسبه می‌کند.

$$Avg\_Sim\_Graph = \frac{\sum_{c \in C} c_{avgsim}}{|C|} \qquad (۱۲)$$

که در آن C مجموعه خوشه‌های حاصل از خوشه‌بندی، c_{avgsim} میانگین شباهت درون خوشه‌ای خوشه c و |C| تعداد خوشه‌های حاصل از خوشه‌بندی را نشان می‌دهد. درواقع میانگین شباهت، مجموع میانگین شباهت داخلی خوشه‌ها به روی تعداد خوشه‌ها است. همان‌طور که در روش ICS-Cluster توضیح داده شد، در مرحله اولیه این الگوریتم، ابتدا گراف به یک گراف ساختاری-محتوایی تبدیل می‌شود که در آن نقش شباهت محتوایی دیده شده است. در روش CS-Cluster در هنگام تعیین مراکز خوشه‌ها و همچنین در زمان توسعه خوشه‌ها میزان شباهت گره‌ها در نظر گرفته می‌شود. درنتیجه همان‌طور که در نمودار شکل (۱۷) مشاهده می‌شود این دو روش بیش‌ترین میانگین شباهت را دارند.

شکل ۱۷. مقایسه میانگین شباهت روش ICS-Cluster با سایر روش‌ها

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

[1] Average Similarity

معیارهای ساختاری معیار تراکم، پیمانگی، میانگین درجه، نسبت برش، هدایت، خطای یال را می‌توان نام برد [۱۳]. برخی معیارهای کارایی نیز وجود دارد که محتوای گره‌ها را مدنظر دارند ازجمله فراخوانی، دقت و صحت که البته این دو معیار در مورد خوشه‌بندی‌هایی که برای هر خوشه یک توصیف مشخص و خاص مدنظر است کاربرد دارد [۳۵]. معیار CS-Measure هر دو جنبه ساختاری و محتوایی را با یکدیگر ترکیب می‌کند.

همان‌طور که گفته شد، خطای یال خوشه به‌صورت رابطه‌های (۱۱-۹) تعریف می‌شود. چون خطای یال‌های خارجی برای هر دو خوشه‌ای که شامل گره‌های مبدأ و مقصد یال هستند محاسبه می‌شود، خطای یال خارجی تقسیم‌بر ۲ می‌شود [۳۴]. 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 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 ایدئال 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.

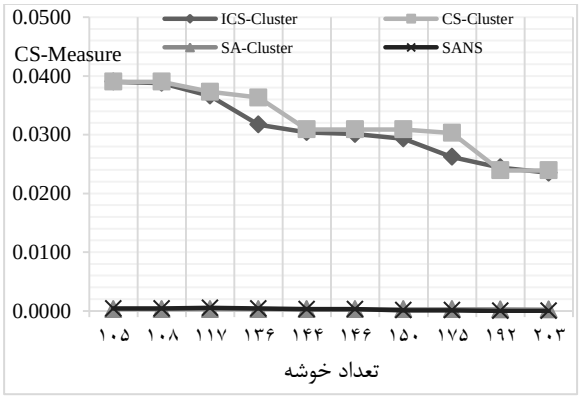
اگر وزن یال‌های خروجی زیاد باشد این کاهش امتیاز زیاد خواهد بود و اگر وزن یال‌های خروجی کم باشد این کاهش امتیاز کم‌تر خواهد بود. وزن یال‌های ناموجود برابر صفر در نظر گرفته می‌شود. در ادامه میانگین وزن یال‌ها (شامل یال‌های موجود در خوشه، یال‌های ناموجود و یال‌های خروجی خوشه) محاسبه‌شده و به‌عنوان امتیاز خوشه در نظر گرفته می‌شود. پس از محاسبه امتیاز هر خوشه، میانگین امتیاز خوشه‌ها به‌عنوان امتیاز کلی خوشه‌بندی مطابق با رابطه‌های (۱۳) و (۱۴) محاسبه می‌شود [۳۲].

$$\mathrm{EScore}(\mathcal{C}) = \frac{\sum_{e\in E_{in\_\mathcal{C}}} we + \frac{\sum_{e\in E_{ex\_\mathcal{C}}}(-we)}{2}}{|E_{in\_\mathcal{C}}|+|E_{ex\_\mathcal{C}}|} \qquad (۱۳)$$

$$\mathrm{CS-Measure} = \frac{\sum_{\mathcal{C}\in C} \mathrm{EScore}(\mathcal{C})}{|C|} \qquad (۱۴)$$

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

همان‌طور که توضیح داده شد در این معیار، وزن یال‌های خارجی در (۱-) ضرب شده و در محاسبه میانگین شباهت در نظر گرفته می‌شود. درنتیجه 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، درنتیجه نسبت به معیار خطای یال دقیق‌تر عمل می‌کند. با توجه به اینکه این معیار، شباهت ساختاری و محتوایی را به‌طور هم‌زمان در نظر می‌گیرد، معیار مطلوب‌تری برای ارزیابی کیفیت خوشه‌های نهایی است؛ بنابراین، با توجه به شکل (۱۸)، می‌توان گفت روش CS-Measure نسبت به سایر روش‌ها درمجموع بهتر عمل می‌کند. با توجه به اینکه در مرحله اول الگوریتم روش ICS-Cluster گراف ایجاد شده گراف ساختاری-محتوایی بوده و در ایجاد و توسعه خوشه‌ها هم‌زمان هر دو جنبه در نظر گرفته می‌شود، این روش نیز دارای مقادیر مناسبی بر اساس این معیار است، درحالی‌که روش SA-Cluster و SANS به لحاظ این معیار دارای مقدار کمی هستند.



شکل ۱۸. مقایسه CS-Measure روش ICS-Cluster با سایر روش‌ها

۴-۲-۵. پیمانگی

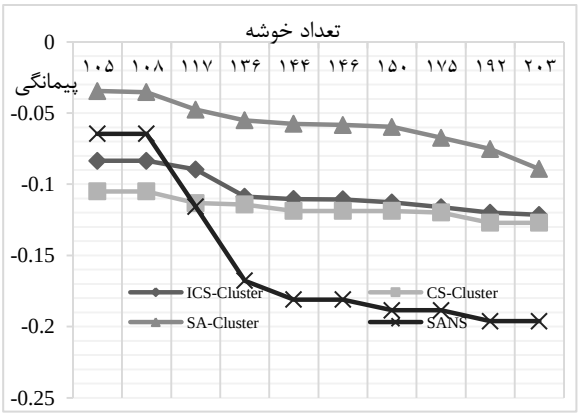
این معیار که درواقع خوشه‌بندی را ازنظر ساختاری ارزیابی می‌کند به‌صورت رابطه (۱۵) محاسبه می‌شود [۳۶]:

$$Q = \frac{1}{4m}\sum_s (m_S - E(m_S)) \qquad (۱۵)$$

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 این معیار از طریق مجموع تفاضل یال‌های موجود در خوشه‌ها به یال‌های ممکن آن به روی مجموع یال‌های موجود در گراف محاسبه می‌شود.

همان‌طور که در شکل (۱۹) مشاهده می‌شود، با توجه به اینکه در روش SA-Cluster مراکز خوشه‌ها در نقاط متراکم گراف انتخاب می‌شود و سپس گره‌های همسایه این گره‌های مرکزی به هر خوشه تخصیص داده می‌شود. این عمل باعث می‌شود درنهایت خوشه‌ها پیمانگی بالایی داشته باشند.

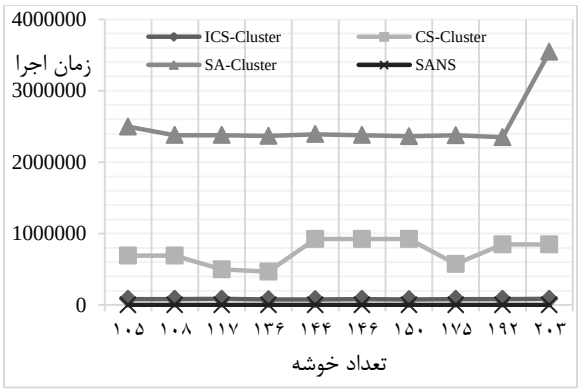
روش ICS-Cluster نسبت به روش CS-Cluster دارای پیمانگی بالاتری بوده و روش SANS نیز کمترین پیمانگی را دارد. علت پیمانگی پایین در روش SANS این است که بعد از تعیین مراکز خوشه‌ها، گره‌های مشابه در هر نقطه‌ای از گراف به خوشه اضافه می‌شوند و این باعث می‌شود درنهایت پیمانگی کلی خوشه‌ها پایین باشد.



شکل ۱۸. مقایسه پیمانگی روش ICS-Cluster با سایر روش‌ها

۴-۲-۶. زمان اجرا

با استفاده از این معیار مدت‌زمان خوشه‌بندی محاسبه می‌شود. با توجه به نمودار حاصل از آزمایش‌های انجام‌شده که در شکل (۲۰) مشاهده می‌شود؛ از بین روش‌های مقایسه شده، روش‌های SANS و ICS-Cluster نسبت به سایر روش‌ها سریع‌تر هستند. پس از آن‌ها، روش CS-Cluster دارای زمان اجرای قابل قبولی است. روش SA-Cluster به دلیل تکرارشوندگی، نسبت به بقیه کندتر عمل خوشه‌بندی را انجام می‌دهد.

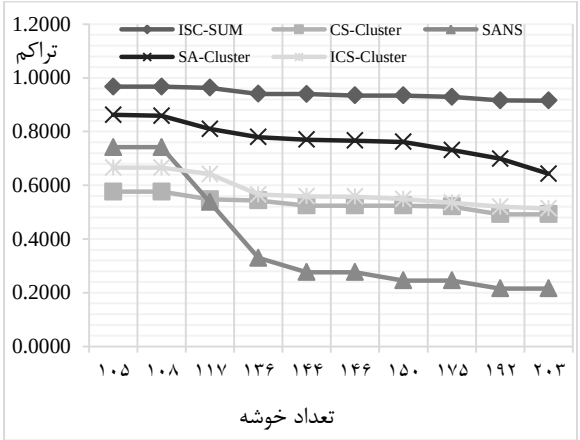


شکل ۲۰. مقایسه زمان اجرای روش ICS-Cluster با سایر روش‌ها (میلی‌ثانیه)

با توجه به تحقیقات صورت گرفته، با تغییر رابطه (۷) در بخش چهار، برای برخی از معیارهای محاسبه‌شده بهبودهایی حاصل شد. در رابطه (۷) از میانگین حسابی برای ادغام گره‌ها استفاده شده است، درحالی‌که اگر به‌جای آن از جمع یال‌های ادغام شونده استفاده شود، نتایج پیمانگی و تراکم نسبت به محاسبه میانگین حسابی بهبود خواهند یافت. بدین معنا که مثلاً اگر دو رأس قرار است با یکدیگر ادغام شوند، وزن یال جدید مطابق با رابطه (۱۶) حاصل جمع وزن دو یال قبلی است.

$$W_{new} = \sum_{i=1}^{n} W_i \qquad (۱۶)$$

که در آن، W_{new} وزن یال ادغام شده است، n تعداد یال‌هایی است که ادغام شده و W_i وزن هرکدام از آن‌ها است. همان‌طور که در شکل (۲۱) مشاهده می‌شود، این روش با نام ICS-SUM نشان داده شده و دارای تراکم بهتری نسبت به سایر روش‌ها است.

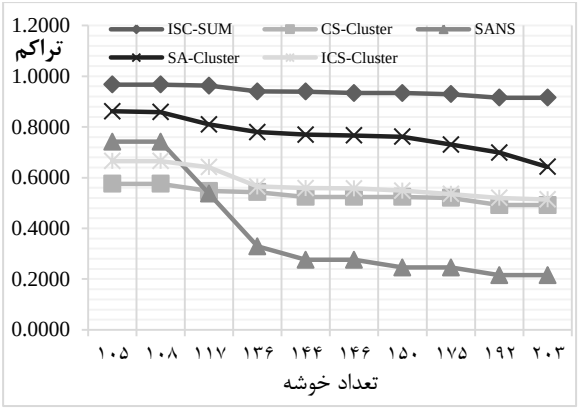


شکل ۲۱. مقایسه تراکم روش‌های خوشه‌بندی بر اساس رابطه (۱۶) (ICS-SUM)

مقایسه خاصیت پیمانگی برای الگوریتم‌ها در شکل (۲۲) نشان داده شده است که در آن روش ICS-SUM نسبت به سایرین دارای خاصیت پیمانگی بالاتری است. درواقع نتایج نشان می‌دهد که اگر در خوشه‌بندی نیاز به بالا بودن دو معیار تراکم و

پیمانگی بود، بهتر است به‌جای میـانگین حسـابی از رابطـه (۱۶) استفاده کرد.

کاربرد زیادی در عصر ارتباط دارد که بـا افـزایش اسـتفاده از اینترنت حجم گراف‌های تشکیل‌شده روبه افزایش است، به همین دلیل خوشه‌بندی این گراف‌ها از اهمیت بالایی برخوردار است.



شکل ۲۲. مقایسه پیمانگی روش‌های خوشه‌بندی بر اساس رابطه (ICS-SUM) (۱۶)

بیشتر روش‌های خوشه‌بندی که تاکنون ارائه شده است فقـط یک جنبه ساختاری یا محتوایی را در نظر گرفته‌انـد و روش‌هـای خوشه‌بندی که ساختار و محتوا را باهم نظر بگیرند کمتر ارائه‌شده است. در این مقاله روش ICS-Cluster ارائه شد که خوشه‌بندی را بر اساس ساختار و محتوا انجام می‌دهد. در این روش ابتدا گـراف اولیه به یک گراف ساختاری-محتوایی تبدیل شده، سـپس عمـل خوشه‌بندی بر اساس یال‌هـای مهـم آن گـراف انجـام مـی‌شـود. کارایی روش ارائه‌شده در مقاله با سـایر روش‌هـای خوشـه‌بنـدی ساختاری-محتوایی ارائه‌شده مورد ارزیابی قرار گرفـت کـه نتـایج کلی آن در جـدول (۱) آمـده اسـت. همـان‌طـور کـه گفتـه شـد خوشه‌بندی که بین ساختار و محتوا تعادل خوبی برقرار کنـد، بـه لحاظ خوشه‌بندی ساختاری-محتوایی عملکرد خوبی از خود نشان داده است. لذا روش خوشه‌بندی که بتواند به لحاظ تمامی معیارها (ساختاری، محتوایی و ساختاری-محتوایی) نتایج مناسـبی ارائـه دهد، می‌توان گفت که بهترین روش است.

جدول ۱. مقایسه کلی روش ICS-Cluster با سایر روش‌ها بر اساس معیارهای مختلف ارائه شه در این مقاله

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

گراف کاربرد زیادی در عصر ارتباط دارد که با افزایش اسـتفاده از اینترنت حجم گراف‌های تشـکیل شـده روبـه افـزایش اسـت، بـه همین دلیل خوشه بندی این گراف ها از اهمیت بـالایی برخـوردار است. بیشتر روش‌های خوشه‌بندی که تاکنون ارائه‌شده است فقط یک جنبه ساختاری یا محتوایی را در نظر گرفته‌انـد و روش‌هـای خوشه‌بندی که ساختار و محتوا را باهم نظر بگیرند کمتر ارائه‌شده است. در این مقاله روش ICS-Cluster ارائه شد که خوشه‌بندی را بر اساس ساختار و محتوا انجام می‌دهد. در این روش ابتدا گـراف اولیه به یک گراف ساختاری-محتوایی تبدیل شده، سـپس عمـل خوشه‌بندی بر اساس یال‌هـای مهـم آن گـراف انجـام مـی‌شـود. کارایی روش ارائه‌شده در مقاله با سـایر روش‌هـای خوشـه‌بنـدی ساختاری-محتوایی ارائه‌شده مورد ارزیابی قرار گرفـت کـه نتـایج کلی آن در جـدول (۱) آمـده اسـت. همـان‌طـور کـه گفتـه شـد خوشه‌بندی که بین ساختار و محتوا تعادل خوبی برقرار کنـد، بـه لحاظ خوشه‌بندی ساختاری-محتوایی عملکرد خوبی از خود نشان داده است. لذا روش خوشه‌بندی که بتواند به لحاظ تمامی معیارها (ساختاری، محتوایی و ساختاری-محتوایی) نتایج مناسـبی ارائـه دهد، می‌توان گفت که بهترین روش است.

۵. نتیجه‌گیری

نتایج ارائه‌شده در مقایسـه‌هـای انجـام‌شـده نشـان مـی‌دهـد کـه الگوریتم ICS-Cluster به لحاظ ساختاری و محتوایی دارای نتایج متعـادلی اسـت، بـدین معنـا کـه بـرخلاف روش‌هـای SANS و SA-Cluster که در آن‌ها در برخی از معیارها جواب مناسبی ارائه شده و در برخی دیگر جواب نامناسـبی ارائـه مـی‌شـود، در روش ICS-Cluster پاسخ‌هـا بـه نسـبت خـوب و متعـادل اسـت. روش SA-Cluster بر اساس معیارهای سـاختاری عملکـرد خـوبی دارد برخلاف روش CS-Cluster که به لحاظ محتوایی عملکرد مناسبی از خود نشان می‌دهد. در کـل تنهـا مزیـت روش SANS سـرعت اجرایی آن است. روش ICS-Cluster به لحاظ معیارهای ساختاری

و محتوایی پاسخ مناسبی ارائه می‌دهد. این تحقیق ضمن ارائه روش خوشه‌بندی ساختاری-محتوایی جدید، منجر به ارائه نکاتی شد که حائز اهمیت هستند:

- تبدیل گراف اولیه به یک گراف ساختاری-محتوایی که وزن یال‌های آن بیانگر شباهت ساختار و محتوایی است.
- دادن وزن موردنظر کاربر به ساختار یا محتوا با توجه به اهمیت هرکدام نسبت به دیگری که منجر به افزایش کیفیت خوشه‌های نهایی مدنظر کاربر می‌شود.
- خوشه‌بندی بر اساس یال‌های حساس (با توجه به اینکه در سایر روش‌ها بر اساس گره حساس خوشه‌بندی انجام می‌شود.)

در تحقیقات معیارهای جدیدی به‌منظور ارزیابی کیفیت خوشه‌ها در گراف ارائه می‌شود [۳۷]، اما تنها یک معیار ساختاری-محتوایی برای ارزیابی خوشه‌ها وجود دارد [۳۲]، برای کارهای آینده پیشنهاد می‌شود معیارهای ارزیابی بیشتر و بهتری برای ارزیابی کیفیت و کارایی این روش‌های خوشه‌بندی پیشنهاد داد تا ارزیابی‌های دقیق‌تر و کامل‌تری بر روی روش‌های خوشه بندی انجام شود. درزمینه خوشه‌بندی ساختاری-محتوایی گراف می‌توان روش‌هایی را پیشنهاد داد که این عمل را با سرعت و دقت بالاتری انجام دهند.

۶. مرجع‌ها

[1] Sambhoos, K.; Nagi, R.; Sudit, M.; Stotz, A. "Enhancements to High Level Data Fusion Using Graph Matching and State Space Search"; Inform. Fusion 2010, 11, 4, 351-364.

[2] Zhou, L.; Wang, Q.; Fujita, H. "One Versus One Multi-Class Classification Fusion Using Optimizing Decision Directed Acyclic Graph for Predicting Listing Status of Companies"; Inform. Fusion 2017, 36, 80-89.

[3] Gross, G. A.; Nagi, R. "Precedence Tree Guided Search for the Efficient Identification of Multiple Situations of Interest – AND/OR Graph Matching"; Inform. Fusion 2016, 27, 240-254.

[4] Guedes, G. P.; Ogasawara, E.; Bezerra, E.; Xexeo, G. "Discovering Top-Non-Redundant Clustering in Attributed Graphs"; Neurocomputing 2016, 210, 45-54.

[5] Schoch, D.; Valente, T. W.; Brandes, U. "Correlations Among Centrality Indices and a Class of Uniquely Ranked Graphs"; Soc. Networks 50, 46-54.

[6] Johnson, M.; Paulusma, D.; Leeuwen, E. J. V. "Algorithms for Diversity and Clustering in Social Networks Through Dot Product Graphs"; Soc. Networks 2015, 41, 48-55.

[7] Ertem, Z.; Veremyev, A.; Butenko, S. "Detecting Large Cohesive Subgroups with High Clustering Coefficients in Social Networks"; Soc. Networks 2016, 46, 1-10.

[8] Vörös, A.; Snijders, T. A. B. "Cluster Analysis of Multiplex Networks: Defining Composite Network Measures"; Soc. Networks 2017, 49, 93-112.

[9] Opsahl, T.; Panzarasa, P. "Clustering in Weighted Networks"; Soc. Networks 2009, 31, 155-163.

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

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

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

[13] Newman, M. "Community Detection in Networks: Modularity Optimization and Maximum Likelihood Are Equivalent"; Phys. Rev. 2016, 94, 1-8.

[14] Bhatia, V.; Rani, R. "A Parallel Fuzzy Clustering Algorithm for Large Graphs Using PREGEL"; Expert Syst. Appl. 2017, 78, 135-144.

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

[16] Khatoon, M.; Banu, A. "A Survey on Community Detection Methods In Social Networks"; IJEME. 2015, 1, 8-18.

[17] Elhadi, H.; Agam, G. "Structure and Attributes Community Detection: Comparative Analysis of Composite Ensemble and Selection Methods"; Proc. 7th Workshop on Social Network Mining and Analysis 2013, 1-7.

[18] 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"; Computation Stat. 2014, 6, 426-439.

[19] Bu, Z.; Gao, G.; Li, H. J.; Cao, J. "CAMAS: A Cluster-Aware Multiagent System for Attributed Graph Clustering"; Inform. Fusion 2017, 37, 10-21.

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

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

[22] Skabar, A. "Clustering Mixed-Attribute Data Using Random Walk"; Procedia Comput. Sci. 2017, 108, 988-997.

[23] Xu, Z.; Cheng, J.; Xiao, X.; Fujimaki, R.; Muraoka, Y. "Efficient Nonparametric and Asymptotic Bayesian Model Selection Methods for Attributed Graph Clustering"; Knowl. Inform. Syst. 2017, 53, 239-268.

[24] Gross, G.; Nagi, R.; Sambhoos, K. "A Fuzzy Graph Matching Approach in Intelligence Analysis and Maintenance of Continuous Situational Awareness"; Inform. Fusion 2014, 18, 43-61.

[25] Boobalan, M. P.; Lopez, D.; Gao, X. Z. "Graph Clustering Using K-Neighbourhood Attribute Structural Similarity"; Appl. Soft Comput. 2016, 47, 216-223.

[26] Zhou, H.; Li, J.; Li, J.; Zhang, F.; Cui, Y. "A Graph Clustering Method for Community Detection in Complex Networks"; Physica A 2017, 469, 551-562.

[27] Bai, L.; Cheng, X.; Liang, J.; Guo, Y. "Fast Graph Clustering with a New Description Model for Community Detection"; Inform. Sci. 2017, 388, 37-47.

[28] Ruan, Y.; Fuhry, D.; Parthasarathy, S. "Efficient Community Detection in Large Networks Using Content and Links"; Proc. 22nd Int. Conf. World Wide Web WWW '13). ACM, New York, NY, USA, 2013, 1089-1098.

[29] Zhou, Y.; Cheng, H.; Xu, Y. J. "Graph Clustering Based on Structural/Attribute Similarities"; Proc. VLDB Endowment 2009, 2, 718-729.

[30] Parimala, M.; Daphne, L. "Graph Clustering Based on Structural Attribute Neighborhood Similarity (SANS)"; IEEE Int. Conf. Elec. Comput. Commun. Technol. 2015, 1-5.

[31] Pool, S.; Bonchi, F.; Leeuwen, M. "Description-Driven Community Detection"; ACM Trans. Intel. Syst. Technol. 2014, 03, 201-210.

[32] Rahmati, K.; Naderi, H.; Keshvari, S. "Content-Structural Graph Clustering and a New Measure For Its Evaluation"; Adv. Defence Sci. Technol. 2018, 03, 201-210. (In Persian)

[33] Halloush, R. "Overhearing-Aware Modified Dijkstra's Algorithm for Multicasting over Multi-Hop Wireless Networks"; Int. J. Commun. Netw. Distrib. Syst. 2016, 16, 240-260.

[34] Pool, S.; Bonchi, F.; Leeuwen, M. "Description-Driven Community Detection"; ACM Trans. Intel. Syst. Technol. 2014, 5, 28.

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

[36] Yang, J.; Leskovec, J. "Defining and Evaluating Network Communities Based on Ground-Truth"; Knowl. Inf. Syst. 2015, 42, 181-213.

[37] Biswas, A.; Biswas, B. "Defining Quality Metrics for Graph Clustering Evaluation"; Expert Syst. Appl. 2017, 71, 1-17.