استفاده از معیارهای شباهت و ویژگی‌های دموگرافیک کاربران برای تشخیص اجتماعات در شبکه‌های اجتماعی

احد فخاری[۱]، موسی مجرد[۲و۳] و حسن ارفعی‌نیا[۳]

[۱] دانشجو کارشناسی ارشد، موسسه آموزش عالی غیرانتفاعی-غیر دولتی 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

چکیده: امروزه شبکه‌های اطلاعاتی در حوزه‌های مختلفی مورداستفاده قرار می‌گیرد. یکی از انواع شبکه‌های اطلاعاتی که امروزه به‌صورت گسترده‌ای مورداستفاده قرار می‌گیرد، شبکه‌های اجتماعی است. شبکه‌های اجتماعی با کاربردهای متنوع دربرگیرنده تعداد بالایی از کاربران است که با واسطه یا بدون واسطه با یکدیگر در ارتباط هستند. در این راستا، با توجه به حجم بالای کاربران شبکه، یکی از قابلیت‌هایی که شبکه می‌تواند در اختیار کاربران قرار دهد، تشخیص اجتماعات بر اساس معیارهای تشابه بین کاربران است. ساختار کلی سیستم‌های تشخیص اجتماعات در نظر دارد که با استفاده از یکسری معیارها، شباهت بین کاربران را محاسبه کرده و سپس کاربرانی با بالاترین شباهت را به‌عنوان یک اجتماع معرفی کند. این معیارها می‌توانند شباهت بین کاربران را بر اساس اطلاعات کاربران و برخی ویژگی‌های ساختاری گراف محاسبه کنند. در این مقاله ویژگی‌های مورد نیاز جهت خوشه‌بندی و تشخیص اجتماعات از اطلاعات دموگرافیک کاربران استخراج می‌شود. سیستم خوشه‌بندی با استفاده از درخت تصمیم آموزش داده می‌شود تا کاربران جدید را در مناسب‌ترین گروه قرار دهند. درنهایت با ترکیب چند معیار شباهت معروف و کنترل تأثیر هر یک در محاسبه شباهت نهایی، نمره مشابهت بین کاربران مشخص شده و اجتماعات نهایی تشخیص داده می‌شود. به‌منظور ارزیابی عملکرد روش پیشنهادی از مجموعه داده واقعی شبکه اجتماعی پوپک استفاده شده است. نتایج حاصل از آزمایش‌ها عملکرد بالای روش پیشنهادی را در مقایسه با سایر روش‌های مشابه نشان می‌دهد.

کلیدواژه‌ها: *شبکه‌های اجتماعی، تشخیص اجتماعات، معیارهای شباهت، خوشه‌بندی کاربران.*

۱- مقدمه

نخستین بار مفهومی با عنوان شبکه‌های اجتماعی اینترنتی با قالب امروزی در سال ۱۹۶۰ اولین بار در دانشگاه ایلی نویز در ایالت متحده امریکا مطرح شد [۱]. پس از آن در ۱۹۹۷ نخستین سایت شبکه اجتماعی اینترنتی SixDegrees راه‌اندازی شد. این سایت به کاربرانش اجازه ایجاد پروفایل داد تا آن‌ها بتوانند فهرستی از دوستانشان ایجاد کنند [۲]. شبکه‌های اجتماعی ساختاری اجتماعی است که از گره‌هایی که به‌صورت منفرد و یا گروهی تشکیل شده‌اند که با یکدیگر ارتباط دارند؛ بنابراین می‌توان گفت هر شبکه اجتماعی شامل تعدادی از افراد و ارتباط بین آن‌ها می‌باشد [۳]. برای نمایش شبکه‌های اجتماعی و محاسبات اولیه روی آن‌ها می‌توان نگاشتی از شبکه‌ها به ماتریس‌ها و گراف‌ها در نظر گرفت که می‌توان از ماتریس‌های مجاورت یا شباهت استفاده کرد.

شبکه‌های اجتماعی متشکل از افراد (سازمان‌ها) هستند که توسط یک یا چند نوع علاقه خاص به هم گره‌خورده‌اند.

وابستگی‌هایی مثل دوستی، خویشاوندی، منافع مشترک، تبادل مالی، علاقه خاص به یک موضوع باورهای مشترک، دانش و هنر و غیره. تجزیه‌وتحلیل شبکه های اجتماعی و تحلیل نظریه‌های موجود نشان می‌دهد که اصلی‌ترین نکته در هر شبکه‌ای، روابط است [۴-۶]. از آنجا که درک و فهم این روابط در سطح و گستره یک شبکه بسیار مشکل و غیرممکن است، تحلیل شبکه‌های اجتماعی نیز کاری بسیار دشوار می‌نمایاند. شبکه‌های اجتماعی در بسیاری از سطوح فعالیت می‌کنند، از خانواده تا سطح کشورها و بین‌الملل و نقش مهمی در تعیین راه حل مشکلات یا ایجاد آن ها دارند. جامعه مجازی در ساده‌ترین شکل آن، بستری است برای ایجاد روابط، مشخص مانند دوستی، بین گره‌های، گروه‌هایی که به افراد متصل هستند است در نتیجه ارتباطات اجتماعی از آن فرد به شبکه و از شبکه به‌کل افراد سرایت می‌کند [۷].

بررسی‌ها نشان می‌دهد که شبکه‌های اجتماعی در خصوصیات مشترک ساختاری به طرز قابل‌توجهی اشتراک دارند. تحلیل‌های با ارزشی به‌منظور گراف‌کاوی تاکنون ارائه شده است. در دهه‌ی اخیر، بررسی ساختار شبکه‌ها با استفاده از داده‌های واقعی و ابزارهای پرداز شی قوی، خصوصیات مشابه انواع شبکه‌ها با یکدیگر را آشکار کرده است. به‌عنوان مثال کوچک بودن میانگین کوتاه‌ترین فاصله‌ها، بالا بودن خوشه‌بندی و وجود اجتماعات در شبکه‌های پیچیده تنها برخی از خصوصیات مشترک شبکه‌ها در علوم مختلف می‌باشد. حضور تمامی این خصوصیات در شبکه‌های واقعی آنان را از مدل‌های بررسی شده در ریاضیات، مانند تورها و گراف‌های تصادفی، متمایز می‌کند [۸]. توپولوژی شبکه‌های اجتماعی نقش مهمی در استخراج برخی اطلاعات از آن ها ایفا می‌کند. توپولوژی شبکه به ساختار و نحوه اتصال نودهای آن اشاره دارد. قسمت اعظم فرایند استخراج توپولوژی شبکه با استفاده از توزیع درجه نودها انجام می‌شود.

با گسترش روزافزون تعداد کاربران شبکه های اجتماعی، تشخیص اجتماعات موضوع محور بر اساس موضوعات مشترک موردعلاقه کاربران به یکی از زمینه‌های پژوهشی جذاب در حوزه تحلیل شبکه‌های اجتماعی بدل گشته است. یک از چالش های این حوزه وابسته بودن الگوریتم ها و راهکارهای تشخیص اجتماعات بر اساس محوریت موضوع شبکه‌های اجتماعی است [۸]. در اغلب مطالعات تنها روابط ساده مابین کاربران مورد بررسی قرار گرفته است؛ اما اخیراً ویژگی دموگرافیک کاربران برای محاسبه روابط و میزان علاقه بین دو کاربر در شبکه‌های اجتماعی و در نظر داشتن آن در تحلیل‌ها، موردتوجه محققان این حوزه قرار گرفته است. مفهوم اجتماعات کاربران، یک مفهوم اجتماعی است. اجتماعات کاربران شبکه‌های اجتماعی می‌تواند گروه‌های دوستان یا گروه‌هایی از مردمی باشند که علایق مشترکی دارند.

ادامه این تحقیق بدین صورت است؛ پیشینه تحقیق در بخش دوم ارائه می شوند. بخش سوم مرتبط به روش پیشنهادی این تحقیق جهت کشف اجتماعات است و نتایج حاصل از آن در بخش چهارم ارائه می‌گردد. در نهایت در بخش پنجم نتیجه‌گیری و پیشنهادها 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.

روش ECD توسط علیزاده و همکاران (۱۳۹۲) ارائه شد که یک الگوریتم تشخیص اجتماعات ترکیبی برای شبکه‌های اجتماعی می‌باشد [۹]. خوشه‌بندی یا تشخیص اجتماعات، طبقه‌بندی کاربران در شبکه‌های اجتماعی و ارتباطات پنهان بین مؤلفه‌های آن‌ها را آشکار خواهد نمود. اکثر روش‌های رایج تشخیص اجتماعات موجود قطعی نیستند و نتایج آن‌ها به مقادیر اولیه‌ای که در اکثر مواقع به‌صورت تصادفی انتخاب می‌شود بستگی دارد؛ اما خوشه‌بندی ترکیبی، بدون توجه به مقادیر اولیه تصادفی هرکدام از الگوریتم‌های پایه‌اش، با ترکیب آن‌ها، نتایج مستحکم و پایداری تولید می‌کند. روش ECD یک الگوریتم ترکیبی تشخیص اجتماعات با الهام از خوشه‌بندی ترکیبی ارائه می‌دهد. از مشخصه‌های این روش تشخیص اجتماعات ترکیبی، توانایی ترکیب با روش های دیگر است به‌گونه‌ای که می‌توان از الگوریتم‌های دقیق‌تر نیز در چهارچوب پیشنهادی استفاده کرد. نتایج تجربی روش ECD نشان می‌دهند که به متوسط روش‌های تشخیص اجتماعات پایه‌ای

مورد استفاده در آن از کارایی بسیار بالاتری برخوردار بوده و در اکثر موارد حتی از بهترین الگوریتم پایه‌ای خود نیز بهتر عمل کرده است.

روش DLA توسط دادا و همکاران (۲۰۱۸) به‌منظور طبقه‌بندی توئیت‌های خشونت‌آمیز کاربران شبکه‌های اجتماعی ارائه شد [۱۰]. اکثر روش‌های تجارتی اخیر برای تشخیص واژه‌های خشونت‌آمیز از لیست سیاه استفاده می‌کنند، بااین‌وجود این تکنیک کارایی مناسبی در برابر نمونه‌های زیادی که در نظر گرفته نشده‌اند، ندارد. در روش DLA از یک الگوریتم مبتنی بر یادگیری ماشین برای کشف توئیت‌های خشونت‌آمیز در نظرات کاربران آنلاین استفاده می‌کند. تکنیک استفاده شده در DLA، یادگیری عمیق است که به‌صورت پیشرفته‌ای توسعه داده شده است. نتایج این تحقیق روی مجموعه داده توئیتر برای طبقه‌بندی کاربران اعمال شده است که اطلاعات مؤثری را در طول زمان گزارش می‌دهد.

فصیح فر و همکاران (۱۳۹۶)، روش‌های مبتنی بر هوش مصنوعی جهت تشخیص اجتماعات در شبکه‌های اجتماعی را مورد بررسی قرار دادند [۱۱]. در این تحقیق روشی توسط الگوریتم‌های خوشه‌بندی، پیشنهاد شده که شبکه را خوشه‌بندی نموده، سپس با کمک روش‌های تحلیل پیوند، اجتماعات را در هر یک از این خوشه‌ها تشخیص می‌دهد.

لو و همکاران (۲۰۱۴)، الگوریتم‌ها و برنامه‌های کاربردی برای تشخیص اجتماع در شبکه‌های وزن‌دار را مورد بررسی قرار دادند [۱۲]. در این مقاله به رفع مشکل تشخیص اجتماع در شبکه‌های وزن‌دار پرداخته شده است که استخراج اجتماع‌ها، جابه‌جایی داده در DNT و محدود کردن فعالیت الگوریتم کرم شب‌تاب در شبکه‌های اجتماعی آنلاین را ساده می‌کند و نشان می‌دهد که اجتماع چگونه می‌تواند این طرح‌های شبکه را تسهیل ببخشد.

ژانگ و همکاران (۲۰۱۳)، کشف جامعه کاربری از شبکه‌های چند ارتباطی را پیشنهاد دادند [۱۳]. در این مطالعه، یک قالب کار واحد برای کشف اجتماع‌های کاربری از شبکه‌های چند ارتباطی، پیشنهاد داده شده است. نتایج این روش نشان می‌دهد که محتویات تولید شده توسط کاربران دارای نقشی اساسی در شناسایی اجتماع‌های کاربری در شبکه‌های اجتماعی دارد.

سان و همکاران (۲۰۱۵)، الگوریتم انتشار برچسب بر مبنای مرکزیت برای شناسایی اجتماع در شبکه‌های اجتماعی را با نام CenLP معرفی کردند [۱۴]. این الگوریتم یک تابع جدید را برای اندازه‌گیری مرکزیت گره‌ها به‌طور کمّی بدون هیچ‌گونه تعامل با کاربر به‌وسیله محاسبه تراکم محلی و شباهت با همسایه‌های با تراکم بالا برای هر گره معرفی می‌کند. بر اساس مرکزیت گره‌ها، یک الگوریتم انتشار برچسب جدید را با ترتیب به‌روزرسانی خاص و اولویت گره برای کشف کردن اجتماعات در شبکه‌های مقیاس بزرگ به‌طور اتوماتیک بدون تحمیل کردن هیچ محدودیت از پیش، ارائه می‌شود.

یو و همکاران (۲۰۱۶)، تشخیص جامعه در شبکه‌های پیچیده با استفاده از الگوریتم خوشه‌بندی مبتنی بر تراکم و یادگیری چند ظرفیتی را مطرح ساختند [۱۵]. در این تحقیق چارچوب خوشه‌بندی مبتنی بر تراکم برای شناسایی ساختار جامعه، چگالی پارتیشن بهبود یافته برای ارزیابی کیفیت جوامع شناسایی، چارچوب به پارامترهای آن حساس است و آسان برای اجرای آن و مقایسات انجام شده بر روی معیارهای مصنوعی و شبکه‌های دنیای واقعی اثربخشی چارچوب پیشنهاد و بررسی شده است.

ژانگ و همکاران (۲۰۱۷)، الگوریتم انتشار برچسب برای تشخیص جامعه بر اساس اهمیت گره و تأثیر برچسب را پیشنهاد دادند [۱۶]. این الگوریتم به خاطر مزایایی که دارد مورد توجه زیادی قرار گرفته است، از جمله اینکه پیچیدگی زمانی خطی دارد و نیازی نیست تابع هدف و تعداد جامعه از قبل معلوم باشد. بااین‌حال، عیب آن نااطمینانی و تصادفی بودن در فرآیند انتشار برچسب است که دقت و پایایی جامعه را تحت تأثیر قرار می‌دهد.

عظوظی و رمضان (۲۰۱۷)، الگوریتم انتشار برحسب بر اساس تأثیر مبتنی بر شواهد برای تشخیص جامعه توزیع‌شده در شبکه‌های اجتماعی را مورد بررسی قرار دادند [۱۷]. در این مدل تخصیص انبوه و میزان معقول بودن بر اساس تأثیر اجتماعی برای تشخیص برچسب دامنه هر گره محاسبه می‌شود. آزمایش این مدل در دنیای واقعی و شبکه‌های LFR مصنوعی، کارایی این الگوریتم را در مقایسه با پیشرفته‌ترین الگوریتم‌ها نشان می‌دهد.

چی و همکاران (۲۰۱۳)، تشخیص جوامع همپوشان در شبکه‌ها: جدیدترین تکنولوژی و مطالعه تطبیقی را بررسی کردند [۱۸]. این مقاله جدیدترین تکنولوژی الگوریتم تشخیص جامعه همپوشان، معیارهای و مقیاس‌های کیفی را مورد تجدیدنظر قرار می‌دهد. پس از در نظر گرفتن اجرای تشخیص در سطح جامعه که با اطلاعات مفعل و انفعال نرمال‌سازی شد ارزیابی‌شده‌اند، شاخص امگا و اجرای تشخیص در سطح گره که با نمره F ارزیابی می‌شود.

ژانگ و همکاران (۲۰۱۷)، دست‌کاری ترندهای توییتر: اولین مطالعه از دیدگاه امنیتی به ترند توییتر به بررسی اینکه آیا ترندهای توییتر از دست‌کاری‌ها توسط حساب‌های مخرب ایمن هستند یا خیر پرداختند [۱۹]. در این روش و در مرحله تحلیل از یک مدل نفوذ خطی استفاده کردند تا تأثیر شبکه بر انتشار یک موضوع در توییتر را بسنجند که آن‌ها را قادر ساخت تا شواهدی از دست‌کاری را پیدا کنند. کاربرد این مدل محدود به سناریوهای خطی است.

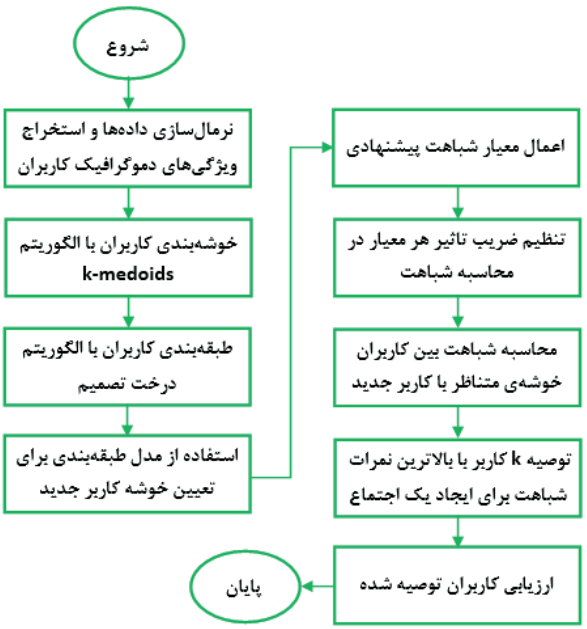
جاین و همکاران (۲۰۱۶)، در خصوص تشخیص اتوماتیک و بلادرنگ اطلاعات غلط در توییتر انجام دادند [۲۰]. از آنجا که اطلاعات غلطی که در توییتر درحال گردش می‌باشد، عمدتاً از منابع معتبر نمی‌باشند ازاین‌رو رویکرد ما بر اساس فرضیه حساب‌های کانال‌های خبری تائید شده به‌عنوان منابع معتبر اطلاعات شروع شد و با مقایسه این اطلاعات با اطلاعات مشابه به‌دست‌آمده از حساب‌های کاربری ناآشنا (عموم مردم) ادامه می‌یابد.

ژائو و همکاران (۲۰۱۵)، پیش‌بینی لینک‌ها و وزن‌ها در شبکه‌ها را با مسیرهای قابل‌اطمینان ارائه داده‌اند [۲۱] . هدف از پیش‌بینی لینک کشف لینک‌های گم‌شده و یا پیش‌بینی ظهور روابط در آینده با توجه به ساختار شبکه جاری می‌باشد. بسیاری از الگوریتم‌های امروزی برای پیش‌بینی لینک در شبکه‌های اجتماعی بدون وزن ارائه شده‌اند. در این تحقیق ایده مسیرهای قابل‌اطمینان مطرح شده است. یک مسیر قابل‌اطمینان به‌منظور توسعه شاخص شباهت محلی بدون وزن به شاخص‌های وزن‌دار ارائه شده است.

۳- روش پیشنهادی

در این تحقیق روش جدیدی برای پیش‌بینی دقیق‌تر اجتماعات در شبکه‌های اجتماعی با استفاده از ویژگی‌های دموگرافیک کاربران و ساختار توپولوژی گراف شبکه‌ها ارائه می‌شود. هدف ایجاد یک سیستم تشخیص اجتماعات است. در روش پیشنهادی استفاده از داده‌های واقعی و پیش‌بینی توصیه به کاربران، به‌منظور ایجاد ارتباط بین کاربرانی با بالاترین شباهت مورد توجه می‌باشد.

روش پیشنهادی در این تحقیق از چند مرحله تشکیل شده است. در مرحله اول ویژگی‌های مورد نیاز از اطلاعات دموگرافیک کاربران استخراج و نرمال می‌شوند. اطلاعات استخراج‌شده به‌منظور قرار دادن کاربران در خوشه‌های مشابهی استفاده می‌شود. این عمل در مرحله دوم و با استفاده از الگوریتم خوشه‌بندی k-medoids انجام می‌گیرد. در مرحله سوم و به‌منظور تعیین خوشه کاربر جدید و تعیین مشابهت آن به یک گروه از کاربران از یک مدل طبقه‌بندی درخت تصمیم استفاده می‌شود. تعیین شباهت بین کاربر جدید و سایر کاربران با استفاده از یک معیار شباهت ترکیبی در مرحله چهارم محاسبه می‌شود. این معیار تنها بین کاربران خوشه‌ی متناظر با کاربر جدید شباهت محاسبه می‌شود. در نهایت در مرحله پنجم الگوریتم پیشنهادی، فهرستی از k کاربر با بالاترین نمرات شباهت را به‌عنوان یک اجتماع معرفی می‌کند و با استفاده از چندین شاخص دقت اجتماعات مورد ارزیابی قرار می‌گیرد. شکل ۱ فلوچارت روش پیشنهادی را نشان می‌دهد.



شکل ۱: معماری روش پیشنهادی

۳-۱- نرمال‌سازی و استخراج ویژگی‌ها از اطلاعات دموگرافیک کاربران

هدف این بخش استخراج اطلاعات دموگرافیک کاربران می‌باشد. در این تحقیق با توجه به شبکه اجتماعی استفاده شده ویژگی-های سن، جنسیت، مکان جغرافیایی، شغل، علاقه به موزیک و علاقه به فیلم از کاربران استخراج می‌شود. به‌منظور استفاده از مقادیر این ویژگی‌ها نیاز به نرمال‌سازی آن‌ها ضروری می‌باشد. با توجه به اینکه در پردازش داده‌ها با اعداد سروکار دارم و برخی از مشخصه‌های مجموعه داده دارای مقادیر کیفی هستند، لذا تغییر آن‌ها به مقادیر کمی برای انجام پردازش الزامی می‌باشد. برای نرمال‌سازی به هر مقدار ممکن و منحصربه‌فرد از هر ویژگی از نمایه موجود، یک عدد یکتا نسبت داده شود. سپس آن‌ها را با استفاده از روش مینیمم – ماکزیمم، نرمال‌سازی می‌کنیم. این روش داده‌ها را در بازه [0-1] قرار می‌دهد. مقدار نرمال شده j-مین ویژگی از کاربر i-ام با $X_{i,j}^{new}$ نشان داده می‌شود و به‌صورت رابطه زیر است.

$$X_{i,j}^{new} = \frac{X_{i,j} - \min\limits_{1<i\leq N}(X_{i,j})}{\max\limits_{1<i\leq N}(X_{i,j}) - \min\limits_{1<i\leq N}(X_{i,j})} \quad (۱)$$

در این رابطه $X_{i,j}$ مقدار ویژگی j-ام از i-مین کاربر است. N تعداد کل کاربران و M تعداد ویژگی‌های استخراج‌شده از اطلاعات دموگرافیک کاربران می‌باشد.

۳-۲- خوشه‌بندی کاربران

در این مقاله برای خوشه‌بندی کاربران الگوریتم k-medoids مورد توجه می‌باشد. در این خوشه‌بندی برای محاسبه گره‌های مشابه از معیارهای شباهت شناخته شده‌ای نظیر فاصله اقلیدسی، فاصله همینگ و غیره استفاده می‌شود. این معیارها، شباهت گره‌ها را بر اساس فاصله آن‌ها محاسبه می‌کنند، یعنی هرچقدر فاصله بین دو گره کمتر باشد، این دو گره شبیه‌تر بوده و درنتیجه در یک خوشه قرار داده می‌شوند. به‌طورکلی الگوریتم k-medoids برخی از مشکلات الگوریتم را برطرف کرده است. در این الگوریتم بجای کمینه نمودن مجموع مجذور اقلیدسی فاصله بین نقاط، مجموع تفاوت‌های فواصل جفت نقاط را کمینه می‌کنند.

۳-۳- تعیین خوشه کاربران جدید با درخت تصمیم

بعد از انجام خوشه‌بندی و قرار دادن کاربران در گروه‌های مشابه، جهت تعیین کلاس کاربر جدید مجموعه داده مربوط به اطلاعات دموگرافیک استخراج‌شده از کاربران طبقه‌بندی می‌شوند. به این صورت که کلاس هر نمونه (خوشه هر کاربر) با توجه به خروجی الگوریتم خوشه‌بندی مشخص می‌شود. در این مقاله از الگوریتم طبقه‌بندی درخت تصمیم ID3 برای تعیین خوشه کاربر جدید استفاده می‌شود. درخت تصمیم یکی از روش‌های قوی و متداول برای دسته‌بندی و پیش‌بینی است.

۳-۴- ایجاد اجتماعات اولیه کاربران

در ساختار توپولوژی شبکه‌های اجتماعی، پیش‌بینی احتمال وجود ارتباط میان دو کاربر بر اساس خصوصیت‌های کاربران و سایر ارتباطات مشاهده شده را پیش‌بینی لینک می‌گویند. پیش-بینی لینک در یک گراف به‌عنوان تعامل بین رأس‌های گراف تعریف می‌شود. یک شبکه اجتماعی به‌صورت $G = (V, E)$ در نظر گرفته می‌شود. جایی که هر یال ارتباط بین و کاربر u و v را به‌صورت $(u, v) \in E$ نشان می‌دهد.

معیار شباهت بین دو کاربر احتمال دوستی بین آن‌ها را در آینده نشان می‌دهد. در سیستم توصیه دوست برای هر کاربر آزمون باید میزان شباهت آن با تمامی کاربران آموزش محاسبه شده و کاربرانی با بالاترین نمره شباهت به‌عنوان دوست به کاربر آزمون توصیه شوند. با توجه به امکان توصیه چند دوست، در اینجا فهرستی از k-top کاربر با بالاترین نمره شباهت پیشنهاد می‌شوند.

همان‌طور که پیش‌تر گفته شد، در این تحقیق از ترکیب معیارها شامل پیوند دوست و کاتز برای محاسبه شباهت استفاده می‌شود. اگر M تعداد معیارهای انفرادی استفاده شده باشد، معیار شباهت ترکیبی پیشنهاد شده با رابطه زیر تعریف می‌شود.

$$sim(u,v)=\sum_{q=1}^{M} a_q \times score_q \qquad (۲)$$

جایی که $sim(u,v)$ شباهت بین دو کاربر u و v ؛ $scoe_q$ شباهت محاسبه شده توسط q-مین شاخص شباهت انفرادی و a_q ضریبی ثابتی برای نشان دادن تأثیر q-مین روش انفرادی در محاسبه شباهت نهایی می‌باشد.

مقدار این ثابت‌ها با توجه به درصد تأثیر هر یک از این معیارها تنظیم می‌شود. این مقادیر با استفاده از خصوصیت بهینه‌سازی این پارامتر و با معرفی یک تابع هدف مستقل، مقادیر پارامترها را با سعی و خطا در بهینه‌ترین حالت جستجو می‌شوند. تابع هدف ارائه شده به فرم $F(a,i)$ است که مقدار برازندگی ضرایب a را در تکرار i-ام نشان می‌دهد. این تابع به‌صورت زیر تعریف می‌شود.

$$F(a,i)=\frac{1}{H}\sum_{h=1}^{H} Precision(ds_h,a) \qquad (۳)$$

جایی که H تعداد مجموعه داده‌های واقعی مورد آزمایش و ds_h به h-مین مجموعه داده اشاره دارد. $Precision(ds_h,a)$ تابعی است که مقدار دقت (درستی) را به ازای h-مین مجموعه داده و ضرایب a محاسبه می‌کند. در اینجا ضرایب ثابت با سعی و خطا به‌صورت جدول ۱ تنظیم شده است.

جدول ۱: تخمین ضرایب با سعی و خطا

ضرایب	a_1	a_2
مقدار بهینه	۰,۷۷۳۰	۱,۰۰۰
معیار شباهت	پیوند دوست	کاتز

مقادیر به‌دست‌آمده حاکی از اهمیت بالای معیار شباهت کاتز در تعیین میزان شباهت بین دو کاربر در روش پیشنهادی می‌باشد. با توجه به گزارش عملکرد بالای معیار کاتز در تحقیقات مختلف به نظر می‌رسد این معیار در شبکه‌های بزرگ و با بکار بردن آن به‌صورت ترکیبی، عملکرد خوبی را نشان می‌دهد.

۳-۵- ایجاد اجتماعات نهایی

در این تحقیق برای ایجاد اجتماعات نهایی از ماتریس شباهت بین کاربران استفاده می‌شود. ازاین‌رو برای هر کاربر k کاربر با بالاترین امتیاز شباهت انتخاب شده و به‌عنوان یک اجتماع در نظر گرفته می‌شوند. اگر لیست کاربران تست را با V^P و لیست کاربران آموزش را با V^T نشان دهیم؛ آنگاه برای هر کاربر تست از مجموعه V^P نمره شباهت آن با تمام کاربران V^T در دسترس است. با مرتب‌سازی نزولی نمره شباهت، کاربرانی با امتیاز تشابه بیشتر در ابتدای لیست قرار می‌گیرند. با توجه به تعداد کاربران مشابه که با پارامتر top-k شناخته می‌شود، k کاربر ابتدایی لیست به‌عنوان یک اجتماع در نظر گرفته می‌شود.

۴- نتایج و بحث

در این تحقیق آزمایش‌ها گسترده‌ای را برای نشان داده برتری الگوریتم پیشنهادی ارائه شده است. به‌منظور بررسی عملکرد و ارزیابی الگوریتم پیشنهادی، مقایسه‌ای با سایر معیارهای شباهت معروف انجام شده است. این معیارها کاتز [۲۲]، پیوند دوست [۲۳]، جاکارد [۲۴] و آدامیک آدار [۲۵] می‌باشند. علاوه بر این، پارامتر top-k برابر ۲۰، حداکثر طول مسیر برابر ۳ و همچنین تعداد خوشه ۴ در نظر گرفته شده است.

پیاده‌سازی و تمامی آزمایش‌ها با استفاده از یک پردازنده اینتل هفت هسته‌ای با فرکانس 2.4 گیگاهرتز، حافظه 8 گیگابایت سیستم‌عامل ویندوز 10 نسخه Enterprise و با استفاده از نرم‌افزار متلب انجام گردیده است. همچنین نتایج به‌دست‌آمده از این بررسی به‌منظور حصول نتایج دقیق‌تر، میانگین ۱۰ مرتبه تکرار تست می‌باشد.

شبیه‌سازی روی مجموعه داده شبکه اجتماعی پوپک انجام شده است. با توجه به حجم بالای این مجموعه داده زیاد بودن تعداد

کاربران و لینک‌های در آن‌ها، در این حاضر به‌منظور ارزیابی الگوریتم پیشنهادی تعداد ۲۰۰۰ کاربر ابتدایی از هر دو مجموعه داده به همراه تمامی لینک‌های آن‌ها در نظر گرفته شده است. جدول ۲ مشخصات مجموعه داده‌های اصلی و استفاده شده را نشان می‌دهد.

جدول ۲: مشخصات مجموعه داده پوپک

مجموعه داده	تعداد گره‌ها	تعداد لینک‌ها
پوپک نسخه اصلی	۱۶۳۲۸۰۳	۳۰۶۲۲۵۶۴
پوپک نسخه استفاده شده	۲۰۰۰	۱۸۳۱۵

در این تحقیق از معیارهای درستی، یادآور، اندازه‌گیری F و AUC جهت ارزیابی و مقایسه نتایج استفاده شده است. معیار درستی نسبت تعداد کاربران مرتبط از لیست top-k (کاربرانی از لیست top-k متعلق به مجموعه E^P) به تعداد واقعی کاربران مرتبط (تعداد لینک‌ها برای کاربر هدف از مجموعه E) می‌باشد. معیار یادآور نسبت تعداد کاربران مرتبط از لیست top-k به تعداد کل کاربران مرتبط (تعداد کل کاربران مجموعه E^P برای کاربر هدف) است. معیار اندازه‌گیری F توصیف‌کننده میانگین هارمونیک نرمال از دو معیار درستی و یادآور است. این معیار با توجه به رابطه زیر محاسبه می‌شود.

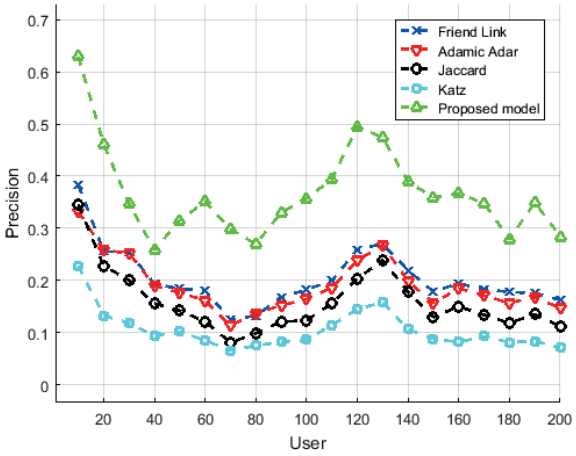
$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (۴)$$

در نهایت معیار AUC معادل ناحیه زیر منحنی ROC است، که هر چه مقدار این عدد مربوط به یک دسته بزرگ‌تر باشد کارایی نهایی مدل مطلوب‌تر ارزیابی می‌شود. رابطه زیر نحوه محاسبه این معیار را نشان می‌دهد.

$$AUC = \frac{n' + 0.5 \times n''}{n} \qquad (۵)$$

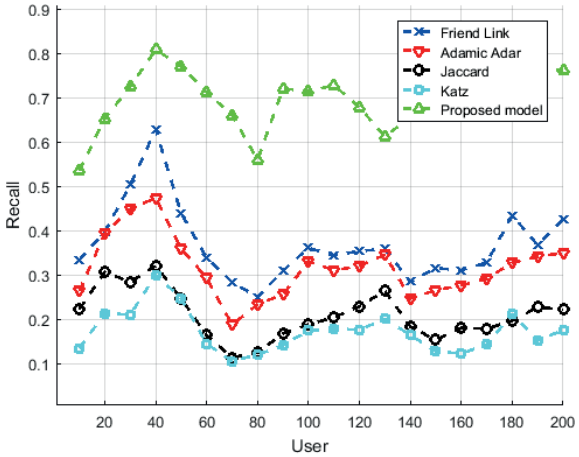
در این رابطه n تعداد کل مقایسه‌ها می‌باشد. درواقع این مقدار برابر تعداد لینک‌های موجود در مجموعه E^P است. در این رابطه منظور از لینک مفقود شده، لینک‌های موجود در E^P و منظور از لینک غیر موجود، لینک‌هایی که در ماتریس شباهت پیشنهاد شده ولی در E^P وجود ندارد. n' در این رابطه تعداد دفعاتی است که لینک مفقود شده مقدار شباهت بیشتری دارد. n'' نیز تعداد دفعاتی است که لینک‌های مفقود شده و لینک‌های غیر موجود مقدار شباهت یکسانی داشته باشند.

در اولین مقایسه معیار ارزیابی درستی برای الگوریتم پیشنهادی و سایر معیارهای شباهت محاسبه شده است. شکل ۲ نتایج این مقایسه را نشان می‌دهد.



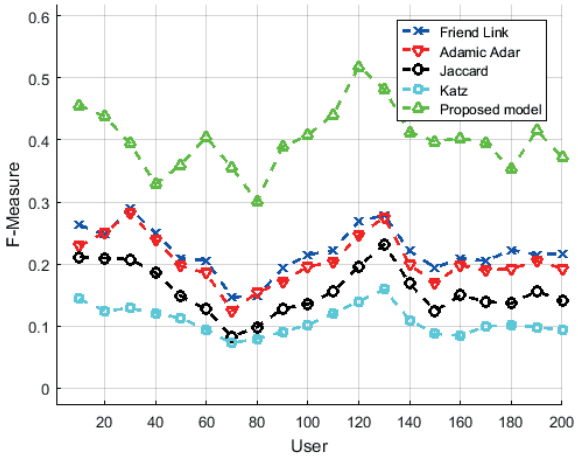
شکل ۲: نتایج مقایسه معیار درستی

نتایج نشان می‌دهد که استفاده از معیار شباهت درون خوشه‌ها باعث تمرکز الگوریتم روی کاربران با علایق مشابه شده و عملکرد الگوریتم ما را در مقایسه با سایر روش‌ها بالا برده است. در ادامه معیار یادآور برای روش‌های مورد مقایسه محاسبه شده و نتایج در شکل ۳ نشان داده شده است. مشاهده می‌شود که مقدار یادآوری به‌دست‌آمده در روش پیشنهادی نسبت به روش‌های دیگر بهتر می‌باشد.

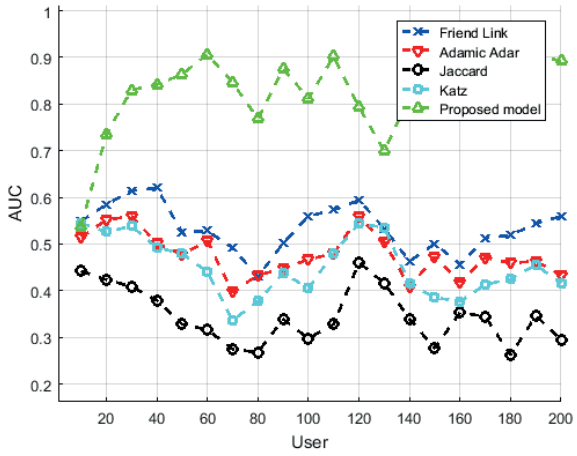


شکل ۳: نتایج مقایسه معیار یادآور

همچنین درصد بهبود الگوریتم پیشنهادی ازنظر معیار اندازه-گیری F نسبت به سایر الگوریتم‌ها در شکل ۴ به‌وضوح مشخص است.



شکل ۴: نتایج مقایسه معیار اندازه‌گیری F



شکل ۵: نتایج مقایسه معیار AUC

بر اساس نمودارها و نتایج حاصل می‌توان نتیجه گرفت که استفاده از خوشه‌بندی کاربران با استفاده از معیار شباهت ارائه شده که ترکیبی از شباهت ساختاری و شباهت دموگرافیک کاربران می‌باشد، باعث می‌شود که کاربرانی که دارای تعاملات، ویژگی‌ها و علائق یکسانی هستند در یک خوشه قرار بگیرند و درنتیجه پذیرش توصیه‌ها از جانب کاربران افزایش می‌یابد.

۵- نتیجه‌گیری

در تجزیه‌وتحلیل شبکه اجتماعی تشخیص اجتماعات یک گام اساسی برای درک ساختار و عملکرد شبکه است. برخی از روش-های تشخیص اجتماعات امروزی عملکرد محدودی دارند، زیرا تنها به ساختار توپولوژی در شبکه تمرکز می‌کنند. علاوه بر توپولوژی، اطلاعات محتوا یکی دیگر از جنبه‌های مهم شبکه‌های اجتماعی است. در این مقاله از یک روش ترکیبی برای تشخیص اجتماعات بر مبنای ویژگی‌های توپولوژیکی و دموگرافیکی استفاده شد. استفاده از الگوریتم معیار شباهت پیشنهادی ترکیبی و اعمال آن به درون خوشه‌ها برای تشخیص اجتماعات، سبب بالا رفتن دقت می‌شود، زیرا معیار شباهت پیشنهادی از ترکیب چند روش با ضریب تأثیر می‌باشد و این روش‌ها از اطلاعات محلی و سراسری گراف بهره می‌گیرند.

در سیستم پیشنهادی برای تعریف معیار شباهت ویژگی‌های نمایه کاربران از ویژگی‌های هر کاربر (مانند سن، جنسیت، شغل و غیره) استفاده شده است. استفاده از تعاملات روزانه کاربران (مانند کامنت‌گذاری در پست‌های افراد، برچسب‌گذاری مشترک عکس‌ها، نرخ‌دهی مشابه به محصولات و غیره) نیز می‌تواند برای تعریف معیار شباهت به کار گرفته شود تا به ایجاد توصیه‌هایی دقیق‌تر کمک کند.

مراجع

[1] Pennacchiotti, M., & Popescu, A. M. (2011). A Machine Learning Approach to Twitter User Classification. Icwsm, 11(1), 281-288.

[2] Malouf, R., & Mullen, T. (2008). Taking sides: User classification for informal online political discourse. Internet Research, 18(2), 177-190.

[3] Bhagat, S., Cormode, G., & Muthukrishnan, S. (2011). Node classification in social networks. In Social network data analytics (pp. 115-148). Springer, Boston, MA.

[4] 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.

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

[6] Sun, H., Liu, J., Huang, J., Wang, G., Yang, Z., Song, Q., & Jia, X. (2015). CenLP: A centrality-based label propagation algorithm for community detection in networks. Physica A: Statistical Mechanics and its Applications, 436, 767-780.

[7] 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.

[8] 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.

[۹] علیزاده، حسین، رسول حسین‌زاده و اسلام ناظمی (۱۳۹۶)، تشخیص اجتماعات ترکیبی در شبکه‌های اجتماعی، مجله مهندسی برق و الکترونیک ایران، ۱۱(۲)، صفحه ۴۹-۶۰.

[10] 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.

[۱۱] فصیح فر، زهره، زهرا عالیشاه، نجمه مشکانی و حمیدرضا رخصتی (۱۳۹۶)، بررسی روش‌های مبتنی بر هوش مصنوعی جهت تشخیص اجتماعات در شبکه‌های اجتماعی، پنجمین کنفرانس بین‌المللی مهندسی برق و کامپیوتر با تأکید بر دانش بومی، تهران، موسسه آموزش عالی مقدس اردبیلی

[12] 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.

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

[14] Sun, H., Liu, J., Huang, J., Wang, G., Yang, Z., Song, Q., & Jia, X. (2015). CenLP: A centrality-based label propagation algorithm for community detection in networks. Physica A: Statistical Mechanics and its Applications, 436, 767-780.

[15] You, T., Cheng, H. M., Ning, Y. Z., Shia, B. C., & Zhang, Z. Y. (2016). Community detection in complex networks using density-based clustering algorithm and manifold learning. Physica A: Statistical Mechanics and its Applications, 464, 221-230.

[16] Zhang, X. K., Ren, J., Song, C., Jia, J., & Zhang, Q. (2017). Label propagation algorithm for community detection based on node importance and label influence. Physics Letters A, 381(33), 2691-2698.

[17] 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.

[18] 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.

[19] Zhang, Y., Ruan, X., Wang, H., Wang, H., & He ,S. (2017). Twitter Trends Manipulation: A First Look Inside the Security of Twitter Trending. IEEE Transactions on Information Forensics and Security, 12(1), 144-156 .

[20] Jain, S., Sharma, V., & Kaushal, R. (2016). Towards automated real-time detection of misinformation on Twitter. Paper presented at the Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on.

[21] 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, 12261.

[22] Katz, L. (1953). A new status index derived from sociometric analysis. Psychometrika, Vol. 18. No. 1. 39-43

[23] Papadimitriou, A., Symeonidis, P., & Manolopoulos, Y. (2012). Fast and accurate link prediction in social networking systems. Journal of Systems and Software, 85(9), 2119-2132.

[24] Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013, March). Using of Jaccard coefficient for keywords similarity. In Proceedings of the International MultiConference

of Engineers and Computer Scientists (Vol. 1, No. 6).

[25] Adamic , L. A., & Adar, E. (2003). Friends and Neighbors on the Web, Social Networks, vol. 25(3), 211–230.