تشخیص جوامع در شبکه‌های اجتماعی با استفاده از الگوریتم‌های فراابتکاری مبتنی بر خوشه‌بندی تجمعی

احد فخاری[۱]، موسی مجرد[۲،۳] و حسن ارفعی‌نیا[۴]

[۱]دانشجو کارشناسی ارشد، موسسه آموزش عالی غیرانتفاعی لیان بوشهر، بوشهر، ایران،fakhari.eng@gmail.com

[۲]استادیار، دانشکده مهندسی برق و کامپیوتر، واحد فیروزآباد، دانشگاه آزاد اسلامی، فیروزآباد، ایران، m.mojarad@iauf.ac.ir

[۳]باشگاه پژوهشگران جوان، واحد فیروزآباد، دانشگاه آزاد اسلامی، فیروزآباد، ایران

[۴]مربی، موسسه آموزش عالی غیرانتفاعی-غیر دولتی لیان بوشهر، بوشهر، ایران، msc.harfaeinia@gmail.com

۱- مقدمه

کشف انجمن در شبکه‌های اجتماعی یکی از موضوعات

چکیده: تحلیل روابط افراد در این شبکه‌ها اهمیت بسیار دارد. با تحلیل شبکه‌های اجتماعی می‌توان به روابط و ارتباطات بین افراد در شبکه‌ها دست یافت. تشخیص اجتماعات کاربران در شبکه‌های اجتماعی به مفهوم تشخیص گروه‌های کاربری می‌باشد. برای شناسایی گروه‌های کاربری باید کاربران را از نواحی مختلف شبکه به نحوی درنظر گرفت که دارای بیشترین چگالی با یکدیگر باشند. به‌طورکلی الگوریتم‌های تشخیص اجتماعات براساس روش‌های مختلف خوشه‌بندی ارائه می‌شوند. در این تحقیق روشی برای افزایش صحت و کیفیت تشخیص اجتماعات با استفاده از الگوریتم‌های ژنتیک، ازدحام ذرات، خوشه‌بندی ***k-means*** و خوشه‌بندی ***louvain*** ارائه شده است. برای این کار از ویژگی‌ها و اطلاعات توپولوژیکی و دموگرافیکی (محتوا) استفاده می‌شود. عملکرد روش پیشنهادی بر روی مجموعه داده شبکه اجتماعی خودمحور فیس‌بوک با الگوریتم‌های ***ACD-TC***، ***HICODE*** و ***TSC*** براساس معیارهای سیلوئت و ماژولاریتی ارزیابی و مقایسه شده است. نتایج برتری روش پیشنهادی را در اغلب آزمایش‌ها نشان می‌دهد.

کلیدواژه‌ها: *تشخیص جوامع، الگوریتم‌های فراابتکاری، خوشه‌بندی تجمعی، شبکه‌های اجتماعی.*

انجمن‌های اجتماعی یکی از انواع رسانه‌های اجتماعی محسوب می‌شوند و محلی برای شکل‌گیری جماعت‌های مجازی و انجمن‌سازی کاربران اینترنتی هستند [۱]. انجمن‌های اجتماعی موفق شده‌اند، تعداد قابل‌توجهی از کاربران اینترنتی را جذب کنند و در سال گذشته میلادی در میان چهار فعالیت اصلی کاربران اینترنتی قرار گرفتند. برخی تحلیلگران وب، آینده فضای مجازی را در گروی این نوع سایت‌ها می‌دانند. صدها سایت انجمن اجتماعی با حوزه‌های تخصصی متفاوت و همچنین با زبان‌های متنوع برای کاربران کشورهای مختلف در اینترنت فعالیت می‌کنند [۲]. کاربردهای انجمن‌های اجتماعی بسیار متنوع است و این سایت‌ها تمایل دارند حداکثر میزان نیاز آنلاین کاربران خود را برآورده کنند.

چالش‌برانگیز و مورد اهمیت در دنیای امروز می‌باشد، به این دلیل که از طریق شناسایی انجمن‌ها در این شبکه‌ها می‌توانیم سرعت ارتباطات افراد و انجمن‌ها را افزایش دهیم [۳]. در سال‌های اخیر تحقیقات زیادی بر روی الگوریتم-های تکاملی انجام گرفته است که از جمله این الگوریتم‌ها می‌توان به الگوریتم ژنتیک و ازدحام ذرات کلونی مورچگان اشاره کرد که سعی دارند میزان سرعت ارتباطات بین افراد در انجمن‌ها را بهبود ببخشند، اما متأسفانه به نتیجه مطلوبی نرسیده است.

تحلیل شبکه‌های اجتماعی باعث کشف اجتماعات و تعاملات بین کاربران می‌شود [۴]. یک اجتماع مجموعه‌ای از کاربران است که تراکم ارتباطات در آن گروه بالا است. اگرچه الگوریتم‌های بسیاری برای تشخیص اجتماعات ارائه

شده است، اما اغلب آن‌ها به لحاظ هزینه پردازشی و زمانی برای شبکه‌های اجتماعی با مقیاس بزرگ ناکارا هستند [۵]. در این تحقیق یک الگوریتم ساده و کارا برای تشخیص اجتماع در شبکه‌های اجتماعی ارائه می‌دهیم که به هیچ دانش قبلی درباره تعداد اجتماعات شبکه نیازی ندارد. بیشتر روش‌های موجود تشخیص اجتماعات به بررسی ساختار گراف شبکه اجتماعی بدون درنظرگرفتن موضوعات و تعاملات بین کاربران می‌پردازند [۶]. در روش پیشنهادی علاوه بر در نظر گرفتن توپولوژی ارتباطی بین کاربران، توییت‌های استفاده شده توسط آن‌ها نیز مورد توجه قرار می‌دهیم.

در شبکه‌های اجتماعی، روابط با اصطلاحات گره و یال بیان می‌شود. گره‌ها کاربران درون شبکه‌ها و یال‌ها روابط میان آن‌ها را نشان می‌دهد [۷-۹]. تحلیل شبکه‌های اجتماعی باعث کشف اجتماعات و تعاملات افراد می‌شود. یک اجتماع مجموعه‌ای از کاربران است که تراکم اتصالات در آن گروه بالا است. خوشه‌بندی یا تشخیص اجتماعات، ساختار گروه-ها در شبکه‌های اجتماعی و ارتباطات پنهان بین مؤلفه‌های آن را آشکار خواهد نمود. بیشتر روش‌های موجود تشخیص اجتماعات به بررسی ساختار گراف شبکه بدون در نظر گرفتن موضوعات می‌پردازند. در این تحقیق از اطلاعات توپولوژیکی و دموگرافیکی کاربران از یک شبکه اجتماعی نظیر فیس‌بوک، جهت شناسایی جوامع استفاده می‌شود.

در ادامه این تحقیق؛ پیشینه تحقیق در بخش دوم ارائه می‌شوند. بخش سوم مرتبط به روش پیشنهادی این تحقیق جهت کشف اجتماعات است و نتایج حاصل از آن در بخش چهارم ارائه می‌گردد. درنهایت در بخش پنجم نتیجه‌گیری و پیشنهادات مطرح می‌شود.

۲- پیشینه تحقیق

در زمینه تشخیص اجتماعات در شبکه‌های اجتماعی، تحقیقات متعددی صورت گرفته است. در این بخش به تحلیل و بررسی برخی از جدیدترین تحقیق‌های مرتبط در زمینه می‌پردازیم.

در [۱۰] تأثیر روش‌های کشف اجتماعات در شناسایی گره-های تأثیرگذار و بیشینه‌سازی اثر در شبکه‌های اجتماعی را مورد بررسی قرار دادند. در این تحقیق با توجه به مفروض بودن ساختار کلی شبکه اجتماعی و مدل انتشار، نه‌تنها بیشینه‌سازی تأثیر در کل شبکه، بلکه پوشش حداکثری اجتماعات در شبکه نیز موردتوجه قرار گرفته است.

در [۱۱] حداکثرسازی شیوع اطلاعات در شبکه‌های اجتماعی مبتنی بر اجتماعات را پیشنهاد دادند. در این تحقیق با استفاده از مدل انتشار اطلاعات آبشاری مستقل و روش‌های شناسایی اجتماعات مختلف، روشی در شناسایی گره‌های تأثیرگذار ارائه شده است.

در [۱۲] یک روش خوشه‌بندی فازی سلسله مراتبی جهت شناسایی ساختار اجتماعات در شبکه‌های اجتماعی ارائه شده است. در این تحقیق یک الگوریتم ترکیبی فازی در سه مرحله ارائه شده است که در مرحله اول سعی بر تشخیص اجتماعات روش خوشه‌بندی فازی دارد و در مرحله دوم اجتماعات را در سطوح بالاتر به روش خوشه-بندی سلسله مراتبی ترکیب کرده و در مرحله سوم با استفاده از معیار ارزیابی کیفیت بهینه‌سازی ماژولاریتی، اجتماعات به‌دست‌آمده از دو مرحله قبل ارزیابی می‌شود.

روش CC-GA در [۱۳] ارائه شد. این روش از یک الگوریتم ژنتیک مبتنی بر ضریب خوشه‌ای برای تشخیص جوامع در شبکه‌های اجتماعی استفاده می‌کند. روش CC-GA از لحاظ سیاست تولید جمعیت اولیه و همچنین عملگر جهش نسبت به سایر روش‌های مشابه برتری دارد که باعث افزایش کارایی و دقت آن شده است.

در [۱۴] الگوریتم انتشار برچسب براساس تأثیر مبتنی بر شواهد برای تشخیص جامعه توزیع شده در شبکه‌های اجتماعی را مورد بررسی قرار دادند. در این مقاله الگوریتم LPAجدیدی برای تشخیص جامعه توزیع‌شده براساس تئوری شواهد ارائه می‌شود که در بررسی اطلاعات راندمان بالایی نشان داده است. در این مدل تخصیص انبوه و میزان معقول بودن بر اساس تأثیر اجتماعی برای تشخیص برچسب دامنه هر گره محاسبه می‌شود.

روش MDCL در [۱۵] ارائه شد. این روش از یک الگوریتم جستجوی چندهدفه فاخته با جستجو محلی برای تشخیص اجتماعات در شبکه‌های اجتماعی پیچیده استفاده می‌کند. در روش MDCL دو تابع هدف نرخ تجمع منفی و نرخ شکاف تعریف شده است که از طریق محدودیت پیمانگی حداقل‌سازی می‌شوند. برای ایجاد جمعیت اولیه از یک

استراتژی جستجو محلی و یک اپراتور کلون استفاده می-شود.

در [۱۶] الگوریتم انتشار برچسب بر مبنای مرکزیت (CenLP) برای شناسایی اجتماع در شبکه‌های اجتماعی معرفی کردند. این الگوریتم یک تابع جدید را برای اندازه‌گیری مرکزیت گره‌ها به‌طور کمی بدون هیچ‌گونه تعامل با کاربر به‌وسیله محاسبه تراکم محلی و شباهت با همسایه‌های با تراکم بالا برای هر گره معرفی می‌کند.

در [۱۷] الگوریتم‌ها و برنامه‌های کاربردی برای تشخیص اجتماع در شبکه‌های وزن‌دار را مورد بررسی قرار دادند. در این مقاله به رفع مشکل تشخیص اجتماع در شبکه‌های وزن‌دار می‌پردازد.

در [۱۸] کشف جامعه کاربری از شبکه‌های چند ارتباطی را پیشنهاد دادند. بررسی‌های اولیه با دو مجموعه اطلاعاتی در دنیای واقعی نشان می‌دهد که این مدل می‌تواند اجتماع-هایی مشهود را مشخص کند، که در میان کاربران آن، دوستی‌های جدی و اشتراک‌های فراوان در زمینه علایق وجود دارد.

در [۱۹] تشخیص جوامع همپوشان در شبکه‌های اجتماعی بررسی شده است. این مقاله جدیدترین تکنولوژی الگوریتم تشخیص جامعه همپوشان، معیارهای و مقیاس‌های کیفی را مورد تجدیدنظر قرار می‌دهد.

روش TSC در [۲۰] ارائه شد که یک الگوریتم تشخیص اجتماعات ترکیبی برای شبکه‌های اجتماعی است. این روش یک خوشه‌بندی سری زمانی را برای تشخیص اجتماعات در نظر می‌گیرد. به‌طورکلی، نقطه اصلی این روش، تبدیل سری زمانی از دامنه زمان فضا به دامنه توپولوژیک است.

روش DLA در [۲۱] به‌منظور تشخیص کاربرانی با توئیت-های خشونت‌آمیز ارائه شد. DLA از یک الگوریتم مبتنی بر یادگیری ماشین با توجه به نظرات کاربران استفاده می-کند. تکنیک استفاده شده در DLA، یادگیری عمیق است که به‌صورت پیشرفته‌ای توسعه داده شده است.

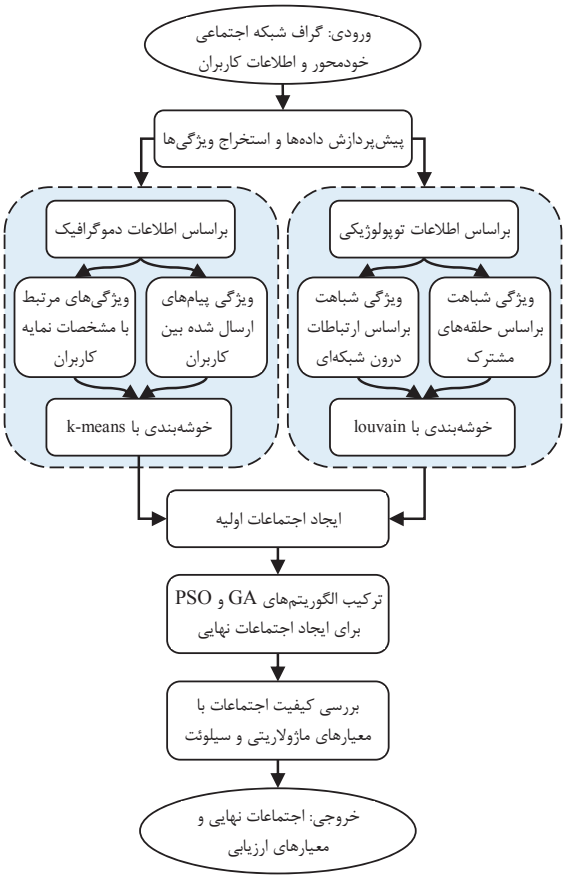
روش ACD-TC در [۲۲] ارائه شد که یک روش شناسایی اجتماعات تطبیقی با ترکیب توپولوژی و محتوا است. در اینجا برای ترکیب توپولوژی و محتوای از یک پارامتر تطبیقی استفاده شده تا تأثیر محتوا را در کشف اجتماعات به‌طور مؤثر کنترل کند. در نهایت یک الگوریتم کشف اجتماعات همپوشانی روی نتیجه پارتیشن‌بندی اجتماعات مرحله قبل اعمال می‌شود.

روش HICODE در [۲۳] برای تشخیص اجتماعات مخفی در شبکه‌های اجتماعی ارائه شد. از نکات برجسته این تحقیق می‌توان به پیشنهاد یک مفهوم جدید از جامعه پنهان برای تجزیه‌وتحلیل شبکه، ارائه متا-روش HICODE برای تشخیص اجتماعات مخفی، پیشنهاد روش‌هایی برای کاهش تأثیر ساختارهای تشخیصی و قابلیت استفاده برای هر دو اجتماعات غالب و اجتماعات مخفی اشاره کرد.

۳- روش پیشنهادی

در این تحقیق از الگوریتم‌های ژنتیک، ازدحام ذرات، خوشه‌بندی k-means و خوشه‌بندی louvain برای ارائه الگوریتم تشخیص اجتماعات پیشنهادی استفاده می‌شود. این الگوریتم مبتنی بر شبکه‌های اجتماعی خودمحور است. به‌طورکلی این شبکه‌ها روابط کاربر ego و دوستانشان را نشان می‌دهد. در گراف این شبکه‌ها کاربر ego در مرکز گراف قرار دارد و حلقه‌های تشکیل شده اطراف این کاربر حاوی روابط فردی خاص آن است. بنابراین در این نوع شبکه‌ها ارتباطات توسط ego معرفی می‌گردد.

ویژگی‌های استفاده شده از کاربران برای تشخیص اجتماعات، اطلاعات توپولوژیکی و دموگرافیکی (محتوا) آن‌ها می‌باشد. اطلاعات توپولوژیکی با ویژگی‌های شبکه-های خودمحور مرتبط است و از حلقه‌های مشترک و ارتباطات درون‌شبکه‌ای خودمحور تشکیل شده است. اطلاعات دموگرافیکی نیز مربوط به مشخصات نمایه کاربران و همچنین کلمات کلیدی استفاده شده در پیام(توئیت)های آن‌ها می‌باشد. روش پیشنهادی برای تشخیص اجتماعات از یک الگوریتم خوشه‌بندی ترکیبی بر مبنای الگوریتم‌های تکاملی بهره می‌گیرد. معماری روش پیشنهادی در شکل ۱ نشان داده شده است.



شکل ۱: معماری روش پیشنهادی

به‌طورکلی کاربران موجود در یک اجتماع باید بیشترین تراکم را با همدیگر داشته باشند و از طرفی باید دارای کمترین تراکم با سایر اجتماعات باشند. در این مسئله تراکم به مفهوم ارتباطات بین کاربران از لحاظ ویژگی‌های مختلفی است. برای مثال کاربران درون یک اجتماع باید دارای بیشترین ارتباط با همدیگر و کمترین ارتباط با سایر کاربران باشند. شاخص‌های اعتبارسنجی برای سنجش میزان صحت نتایج خوشه‌بندی به‌منظور مقایسه بین روش‌های خوشه‌بندی مختلف یا مقایسه نتایج حاصل از یک روش با پارامترهای مختلف مورداستفاده قرار می‌گیرند. ازاین‌رو از معیارهای ارزیابی ماژولاریتی و سیلوئت جهت ارزیابی عملکرد اجتماعات تشخیص داده شده و همچنین مقایسه با سایر روش‌ها استفاده می‌کنیم. معیار ماژولاریتی از جمله معیارهای اندازه‌گیری است که در زمان حداکثر بودن، منجر به ایجاد اجتماعات در یک چنین شبکه‌ی داده شده‌ای می‌شود. معیار سیلوئت یکی دیگر از شاخص‌های ارزیابی درونی خوشه‌بندی به‌منظور ارزیابی اجتماعات است که بر کیفیت اجتماعات و پراکندگی داده‌ها تمرکز دارد.

۳-۱- پیش‌پردازش داده‌ها

در پیش‌پردازش، ابتدا ماتریس هم‌جواری بین هر جفت کاربر ایجاد می‌شود. این ماتریس به‌صورت رابطه (۱) تعریف می‌شود.

$$Adj(u,v)=\begin{cases}0 & if\ u \nrightarrow v\\ 1 & if\ u \rightarrow v\end{cases} \qquad (۱)$$

سپس، با توجه به وجود چندین شبکه ego در ساختار اطلاعات در دسترس، ارتباطات موجود در شبکه ego، i-ام نیز به‌صورت ماتریس هم‌جواری A_i در رابطه (۲) نشان داده می‌شود.

$$A_i(u,v)=\begin{cases}0 & if\ u \overset{i}{\nrightarrow} v\\ 1 & if\ u \overset{i}{\rightarrow} v\end{cases} \qquad (۲)$$

جایی که $u \overset{i}{\rightarrow} v$ وجود ارتباط بین دو کاربر در i-مین شبکه ego را نشان می‌دهد. در این مرحله اطلاعات نمایه کاربران نیز نرمال‌سازی می‌شود. در اینجا فیلدهای مشخصات کاربران که به‌صورت کیفی بیان شده به‌صورت کمی کدگذاری می‌شوند.

۳-۲- استخراج ویژگی‌ها

در این تحقیق چهار ویژگی «شباهت براساس حلقه‌های مشترک» و «شباهت براساس ارتباطات درون‌شبکه‌ای» از اطلاعات توپولوژیکی و «پیام‌های ارسال‌شده بین کاربران» و «فیلدهای مرتبط با مشخصات نمایه کاربران» از اطلاعات دموگرافیکی کاربران استخراج می‌شود.

ویژگی شباهت براساس حلقه‌های مشترک: اغلب در شبکه‌های اجتماعی خودمحور کاربران از طریق مسیرهایی از حلقه‌های مختلف شبکه‌های ego مرتبط هستند. یک مسیر خودمحور به‌صورت $u \xrightarrow{follows} w \xrightarrow{friends} v$ نشان داده می‌شود، جائیکه کاربر w واسط ارتباط بین کاربران u و v است. مقدار ویژگی شباهت براساس حلقه‌های مشترک برای دو کاربر u و v برمبنای تعداد کاربران مشابه w که حداقل با یکی از کاربران u و v در حلقه مشترکی باشند، محاسبه می‌گردد.

ویژگی شباهت براساس ارتباطات درون‌شبکه‌ای: برای محاسبه این ویژگی از ساختار معیار شباهت معروف کاتز به‌صورت محلی درون هر شبکه ego استفاده می‌شود. مقدار این ویژگی براساس تعداد مسیرهای خودمحور نظیر $u \to w \to v$ با طول‌های متفاوت بین کاربران درنظر گرفته می‌شود، جایی که هر کاربر نظیر w باید حداقل با یکی از کاربران u و v در شبکه ego یکسانی باشد. شاخص توسعه‌یافته کاتز برای شبکه‌های خودمحور به‌صورت رابطه (۳) تعریف شده است.

$$KATZ_{ext}(u,v) = \sum_{l=2}^{L} \beta^l . \left|EgoPaths^l_{u,v}\right| \quad (۳)$$

جایی که β ضریب ثابت (کمتر از ۱) جهت کاهش تأثیر مسیرهای طولانی، L حداکثر طول مسیر مورد بررسی (در اینجا ۲، ۳ و ۴) و $\left|EgoPaths^l_{u,v}\right|$ تعداد مسیرهای خودمحور با طول l بین دو کاربر u و v می‌باشد. β^l باعث افزایش ارزش این ویژگی برای مسیرهایی با طول کوتاه‌تر می‌شود، چون هر چه طول مسیر کمتر باشد تراکم بیشتری وجود خواهد داشت.

ویژگی پیام‌های ارسال‌شده بین کاربران: این ویژگی برمبنای کلمات کلیدی استفاده‌شده مشابه در پیام‌ها و توئیت‌های کاربران محاسبه می‌شود. $T(u)$ کلمات کلیدی استفاده‌شده برای کاربر u است و به‌صورت $T(u) = [t_1, t_2, \dots, t_m]$ نشان داده می‌شود. جایی که t_i به‌صورت باینری نشان‌دهنده استفاده یا عدم استفاده از i-مین کلمه کلیدی توسط کاربر u می‌باشد. مقدار ویژگی «پیام‌های ارسال‌شده» برای دو کاربر u و v به‌صورت رابطه (۴) تعریف می‌شود.

$$CM(u,v) = \|T(u) \cap T(v)\| \quad (۴)$$

در این رابطه ∩ اشتراک بین دو بردار $T(u)$ و $T(v)$ است و $\|*\|$ مجموع تعداد عضوهای مشترک را محاسبه می‌کند.

ویژگی فیلدهای مرتبط با مشخصات نمایه کاربران: این ویژگی برمبنای فیلدهای «سن»، «جنسیت»، «میزان تحصیلات»، «شغل» و «موقعیت جغرافیایی» از هر کاربر در قالب یک ماتریس $N \times 5$ به‌منظور تشخیص اجتماعات استخراج می‌شود.

۳-۳- خوشه‌بندی کاربران

در این تحقیق برای کار خوشه‌بندی از دو الگوریتم معروف k-means و louvain به ترتیب برای خوشه‌بندی کاربران براساس اطلاعات دموگرافیکی و توپولوژیکی استفاده می‌شود. در ویژگی‌های بخش دموگرافیک چندین فیلد برای هر نمونه در اختیار است و به همین دلیل با ساختار الگوریتم k-means همخوانی دارد. ویژگی‌های بخش توپولوژیکی یک شباهت برای هر کاربر دارد و ساختار آن گراف است. به همین دلیل برای خوشه‌بندی آن از الگوریتم خوشه‌بندی مبتنی بر گراف louvain استفاده می‌شود.

برای هر چهار ویژگی استخراج‌شده، خوشه‌بندی‌های متفاوتی برای کاربران ایجاد می‌شود که از اطلاعات آن‌ها در جهت ایجاد اجتماعات نهایی استفاده می‌شود.

۳-۴- ایجاد اجتماعات اولیه کاربران

اجتماعات اولیه، شباهت‌های حداکثری بین کاربران را به تصویر می‌کشد و دارای اندازه‌های نسبتاً کوچکی هستند. در اینجا همه خوشه‌ها را یکپارچه و در یک مجموعه به‌صورت $C = \{\{c_1\},\{c_2\},\{c_3\},\{c_4\}\}$ در نظر می‌گیریم که c_i خوشه‌های مربوط به i-مین خوشه‌بندی انجام شده است. سپس شباهت بین هر جفت کاربر را براساس C محاسبه می‌کنیم. شباهت بین دو کاربر u و v به‌صورت $s_{u,v}$ نشان داده می‌شود و مقدار آن برابر است با جمع وزنی تعداد دفعاتی که این دو کاربر در خوشه‌های یکسانی دیده شده است. $s_{u,v}$ با توجه به رابطه (۵) محاسبه می‌گردد.

$$s_{u,v} = \sum_{c} \frac{1}{N_c}, \quad if\ (u,v) \in c \quad (۵)$$

در اینجا N_c تعداد کاربران خوشه c است. هر چه دو کاربر در یک خوشه با تعداد اعضاء کمتر دیده شوند، شباهت بیشتری دارند. این فرایند منجر به ایجاد ماتریس شباهت S می‌شود که یک ماتریس $N \times N$ است. در ادامه براساس تعیین حداکثر شباهت‌ها روی ماتریس S، اجتماعات اولیه ایجاد می‌شود. حداکثر شباهت‌های کاربران به‌صورت رابطه (۶) تعریف می‌گردد.

$$S_{MAX} = \begin{bmatrix} S^1_{max} \\ S^2_{max} \\ \vdots \\ S^N_{max} \end{bmatrix} = \begin{bmatrix} \max[s_{1,1}, & s_{1,2}, & \cdots & s_{1,N}] \\ \max[s_{2,1}, & s_{2,2}, & \cdots & s_{2,N}] \\ \vdots & \vdots & \vdots & \vdots \\ \max[s_{N,1}, & s_{N,2}, & \cdots & s_{N,N}] \end{bmatrix} \quad (۶)$$

برای محاسبه حداکثر شباهت بین کاربران هر سطر از ماتریس S مورد توجه می‌باشد. S_{MAX} نشان می‌دهد که کاربر i-ام با تمام کاربران (در هر سطر) به مقدار S_{max}^{i} رسیده است. اجتماعات اولیه براساس S_{MAX} ایجاد می‌شود. کاربران متعلق به اجتماع اول براساس سطر اول ماتریس یعنی S_{max}^{1} ایجاد می‌شوند. در اینجا همه کاربرانی که مقدار شباهت آن‌ها برابر S_{max}^{1} است، به اجتماع اول ($cluster_1$) تعلق دارند. برای تعیین کاربران متعلق به اجتماعات بعدی به‌صورت زیر عمل می‌کنیم. جایی که فرض می‌شود اجتماع بعدی $cluster_z$ است.

۱. کاربر j جایی که $j = 2,3, \dots N$ مورد بررسی قرار می‌گیرد.
۲. اگر کاربر j قبلاً به اجتماع خاصی تخصیص داده شده، متعلق به $cluster_z$ نمی‌باشد در غیر این صورت کاربر j متعلق به اجتماع $cluster_z$ است.
۳. همه کاربرانی با مقدار شباهت S_{max}^{j} (سطر j-ام از S_{MAX}) که قبلاً به اجتماع خاصی تخصیص داده نشده‌اند، متعلق به اجتماع $cluster_z$ هستند.

براساس مراحل فوق درنهایت Z اجتماع اولیه ایجاد می‌شود، جایی که $cluster_z \in [1,2, \dots, z, \dots, Z]$ می‌باشد.

۳-۵- ایجاد اجتماعات نهایی

در این تحقیق برای ایجاد اجتماعات نهایی از یک رویکرد ترکیبی مبتنی بر الگوریتم‌های ژنتیک و ازدحام ذرات استفاده شده است. تعداد اجتماعات نهایی در طی پروسه بهینه‌سازی و ادغام اجتماعات کوچک‌تر به‌صورت خودکار محاسبه می‌شود. هدف الگوریتم ترکیبی پیشنهادی ادغام اجتماعات کوچک‌تر و ایجاد اجتماعات بزرگ و نهایی به نحوی است که بهترین بهره‌وری حاصل شود. در اینجا اجتماعات کوچک‌تر براساس الگوریتم ژنتیک ادغام شده و به‌منظور افزایش کیفیت اجتماعات یک الگوریتم ازدحام ذرات در هر نسل اعمال می‌شود. الگوریتم ازدحام ذرات وظیفه جابه‌جایی گره‌ها بین اجتماعات ادغام‌شده را بر عهده دارد و این در حالی است که الگوریتم ژنتیک ادغام اجتماعات را انجام می‌دهد.

ساختار راه‌حل پیشنهادی به‌صورت یک بردار به طول Z است، جایی که Z تعداد اجتماع اولیه است. هر عنصر در بردار تعریف شده بیانگر یک اجتماع اولیه و مقدار منتسب به آن شماره اجتماع نهایی را نشان می‌دهد. با توجه به ساختار راه‌حل پیشنهادی جمعیت اولیه به‌صورت تصادفی ایجاد می‌شود.

در این مقاله برای محاسبه برازندگی راه‌حل‌ها از معیارها ماژولاریتی و سیلوئت به‌صورت مقایسه‌ای استفاده شده است. برای انتخاب والدین و تولیدمثل از عملگر انتخاب چرخ رولت استفاده شده و عملگر برش تک-نقطه‌ای با احتمال C_r است. علاوه بر این، عملگر جهش استفاده شده با احتمال M_r روی هر بیت از راه‌حل اعمال شده و مقدار منتسب به آن را یک واحد به‌صورت تصادفی اضافه یا کم می‌کند.

پس از ایجاد جمعیت جدید در هر نسل، یک الگوریتم ازدحام ذرات به‌منظور افزایش کیفیت اجتماعات یک الگوریتم ازدحام ذرات روی جمعیت اعمال می‌شود.

موقعیت هر ذره در فضای جستجو به‌صورت یک بردار به طول تعداد کاربران (N) است که هر عنصر آن شماره اجتماع تشخیص داده شده از راه‌حل الگوریتم ژنتیک است. در اینجا جمعیت اولیه ذرات همان جمعیت اولیه از الگوریتم ژنتیک می‌باشد. هر ذره یک مقدار شایستگی دارد که همان معیارهای سیلوئت و ماژولاریتی هستند. در هر گام، هر ذره با استفاده از دو بهترین مقدار به‌روز می‌شود. اولین مورد، بهترین موقعیتی است که تاکنون ذره موفق به رسیدن به آن شده است (pbest). بهترین مقدار دیگر، بهترین موقعیتی است که تاکنون توسط جمعیت ذرات به‌دست‌آمده است و مصداق هوش جمعی است (gbest).

سرعت برای هر ذره با بردار v نشان داده شده که ابعادی مشابه بردار موقعیت ذرات دارد. مقادیر اولیه بردار v برای هر ذره در ابتدا در بازه ۰ تا ۱ به‌صورت تصادفی انتخاب می‌شود. پس از یافتن بهترین مقادیر (pbest و gbest)، سرعت و مکان هر ذره به‌روزرسانی می‌شوند. با توجه به محیط گسسته در مسئله، موقعیت ذرات به‌صورت زیر تغییر می‌کند:

۱. برای هر اجتماع نهایی یک عنصر (کاربر) از بردار ذره انتخاب می‌شود. این انتخاب بر مبنای بردار سرعت و با استفاده از عملگر چرخ رولت انجام می‌شود.

۲. برای هر عنصر انتخاب‌شده، به‌صورت تصادفی یک شماره اجتماع تخصیص داده می‌شود.
۳. ذره جدید در صورت بهبود معیارهای برازندگی به موقعیت جدید می‌رود.

۳-۶- بررسی کیفیت اجتماعات

اجتماعات شامل کاربرانی است که نسبت به همدیگر بیشترین تراکم (ارتباطات) را دارند و از طرفی کمترین تراکم را با کاربران موجود در سایر اجتماعات دارند. تراکم به ارتباطات میان کاربران براساس ویژگی‌های استخراج شده از گراف شبکه اجتماعی اشاره دارد. از جمله معروف‌ترین معیارهای ارزیابی کیفیت اجتماعات در شبکه‌های اجتماعی بر اساس تراکم، ماژولاریتی و سیلوئت است. این معیارها زمانی استفاده می‌شود که نتایج واقعی خوشه‌های برای هر نمونه را نداشته باشیم.

معیار ماژولاریتی برای اجتماعات $C = \{C_1, C_2, \dots, C_c\}$ با توجه به رابطه (۷) محاسبه می‌شود.

$$M(C) = \sum_{c} \left(\frac{1}{2\xi} \sum_{u \in c,\ v \in c'} \left[A_{u,v} - \frac{k_u \times k_v}{2\xi} \right] \delta(u,v) \right) \quad (۷)$$

جایی که c به کاربران اجتماع c و c' به کاربران سایر اجتماعات غیر از c اشاره دارد. $A_{u,v}$ وزن بین دو کاربر u و v است که به‌صورت میانگین وزنی برای همه ویژگی‌های محاسبه شده بین دو کاربر در نظر گرفته می‌شود، k_u و k_v به ترتیب نشان‌دهنده مجموع وزن‌های یال‌های متصل به کاربران u و v می‌باشد. ξ مجموع همه وزن‌های گراف شبکه اجتماعی است و درنهایت $\delta(u,v)$ یک تابع دلتا ساده برای نشان دادن ارتباط و یا عدم ارتباط بین کاربران u و v است (۰ عدم ارتباط و ۱ ارتباط).

معیار سیلوئت بر محاسبه پراکندگی داده‌ها در اجتماعات مختلف متمرکز است و معیارهای فواصل درون خوشه‌ای و برون خوشه‌ای را به‌طور هم‌زمان لحاظ می‌کند. معیار سیلوئت به‌صورت رابطه (۸) تعریف می‌شود.

$$S = \frac{1}{N} \sum_{i=1}^{N} \frac{b_i - a_i}{\max(b_i, a_i)} \quad (۸)$$

جایی که N تعداد کل کاربران، a_i متوسط فاصله (وزن) کاربر u_i از سایر کاربران در خوشه خودش و b_i متوسط فاصله کاربر u_i از سایر کاربران در سایر خوشه‌ها می‌باشد. b_i برای هر خوشه یک ارزش دارد که مقدار نهایی آن به‌صورت کمترین در نظر گرفته می‌شود.

۴- نتایج و بحث

در این تحقیق آزمایش‌های گسترده‌ای را برای نشان داده برتری الگوریتم پیشنهادی ارائه شده است. نتایج در مقابل الگوریتم‌های ACD-TC [۲۲]، HICODE [۲۳] و TSC [۲۰] با توجه به معیارهای ماژولاریتی [۲۴]، سیلوئت [۲۵] و تعداد اجتماعات مورد مقایسه قرار می‌گیرد. برای انجام شبیه‌سازی از نرم‌افزار MATLAB ورژن 2017a و یک کامپیوتر شخصی با پیکربندی پردازنده اینتل ۵ هسته و فرکانس ۲,۴ گیگاهرتز، حافظه ۱۶ گیگابایت استفاده شده است.

شبیه‌سازی روی مجموعه داده شبکه اجتماعی خودمحور فیس‌بوک انجام شده است. با توجه به حجم بالای تعداد کاربران، در این مقاله تنها از ۳ شبکه ego ابتدایی با ۲۱۶۷ کاربر و ۴۵۴۲۳ لینک استفاده شده است. در هر مرحله از اعتبارسنجی کاربران به دو بخش آموزشی (E^T) و آزمایشی (E^P) تفکیک می‌شوند. این فرایند مطابق 10-fold اجرا شده و نتایج نهایی به‌صورت میانگین ارائه می‌گردد. بنابراین $E = E^T \cup E^P$ و $E^T \cap E^P = \emptyset$ است، جائیکه E ارتباطات کل شبکه را نشان می‌دهد.

در ابتدا از معیار تعداد اجتماعات برای مقایسه روش‌های مختلف بهره می‌گیریم. تعداد اجتماعات نشان می‌دهد که الگوریتم ارائه شده در نهایت با چه تعداد اجتماع، تراکم کاربران را نشان می‌دهد. در تکرارهای متوالی هر چه یک الگوریتم بتواند تعداد اجتماعات تشخیص داده شده مشابهی حاصل کند، الگوریتم پایداری بیشتری دارد. بدین منظور در این آزمایش نتایج تعداد خوشه‌های نهایی برای روش پیشنهادی و سایر روش‌های مورد مقایسه در ۱۵ اجرای مختلف گزارش شده است. شکل ۲ نتایج این آزمایش را نشان می‌دهد.



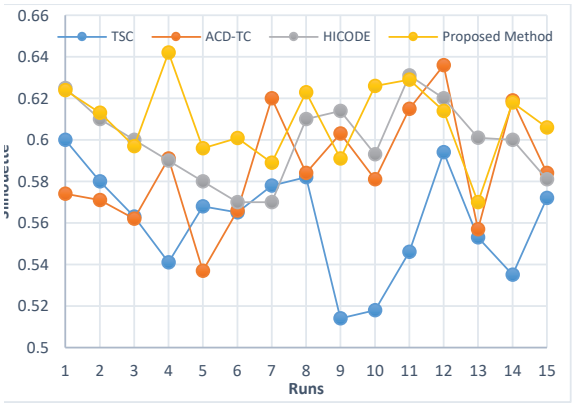
شکل ۲: مقایسه تعداد اجتماعات کشف‌شده در روش‌های مختلف

با توجه به نتایج این آزمایش، روش پیشنهادی کمترین پراکندگی را در بین سایر روش‌ها در تشخیص تعداد اجتماعات دارد. در اینجا تعداد اجتماعات تشخیص داده شده برای روش پیشنهادی تنها در محدوده ۱۹ تا ۲۲ اجتماع در نوسان است و به‌طور میانگین این تعداد حدود ۲۱ است. این در حالی است که روش TSC دارای بیشترین پراکندگی تعداد اجتماعات تشخیص داده شده است. این روش به‌طور میانگین تعداد اجتماعات را حدود ۱۸ تخمین می‌زند و نوسانات آن بین ۱۵ تا ۲۱ است. روش‌های ACD-TC و HICODE نیز عملکرد مناسبی دارند هرچند از روش پیشنهادی نتایج ضعیف‌تری را ارائه می‌دهند. به‌طور خاص میانگین تعداد اجتماعات در این دو روش حدود ۲۰ است. از بین روش‌های ACD-TC و HICODE؛ روش ACD-TC پراکندگی بیشتری را در تعداد اجتماعات تشخیص داده شده گزارش می‌دهد به نحوی این پراکندگی بین ۱۷ تا ۲۳ است. این در حالی است که پراکندگی تعداد اجتماعات روش HICODE بین ۱۸ تا ۲۳ می‌باشد.

بنابراین نتایج این آزمایش نشان داد که روش پیشنهادی به چند مرحله کردن تشخیص اجتماعات و استفاده از دو جنبه مختلف توپولوژی و محتوا توانسته است نسبت به سایر روش‌های مشابه تعداد اجتماعات پایدارتری را ارائه دهد.

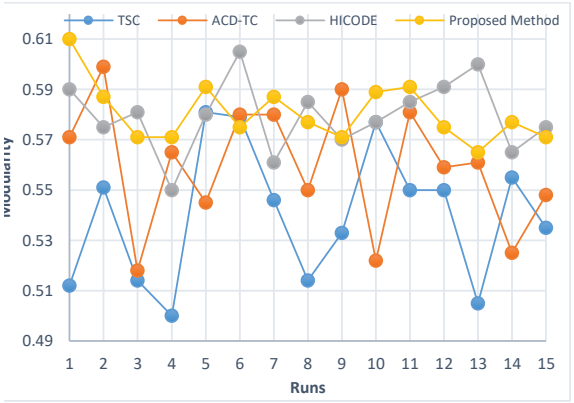
در ادامه روش‌های مختلف از لحاظ معیار سیلوئت بررسی و مقایسه می‌شود. شکل ۳ نتایج این آزمایش را برای روش پیشنهادی و سه روش TSC، ACD-TC و HICODE نشان می‌دهد. در اینجا معیار سیلوئت به‌صورت میانگین برای همه کاربران آزمایشی و برای ۱۵ اجرای متمایز گزارش شده است.

نتایج شاخص سیلوئت نشان می‌دهد که روش پیشنهادی با میانگین ۰,۶۱ بهترین عملکرد را در بین سایر روش‌ها دارد. در رتبه دوم روش HICODE با میانگین سیلوئت ۰,۶ قرار دارد. علاوه بر این، روش‌های ACD-TC و TSC نیز به ترتیب به ۰,۵۹ و ۰,۵۶ در رتبه‌های بعدی قرار دارند. در بهترین اجرا روش پیشنهادی به درجه سیلوئت ۰,۶۴۲ رسیده است، این در حالی است که بهترین معیار سیلوئت روش HICODE مقدار ۰,۶۳۱ است. هرچند در برخی از اجراها روش HICODE نسبت به روش پیشنهادی برتری دارد اما در حالت میانگین روش پیشنهادی اجتماعات بهتری را ارائه می‌دهد، هرچند این برتری ناچیز است.



شکل ۳: مقایسه روش‌های مختلف در شاخص سیلوئت

در آزمایش دیگری روش‌های مختلف را از لحاظ معیار ماژولاریتی مورد مقایسه قرار داده‌ایم. ماژولاریتی به چگالی (تراکم) ارتباطات اجتماعات ایجاد شده تمرکز دارد. هر چه ماژولاریتی برای یک روش بیشتری باشد، این امر نشان‌دهنده عملکرد مناسب و ایجاد اجتماعات قوی‌تری است. نتایج این مقایسه در شکل ۴ برای ۱۵ اجرای مختلف ارائه شده است.



شکل ۴: مقایسه روش‌های مختلف در شاخص ماژولاریتی

نتایج این آزمایش نیز برتری نسبی روش پیشنهادی را نشان می‌دهد. میانگین نتایج شاخص ماژولاریتی برای روش پیشنهادی ۰,۵۸ است، این در حالی است که روش HICODE با اختلاف ناچیزی به شاخص ماژولاریتی ۰,۵۷۹ رسیده است. در برخی از اجراها روش HICODE برتری دارد اما درمجموع روش پیشنهادی نتایج بهتری را گزارش کرده است. بهترین شاخص ماژولاریتی برای روش پیشنهادی ۰,۶۱ است درحالی‌که برای روش HICODE این عدد برابر ۰,۶۰۵ است. علاوه بر این روش پیشنهادی نسبت به دو روش TSC و ACD-TC برتری قابل‌توجهی دارد. شاخص ماژولاریتی برای این دو روش به ترتیب ۰,۵۴ و ۰,۵۶ است. به‌طورکلی می‌توان نتیجه گرفت که روش پیشنهاد با استفاده از خوشه‌بندی ترکیبی موفق به ایجاد اجتماعات متراکم‌تری نسبت به سایر روش‌های مشابه شده است.

در انتها نتایج مقایسه برای معیارهای تعداد اجتماعات، سیلوئت و ماژولاریتی برای روش‌های TSC، ACD-TC، HICODE و همچنین روش پیشنهادی برای مقایسه دقیق‌تر به‌صورت عددی گزارش می‌شوند. نتایج این مقایسه در جدول ۱ نشان داده شده است. نتایج ارائه شده بر مبنای میانگین ۱۵ اجرای متمایز اندازه‌گیری شده است. نتایج اثبات می‌کنند که روش پیشنهادی به اجتماعات پایدارتری در میانگین اجراها رسیده است.

جدول ۱: مقایسه روش پیشنهادی نسبت به سایر روش‌ها

روش‌ها	ماژولاریتی	سیلوئت	تعداد اجتماعات
TSC	۰,۵۴۰	۰,۵۶۰	۲۰,۳۳
ACD-TC	۰,۵۵۹	۰,۵۸۵	۱۷,۸۷
HICODE	۰,۵۷۹	۰,۶۰۱	۲۰,۴۵
روش پیشنهادی	۰,۵۸۱	۰,۶۰۸	۲۱,۰۶

به‌طورکلی میزان بهبود روش پیشنهادی نسبت به روش HICODE در شاخص سیلوئت ۱,۲ و برای ماژولاریتی ۰,۵ درصد است. هرچند این برتری ناچیز است اما روش پیشنهادی به‌طور میانگین نتایج بهتری را گزارش داده است. علاوه بر این برتری روش پیشنهادی در شاخص سیلوئت نسبت به روش‌های TSC و ACD-TC به ترتیب ۷,۵ درصد و ۳,۹ درصد و برای شاخص ماژولاریتی به ترتیب ۸,۵ درصد و ۳,۸ درصد است.

۵- نتیجه‌گیری

یکی از چالش‌های اصلی در آنالیز شبکه‌های اجتماعی، تشخیص اجتماعات (جوامع) است. در این مقاله از یک روش ترکیبی برای تشخیص اجتماعات بر مبنای ویژگی‌های توپولوژیکی و دموگرافیکی استفاده شد. اطلاعات توپولوژیکی با ویژگی‌های شبکه‌های خودمحور مرتبط است و از حلقه‌های مشترک و ارتباطات درون‌شبکه‌ای خودمحور تشکیل شده است. اطلاعات دموگرافیکی نیز مربوط به مشخصات نمایه کاربران و همچنین توئیت‌های مشترک آن‌ها می‌باشد. اجتماعات اولیه براساس شباهت‌های حداکثری و اجتماعات نهایی براساس ترکیب دو الگوریتم ژنتیک و ازدحام ذرات ایجاد شد. الگوریتم ترکیبی ارائه شده با هدف حداکثر بهره‌وری سعی در ادغام خوشه‌ها و اجتماعات کوچک دارد.

استفاده از سایر شاخص‌های ارزیابی درون خوشه‌ای نظیر جاکارد برای بررسی اجتماعات نهایی کاربران از پیشنهادات این تحقیق برای کارهای آتی است.

مراجع

[1] Bliss, C.A. , Frank, M.R. , Danforth, Ch.M. , Dodds, .P.Sh. (2014). An evolutionary algorithm approach to link prediction in dynamicsocial networks. Journal of Computational Science,5(5), 750-764.

[2] Cui, W., Pu, C., Xu, Z., Cai, S., Yang, J., & Michaelson, A. (2016). Bounded link prediction in very large networks. Physica A: Statistical Mechanics and its Applications, 457, 202-214.

[3] Han, S., & Xu, Y. (2016). Link Prediction in Microblog Network Using Supervised Learning with Multiple Features. JCP, 11(1), 72-82.

[4] Liu, F., Liu, B., Sun, C., Liu, M., & Wang, X. (2015). Deep belief network-based approaches for link prediction in signed social networks. Entropy, 17(4), 2140-2169.

[5] Naderipour, M., Bastani, S., & Zarandi, M. F. (2016). A Type-2 Fuzzy Model for Link Prediction in Social Network. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 10(7), 1355-1360.

[6] Nguyen, Trung LT, and Tru H. Cao. (2014). Multi-group-based User Perceptions for Friend Recommendation in Social Networks. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Cham. 525-534

[7] Sun, Z., Han, L., Huang, W., Wang, X., Zeng, X., Wang, M., & Yan, H. (2015). Recommender systems based on social networks. Journal of Systems and Software, 99, 109-119.

[8] Laishram, R., Mehrotra, K., & Mohan, C. K. (2016, November). Link Prediction in Social Networks with Edge Aging. In Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on (pp. 606-613). IEEE.

[9] Jalili, M., Orouskhani, Y., Asgari, M., Alipourfard, N., & Perc, M. (2017). Link prediction in multiplex online social networks. Royal Society Open Science, 4(2), 160863.

[10] Szczepański, P. L., Barcz, A. S., Michalak, T. P., & Rahwan, T. (2015, June). The game-theoretic interaction index on social networks with applications to link prediction and community detection. In Twenty-Fourth International Joint Conference on Artificial Intelligence.

[11] Zhao, J., Miao, L., Yang, J., Fang, H., Zhang, Q. M., Nie, M., ... & Zhou, T. (2015). Prediction of links and weights in networks by reliable routes. Scientific reports, 5.

[12] L. L¨u, T. Zhou, Link prediction in complex networks: A survey, Physica A: Statistical Mechanics and its Applications 390 (2011) 1150–1170.

[13] Said, A., Abbasi, R. A., Maqbool, O., Daud, A., & Aljohani, N. R. (2018). CC-GA: A clustering coefficient based genetic algorithm for detecting communities in social networks. Applied Soft Computing, 63, 59-70.

[14] Azaouzi, M., & Romdhane, L. B. (2017). An evidential influence-based label propagation algorithm for distributed community detection in social networks. Procedia Computer Science, 112, 407-416.

[15] Zhou, X., Liu, Y., & Li, B. (2016). A multi-objective discrete cuckoo search algorithm with local search for community detection in complex networks. Modern Physics Letters B, 30(7), 1650080.

[16] Sun, H., Liu, J., Huang, J., Wang, G., Yang, Z., Song, Q., & Jia, X. (2016). CenLP: A centrality-based label propagation algorithm for community detection in networks. Physica A: Statistical Mechanics and its Applications, 436, 767-780.

[17] Lu, Z., Sun, X., Wen, Y., Cao, G., & La Porta, T. (2015). Algorithms and applications for community detection in weighted networks. IEEE Transactions on Parallel and Distributed Systems, 26(11), 2916-2926.

[18] Zhang, Z., Li, Q., Zeng, D., & Gao, H. (2013). User community discovery from multi-relational networks. Decision Support Systems, 54(2), 870-879.

[19] Xie, J., Kelley, S., & Szymanski, B. K. (2013). Overlapping community detection in networks: The state-of-the-art and comparative study. Acm computing surveys (csur), 45(4) 43.

[20] Ferreira, L. N., & Zhao, L. (2016). Time series clustering via community detection

in networks. Information Sciences, 326, 227-242.

[21] Dadhe, M. M. S., Masidkar, M. P. S., Vaidya, M. V., & Jalan, P. A. (2018). Detection of Abusive Language from Tweets in Social Networks. International Journal on Recent and Innovation Trends in Computing and Communication, 6(3), 148-151.

[22] Qin, M., Jin, D., Lei, K., Gabrys, B., & Musial-Gabrys, K. (2018). Adaptive community detection incorporating topology and content in social networks. Knowledge-Based Systems, 161, 342-356.

[23] He, K., Li, Y., Soundarajan, S., & Hopcroft, J. E. (2018). Hidden community detection in social networks. Information Sciences, 425, 92-106.

[24] Azaouzi, M., & Romdhane, L. B. (2017). An evidential influence-based label propagation algorithm for distributed community detection in social networks. Procedia Computer Science, 112, 407-416.

[25] Xie, J., Kelley, S., & Szymanski, B. K. (2013). Overlapping community detection in networks: The state-of-the-art and comparative study. Acm computing surveys (csur), 45(4), 43.