



# ارائه روشی با استفاده از الگوریتم ژنتیک در تشخیص موضع افراد جامعه در رسانه‌های خبری و اجتماعی

\*مهدی سالخورده حقیقی      \*\*سیدمحمد ابراهیمی

\*استادیار، دانشکده مهندسی کامپیوتر و فناوری اطلاعات - دانشگاه صنعتی سجاد - مشهد - ایران

\*\*دانشکده مهندسی کامپیوتر و فناوری اطلاعات - دانشگاه صنعتی سجاد - مشهد - ایران

تاریخ پذیرش: ۱۳۹۹/۰۵/۰۲

تاریخ دریافت: ۱۳۹۸/۰۲/۲۹

## چکیده

گزارش‌های خبری ارائه شده در رسانه‌های اجتماعی و خبری با انواع اسناد و مدارک ارائه می‌شوند و شامل موضوعاتی هستند که جوامع و نظرات مختلف را در برمی‌گیرند. آگاهی از رابطه میان افراد در اسناد می‌تواند به خوانندگان کمک کند تا یک دانش اولیه در خصوص موضوع و هدف در اسناد مختلف به دست آورند. در این مقاله، روشهای تشخیص جوامع بررسی شده و تکنیک‌های مختلف خوشه‌بندی افرادی که نام آن‌ها در مجموعه‌ای از اسناد خبری آورده شده است نیز مورد بررسی قرار می‌گیرد. این افراد در جوامعی خوشه‌بندی می‌شوند که مواضع مرتبط با یکدیگر دارند. در این مقاله یک روش تشخیص موضع افراد جامعه مبتنی بر یک شبکه دوستی به عنوان مکانیزم پایه معرفی شده و مکانیزم تشخیص جوامع بهبود یافته‌ای بر مبنای آن ارائه گردیده است. در روش پیشنهادی از ساختار الگوریتم ژنتیک جهت بهبود نرخ تشخیص جوامع بهبود یافته‌ای بر مبنای آن ارائه گردیده است. در روش در نظر گرفته شده است که برای رسیدن به این هدف شاخص رند نیز استفاده گردیده است. نتایج حاصل از آزمایش‌ها که بر مبنای پایگاه‌های داده‌ی واقعی اسناد انتشار یافته در رسانه خبری گوگل نیوز در رابطه با یک موضوع خاص به دست آمده‌اند، حاکی از کارآمدی و بهره‌وری مطلوب روش پیشنهادی است.

**واژه‌های کلیدی:** رسانه‌های اجتماعی، رسانه‌های خبری، تشخیص جوامع، تشخیص موضع افراد

## ۱- مقدمه

با توجه به گستردگی استفاده از این شبکه‌ها، تبادل نظرها بین کاربران این شبکه‌ها نیز به سرعت و با حجم زیادی صورت می‌گیرد. از این رو کاربران این شبکه‌ها به دنبال افراد و گروه‌هایی هستند که نظرات مشابه یا متضادی دارند تا به ارزیابی این

امروزه شبکه‌های اجتماعی در بسیاری از زمینه‌های اجتماعی، سیاسی، اقتصادی، فرهنگی، هنری، ورزشی، آموزشی و زمینه‌های دیگر بطور گسترده‌ای مورد استفاده قرار می‌گیرند. کاربران این شبکه‌ها نیز طیف‌های وسیعی از افراد با سطوح دانش متفاوتی از این شبکه‌ها را تشکیل می‌دهند.

گسترده هستند و قابلیت ارتباط آن‌ها را به روش‌های مختلف فراهم می‌کند. آن‌ها همچنین تعامل و اشتراک اطلاعات را با افراد مختلف از جمله بستگان، همکاران، خانواده، دوستان، طرفداران و دیگران تسهیل می‌نمایند [3]. علاوه بر تسهیل ارتباطات، رسانه‌های مذکور، به‌روزرسانی، دوست داشتن، دوست نداشتن، ایجاد پروفایل‌ها و به اشتراک‌گذاری اطلاعات شخصی و عمومی را فراهم می‌کنند [4].

یکی از چالش‌های اساسی در مورد تجزیه و تحلیل رسانه-ها/شبکه‌های اجتماعی، کشف خودکار جوامع است [5]. جوامع به‌عنوان گروه‌ها، خوشه‌ها، یا زیرگروه‌ها در موقعیت‌های مختلف می‌باشند و کشف جامعه در یک رسانه اجتماعی و خبری به معنی شناسایی مجموعه‌ای از گره‌ها است که با یکدیگر ارتباطات بیشتری نسبت به سایرین داشته باشند.

اکثر محققان بر این باورند که تشخیص جوامع در شبکه-های اجتماعی با مسائل خوشه‌بندی در داده‌کاوی قابل مقایسه است. خوشه‌بندی در داده‌کاوی نوعی یادگیری بدون نظارت است که هدف آن تقسیم یک مجموعه داده بزرگ به گروه‌های همگن (خوشه) می‌باشد [6]. در حقیقت، کشف یک جامعه می‌تواند داده‌کاوی بر روی گراف‌ها باشد. علاوه بر این، تشخیص جوامع، بزرگ‌ترین حوزه مطالعه از کاربردهای داده‌کاوی در رسانه‌های اجتماعی و خبری است. کاربردهای دیگر مانند گراف‌کاوی در مراحل اولیه توسعه قرار دارند [7]. روش‌های تحلیل گراف می‌توانند درک عمیق‌تری از ساختار شبکه‌های اجتماعی ارائه نمایند. البته روش‌های مبتنی بر گراف هزینه محاسباتی بالایی از نظر زمان و فضا دارند که استفاده از آن‌ها را در این رابطه محدود می‌نماید [8].

در این پژوهش، هدف اصلی تشخیص جوامع بر اساس موضع افراد در مجموعه‌ای از اسنادی است که در رسانه‌های اجتماعی و خبری منتشر می‌شوند به‌طوری‌که در نهایت افراد در گروه‌های مختلف با مواضع مرتبط خوشه‌بندی می‌شوند. در همین راستا، در این پژوهش شکل بهبود یافته یک روش خاص تشخیص جوامع به نام شناسایی جامعه موضع افراد براساس تحلیل شبکه دوستی ارائه شده است [9]. این بهینه‌سازی در جهت بهبود تشخیص جوامع و افزایش دقت خوشه‌بندی و همچنین کار بروی چالش مقداردهی اولیه و محلی بودن در روش

بپردازند. در نتیجه یافتن اطلاعات مورد نظر کاربران در میان انبوه اطلاعاتی که بطور پیوسته در این شبکه‌ها مبادله می‌شود یکی از چالش‌هایی است که کاربری این شبکه‌ها با آنها مواجه هستند.

علاوه بر موارد ذکرشده، تحلیل اطلاعات مبادله شده بین کاربران این شبکه‌ها می‌تواند راهی برای تحلیل رفتار آنها باشد تا بتوان از انجام برخی رفتارهای ناهنجار در این شبکه‌ها اطلاع حاصل نمود و از آنها جلوگیری کرد. لذا تشخیص وابستگی افراد به جوامع و گروه‌های مختلف از جمله زمینه‌های تحقیق جذاب در این شبکه‌ها می‌باشد.

روش پیشنهادی در این پژوهش یک روش بهبودیافته بر مبنای تشخیص جوامع بر اساس یک شبکه دوستی (گراف اولیه ارتباط بین گره‌ها) است. به عنوان نمونه، این روش گروه‌های مختلف افراد با گرایش‌ها و تمایلات یکسان را تشخیص می‌دهد. جوامع تشخیص داده شده برای آشنایی با پدیده‌های اجتماعی و تعاملات انسانی کاربرد خواهند داشت. اساساً بر طبق این روش، یک شبکه دوستی براساس نام افراد در ساختار الگوریتم ژنتیک تعریف شده و براساس یک تابع برازندگی، افراد در چند خوشه دسته بندی می‌شوند. به‌طور خلاصه، این پژوهش به بررسی مسئله‌ی تشخیص جامعه یا گروه افراد هدف (بر مبنای موضع آن‌ها) می‌پردازد تا در نهایت بتواند افراد هدف در جوامع با مواضع یکسان را به‌صورت بهبودیافته خوشه‌بندی نماید.

رسانه‌های اجتماعی به عنوان گروهی از برنامه‌های مبتنی بر اینترنت تعریف شده‌اند و امکان ایجاد و تبادل محتوای تولید شده توسط کاربران را می‌دهند. رسانه‌های اجتماعی متفاوت از یکدیگرند؛ انواع پایگاه‌های رسانه‌های اجتماعی شامل رسانه‌های سنتی مانند روزنامه، رادیو و تلویزیون و رسانه‌های غیر سنتی مانند فیس‌بوک، توئیتر و غیره می‌باشند [1]. به‌طور جزئی‌تر، از رسانه‌های اجتماعی با عنوان شبکه‌های اجتماعی نیز یاد می‌شود. رسانه‌های خبری، نوعی از مجموعه رسانه‌های اجتماعی هستند که هدف آن‌ها ارائه اخبار به عموم مردم یا جامعه‌ی هدفی از مردم است. این موارد شامل رسانه‌های چاپی (روزنامه، مجله‌های خبری)، پخش اخبار (رادیو و تلویزیون و اخیراً اینترنت)، روزنامه‌های برخط، وبلاگ‌های خبری، و غیره است [2].

رسانه‌های اجتماعی و خبری توان بالقوه‌ی قابل‌توجهی در ارتباطات و تعامل بین افرادی دارند که از لحاظ جغرافیایی

استخراج شده و برای استفاده بعدی به یک ساختار قابل درک تبدیل می‌گردد. نویسندگان از هفت ویژگی کاربر (مانند سن، جنس و غیره) برای خوشه‌بندی و تجزیه و تحلیل استفاده کردند. در نتیجه، مشخص گردید ویژگی‌های فیس بوکی انتخاب شده، تعیین کننده خوشه‌های مهم و معنا داری بودند.

تحقیق‌های مختلفی در زمینه تشخیص جوامع در شبکه‌های اجتماعی صورت گرفته است که در ادامه برخی از آنها بطور خلاصه مورد بررسی قرار می‌گیرند. در تحقیق‌های جداگانه‌ای، چن و همکارانش [15] [16] به مسئله تشخیص جوامع همت گمارده‌اند. آن‌ها در این راستا، استفاده از روش تحلیل مؤلفه‌های اصلی [17] را معرفی و تشریح کرده‌اند. به بیان مشخص، آن‌ها به بررسی نشانه‌های ورودی‌های بردار ویژه‌ای که با بزرگ‌ترین مقدار ویژه در ارتباط است، پرداخته‌اند تا بتوانند گروه‌های مربوط به موضع افراد را شناسایی نمایند. این روش می‌تواند تنها مباحث و موضوعاتی که دو موضع در آن‌ها مطرح است را تحت پوشش قرار دهد؛ اما در عمل، در بسیاری از موارد، بیش از دو موضع در مباحث و موضوعات مطرح می‌شود.

چن و همکارانش [9] روشی برای شناسایی گروه یا اجتماعات ارائه نمودند که مبتنی بر موضع و جایگاه آنها می‌باشد. به موجب آن، به‌طور خودکار و با استناد به سندها و مدرک‌های مورد نظر، یک شبکه‌ی دوستی برای افراد هدف تهیه می‌شود. روش‌های توسعه جامعه و اصلاح جامعه با هدف تعریف جوامعی با مواضع و جایگاه‌های مشخص از افراد هدف در شبکه‌ی دوستی طراحی شده‌اند. همچنین شناسایی افرادی که موضع آن‌ها با موضع هدف نامرتبط است نیز با این روش‌ها قابل انجام است. نتایج حاصل از آزمایش‌ها و تجربیاتی که بر مبنای پایگاه‌های داده‌ی واقعی و حقیقی به عمل آمده است، حاکی از کارآمدی و بهره‌وری مطلوب این روش دارند و نشان می‌دهند که این روش، کارایی بهتری از بسیاری از رویکردهای معروف تشخیص جوامع و الگوریتم‌های خوشه‌بندی دارد.

یادهوتموجو و همکارانش [18] یک روش تشخیص جامعه در یک شبکه استنادی با استفاده از الگوریتم رتبه پیوند ارائه نمودند. مقاله مذکور کاربرد یک الگوریتم تشخیص جامعه در یک شبکه استنادی را توصیف می‌نماید. هدف، تشخیص جوامع بر اساس روابط استنادی و تجزیه و تحلیل شباهت‌های موضوعات در هر جامعه است. نتایج آزمایش‌ها

مذکور است. در نهایت نیز روش پیشنهادی با چند روش دیگر خوشه‌بندی مورد مقایسه قرار گرفته که نتایج آزمایش‌ها نشان دهنده برتری روش پیشنهادی نسبت به سایر روش‌ها است.

## ۲- پیشینه تحقیق

اطلاعات رسانه‌های اجتماعی و خبری در قالب داده‌های متنی بدون ساختار مانند پیام‌های چت و اسناد متنی است. استخراج دستی اطلاعات از منابع متنی، که می‌تواند به یک پایگاه داده‌ی ساخت یافته و مناسب برای تجزیه و تحلیل بیشتر تبدیل شود، به‌ویژه هنگامی که داده‌ها حجیم باشند، قطعاً ناکارآمد و گاهی غیر ممکن است. بنابراین، استخراج خودکار اطلاعات یک رویکرد کارآمدتر در داده‌کاوی رسانه‌های اجتماعی برای تشخیص فعالیت‌های مجرمانه و همچنین روابط بالقوه میان شبکه است [10].

یکی از چالش‌های اساسی در مورد تجزیه و تحلیل رسانه‌های اجتماعی و خبری، کشف خودکار جوامع است [5]. جوامع به‌عنوان گروه‌ها، خوشه‌ها یا زیرگروه‌ها در مناطق مختلف دیده می‌شوند و کشف جامعه در یک رسانه اجتماعی به معنی شناسایی مجموعه‌ای از گره‌ها است که با یکدیگر ارتباطات بیشتری نسبت به سایرین داشته باشند. تشخیص جوامع در زمینه‌های مختلف در رسانه‌های اجتماعی مانند پیشنهاد‌های دوست، تقسیم مشتری، استخراج پیوند، رؤس برچسب‌گذاری و تحلیل نفوذ اجتماعی مفید است [11].

تعدادی از آثار به تشخیص جوامع با یک فرض بسیار محکم به نام جامعه اشاره می‌کنند. در یک جامعه، گروهی از رأس‌ها باید از یک ویژگی ساختاری بسیار محکم پیروی کنند. این کار شبیه مسئله معروف داده‌کاوی در تجزیه و تحلیل شبکه است (گراف کاوی). بعضی از نمونه‌های الگوریتم گراف کاوی در تعدادی از مقالات ارائه شده است [12] [6] [7] [13]. در تشخیص جامعه، تنها یک ساختار مهم وجود دارد و نتیجه دلخواه فهرستی از تمام گره‌های رأس است که ساختاری را در شبکه تشکیل می‌دهند.

تروساس و همکارانش [14] تشخیص جامعه را برای روابط بین خوشه‌های کاربر در فیس بوک برای آموزش زبان‌ها توصیف کردند. الگوریتم خوشه‌بندی K-means برای تعیین گروهی از کاربران با قابلیت‌ها و سبک‌های یادگیری یکسان، استفاده شد. اطلاعات از داده‌های کاربر

دو مرحله اطلاعات پیوندهای شبکه و اطلاعات گره‌ها را استفاده می‌کند. اطلاعات پیوندهای شبکه به جزءبندی ساختار شبکه به جوامع مختلف کمک می‌کند. اطلاعات گره‌ها، به تشخیص ناهنجاری‌های انفرادی کمک می‌کند. گائو و همکارانش [21] یک بهبود در روش بالا به‌وسیله یکپارچه‌سازی شبکه و اطلاعات گره‌ها برای تشخیص ناهنجاری‌های جوامع، ارائه کردند. روش پیشنهاد شده برای الگوریتم تشخیص ناهنجاری جامعه، از مدل ترکیب احتمالی طراحی شده برای شی داده چند متغیره (شیء با ویژگی‌های چندگانه) استفاده می‌نماید.

چن و همکارانش [22] یک روش تشخیص جامعه در شبکه‌های پیچیده با استفاده از حذف یال ارائه نموده‌اند. به منظور تشخیص ساختار جامعه در شبکه‌های پیچیده‌تر، الگوریتم جدیدی که به طور تکرار شونده یال‌ها را در جهت محدودیت‌ها حذف می‌کند، در مقاله مذکور ارائه شده است. این الگوریتم ابتدا از قدرت اتصال بین رأس‌ها برای تقسیم شبکه اصلی به برخی از جوامع استفاده می‌کند. همچنین قدرت اتصال بالا با قابلیت پیمان‌های بهینه با بهبود فرایند حذف یال را به کار می‌گیرد و در نهایت رئوس جدا شده را به جوامع اولیه برای بهینه‌سازی ساختار جامعه متصل می‌نماید.

یکی از چالش‌ها در تشخیص جوامع، وجود جوامع هم-پوشان است. در مقاله ارائه شده توسط رضوانی و همکارانش [23]، آنها ابتدا کیفیت جوامع هم‌پوشان را بررسی نموده و اثرات جداسازی آنها را نیز مورد توجه قرار می‌دهند. در تحقیق دیگری در رابطه با تشخیص جوامع هم‌پوشان، لیو و همکارانش [24] الگوریتمی دو مرحله‌ای را ارائه نمودند. در مرحله اول جوامع شناسایی شده و در مرحله دوم بر مبنای یک معیار شباهت، گره‌های پرت شناسایی می‌شوند. استفاده از توابع عضویت فازی نیز می‌تواند برای تشخیص این جوامع استفاده شود [25].

مسعودی و همکارانش [26] الگوریتم خفاش چند هدفه را برای تشخیص جامعه در شبکه‌های اجتماعی پویا به کار برده‌اند. بسیاری از الگوریتم‌های تکاملی برای حل مشکل تشخیص جامعه در شبکه‌های اجتماعی پویا پیشنهاد شده‌اند. بعضی از الگوریتم‌ها نیاز به پارامترهای پیش فرض دارند. دیگران از یک فرآیند تصادفی برای تولید جمعیت اولیه و برای اعمال اپراتورهای الگوریتم استفاده می‌کنند. این روشها فضای جستجو را افزایش می‌دهند و

نشان می‌دهند که الگوریتم ارائه شده قادر به شناسایی ۱۰۴۴۲ جامعه از ۱۸۸۵۱۴ گره است. پس از شناسایی جوامع، سه گروه برتر انتخاب شدند (کسانی که بیشترین تعداد اعضا را داشتند) و ۱۰ گره بالاترین امتیاز رتبه صفحه را در هر یک از این جوامع گرفتند. نمونه‌ها نشان می‌دهند که اکثر گره‌ها موضوع مشابهی دارند، اما هنوز گره‌هایی با موضوعات مختلف در یک جامعه مشابه وجود دارد. بررسی‌ها نشان داد که هفتاد درصد از گره‌ها موضوع مشابهی دارند، در حالی که سی درصد دیگر موضوعات متفاوتی دارند. با این وجود، مطالعه مذکور تأیید می‌کند که الگوریتم رتبه پیوند را می‌توان برای تشخیص جامعه در شبکه‌های جهت‌دار مورد استفاده قرار داد.

یک روش تشخیص ناهنجاری آماری بدون ناظر توسط ویسوانات و همکارانش [19] ارائه شده است که برای تشخیص رفتار غیرعادی در افراد کاربرد دارد. مجموعه داده‌های بدون برچسب فیس بوک در این تحقیق استفاده شده است و تعدادی از کاربران جعلی شناسایی شده‌اند. معیار استفاده شده برای جدا نمودن ناهنجاری‌ها و نرمال‌ها مقدار "پسندیدن"<sup>۱</sup> کاربران است.

در بسیاری از روش‌ها یک گراف ایجاد شده برای یک رسانه اجتماعی، به گروه‌ها یا جوامع مختلف تقسیم‌بندی می‌شود. این تقسیم‌بندی به‌وسیله حذف پیوندها از گره‌های مختلف یا به کاربردن الگوریتم‌های خاص خوشه‌بندی یا طبقه‌بندی، انجام می‌شود. برای مثال، ساختارهای جوامع به‌وسیله گیون و نیومن [20] بررسی شده‌اند. جوامع در قالب گروه‌های دوستی مختلفی هستند که در آنها قدرت پیوندها در بین گره‌ها در یک جامعه یا گروه دوستانه متراکم است؛ در حالی که در میان گره‌های دیگر خلوت است.

گائو و همکارانش [21] در زمینه‌ی تشخیص پرت‌های جوامع محلی و عمومی بسیار خوب کار کردند. یک روش ساده برای تشخیص پرت‌های جوامع روشی است که برای پرت‌های محلی و سراسری به کار می‌رود. برای مثال الگوریتم تشخیص پرت همسایه مستقیم، برای ناهنجاری‌های محلی، و الگوریتم تشخیص پرت سراسری برای ناهنجاری‌های سراسری به کار رفته‌اند. برای اینکه ناهنجاری‌های جوامع تشخیص داده شوند از اطلاعات موجود و هر گره جاری و همسایه‌هایش استفاده می‌گردد که این رویکرد، الگوریتم همسایه جامعه نامیده می‌شود که

منظور ابتدا یکی از روش‌های تشخیص جوامع معرفی می‌شود. روش مذکور که توسط چن و همکاران ارائه شده است [9]، تشخیص جامعه موضع افراد هدف با استفاده از تحلیل شبکه‌ی دوستی نام دارد که در این مقاله از آن برای ایجاد شبکه دوستی اولیه استفاده شده است تا در مراحل بعدی بهینه‌سازی بر روی آن انجام شود. در این روش، افرادی که نام آن‌ها در مجموعه‌ای از اسناد هدف آورده شده، در جوامعی با مواضع مرتبط خوشه‌بندی می‌شوند. این روش به این دلیل انتخاب شده است که خوشه‌بندی افراد براساس موضع آن‌ها در یک مجموعه از اسناد را به شکل مناسبی انجام می‌دهد بطوری که قابلیت بهبود بیشتر و بهینه‌سازی آن وجود دارد. اگرچه این روش انتخابی برای ایجاد شبکه دوستی اولیه بهترین روش نمی‌باشد، ولی انتخاب آن برای استفاده در روش پیشنهادی در این تحقیق با هدف نشان دادن قابلیت روش پیشنهادی برای بهینه‌سازی در چنین مسائلی است. نتایج حاصل از آزمایش‌ها که بر مبنای پایگاه‌های داده‌ی واقعی به دست آمده‌اند، حاکی از کارآمدی و بهره‌وری مطلوب روش پیشنهادی نسبت به روش پایه و دیگر روش‌های تشخیص جوامع است و نشان می‌دهد که این روش، کارایی بهتری نسبت به بسیاری از رویکردهای معروف تشخیص جوامع و الگوریتم‌های خوشه‌بندی دارد.

### ۳-۱- معرفی روش تشخیص جوامع

#### SCIFNET

در این روش تشخیص موضع افراد جامعه، آنهایی که نام آن‌ها در اسناد هدف آورده شده است، در جوامعی با مواضع یکسان طبقه‌بندی می‌شوند. شکل ۱ معماری این روش را نشان می‌دهد که از سه بخش تشکیل شده است: ساختار شبکه‌ی دوستی، توسعه جامعه موضع و اصلاح جامعه موضع. به بیان مشخص، روش مذکور ضمن استناد به یک مجموعه از اسناد و اخبار که گزارش‌هایی پیرامون یک مبحث با جوامع موضع خاص در آن‌ها آمده است، افراد هدف نامبرده در اسناد و اخبار را تشخیص می‌دهد. متعاقباً، یک شبکه‌ی دوستی برای افراد هدف ساخته می‌شود که مبتنی بر وقوع هم‌زمان نام افراد در اسناد و گرایش مواضع آن‌ها است. سپس، فرآیند توسعه جامعه موضع روی می‌دهد که به موجب آن، شناسایی جامعه موضع افراد هدف مانند یک فرآیند تشخیص جامعه تلقی می‌شود و گروه‌های موضع مشخص به‌طور مکرر در شبکه‌ی دوستی تکرار

باعث پیچیدگی فضایی می‌شوند. برای غلبه بر این نقاط ضعف، الگوریتم جدید چند هدفه‌ای با استفاده از الگوریتم انتقال میانگین برای تولید جمعیت اولیه برای به دست آوردن راه حل‌های با کیفیت بالا پیشنهاد شده است.

در شبکه‌های بزرگ و پیچیده، دستیابی به اطلاعات سراسری شبکه به منظور تشخیص جوامع بسیار دشوار یا حتی در بعضی موارد غیر ممکن است. در نتیجه از روش‌های تشخیص جوامع محلی برای این منظور استفاده می‌شود. در تازه‌ترین تحقیق انجام شده توسط لو و همکارانش [27] یک روش تشخیص جوامع محلی با استفاده از اطلاعات نزدیک‌ترین گره‌ها با مرکزیت بالا معرفی شده است. در زمینه تشخیص جوامع محلی، لو و همکارانش در ضمن تشکیل جامعه، به صورت پویا بر آن نظارت کرده و تابع عضویت مناسب را پیشنهاد می‌دهند [28].

روش‌های مبتنی بر جمعیت نیز برای تشخیص جوامع مورد استفاده قرار می‌گیرند. در تحقیقی که اخیراً توسط ون و همکارانش صورت گرفته [29]، چالش هم‌پوشانی جوامع در روش‌های تشخیص جوامع مورد بررسی قرار گرفته است. آنها از چند معیار و تابع چند هدفه برای تشخیص هم‌پوشانی در یک الگوریتم تکاملی مبتنی بر جمعیت استفاده نمودند. در بین الگوریتم‌های تکاملی تشخیص جامعه، استفاده از الگوریتم وال است که توسط زانگ و همکارانش اخیراً ارائه گردیده است [30].

در مقاله ارائه شده توسط پن و همکارانش [31]، روشی برای تشخیص جوامع ارائه شده است که از یک معیار شباهت محلی برای شناسایی نزدیک‌ترین همسایه‌های هر گره استفاده می‌نماید. سپس با ترکیب زوج گره‌ها زیرگراف‌های مرتبطی را ایجاد می‌کند و در نهایت جوامع را ایجاد می‌کند.

پژوهش حاضر، به بررسی مسئله‌ی تشخیص جامعه یا گروه افراد هدف (بر مبنای موضع آن‌ها) می‌پردازد تا در نهایت بتواند افراد هدف را در جوامع با مواضع یکسان، به صورت بهینه‌سازی شده خوشه‌بندی نماید.

### ۳- روش پیشنهادی

در این بخش روش پیشنهادی تشریح خواهد شد، که به موجب آن افراد هدفی که نام آن‌ها در مجموعه‌ای از اسناد آورده شده است، براساس یک تابع برازندگی و ساختار الگوریتم ژنتیک خوشه‌بندی خواهند شد. به این

حد واقعی نشان داده می‌شود. افراد هدفی که نام آن‌ها مکرراً در اسناد و اخبار مربوط به روابط موافق (مخالف) در کنار یکدیگر ظاهر می‌شوند، احتمالاً یک رابطه‌ی دوستانه (مخالفانه) دارند. روش پایه انتخابی، برای تعیین کمیت گرایش موضع در یک متن یا سند هدف، از روش ترنی و لیست‌من<sup>۵</sup> [35] استفاده نموده و ارزش (وزن) موضع، براساس رابطه (۱) محاسبه می‌شود:

$$(1)$$

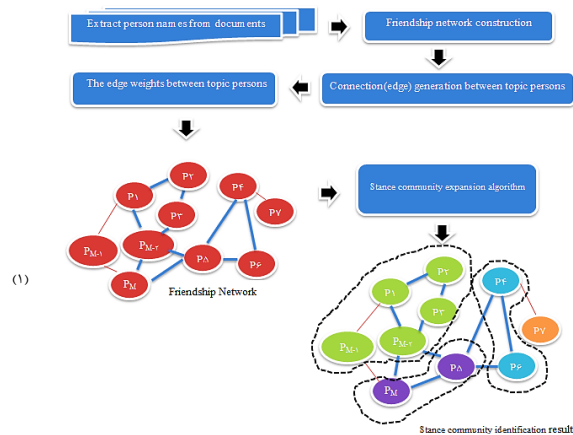
در این عبارت،  $SW_d$  نشان‌دهنده وزن یا ارزش موضع در سند  $D$  است؛  $Fwords$  و  $Owords$  به ترتیب مجموعه‌های لغات با مفاهیم دوستانه و مخالف را نشان می‌دهند که توسط متخصصان زبان‌شناسی شناسایی شده‌اند. تابع  $count(word_i, word_j)$  تعداد اسناد و اخباری که در آن‌ها،  $word_i$  و  $word_j$  به‌طور هم‌زمان ظاهر شده‌اند را محاسبه می‌کند. اساساً، در این معادله از اطلاعات دوسویه‌ی نقطه‌به‌نقطه<sup>۶</sup> (PMI) برای محاسبه وزن موضع یک سند استفاده شده است. وزن موضع  $SW_d$  مثبت خواهد بود اگر محتوای  $D$  ارتباط قوی و فراوانی با  $Fwords$  داشته باشد. چن و همکاران [9]، ضریب همبستگی موضع محور (SOCOR) زیر را که مقدار/ ارزش یا وزن موضع را در ضریب همبستگی تلفیق می‌کند، طراحی کرده‌اند (رابطه ۲):

$$A = \left[ \sum_{d \in D_{friendly}} SW_d * (p_{i,d} - \bar{p}_{i,friendly}) * (p_{j,d} - \bar{p}_{j,friendly}) + \sum_{d \in D_{opposing}} SW_d * (p_{i,d} - \bar{p}_{i,opposing}) * (p_{j,d} - \bar{p}_{j,opposing}) \right]$$

$$B = \sqrt{\sum_{d \in D_{friendly}} [\sqrt{SW_d} * (p_{i,d} - \bar{p}_{i,friendly})]^2 + \sum_{d \in D_{opposing}} [\sqrt{|SW_d|} * (p_{i,d} - \bar{p}_{i,opposing})]^2}$$

$$C = \sqrt{\sum_{d \in D_{friendly}} [\sqrt{SW_d} * (p_{j,d} - \bar{p}_{j,friendly})]^2 + \sum_{d \in D_{opposing}} [\sqrt{|SW_d|} * (p_{j,d} - \bar{p}_{j,opposing})]^2}$$

می‌گردند. در مرحله‌ی آخر، الگوریتم اصلاح جامعه شروع به کار کرده و نتیجه‌ی شناسایی را بر طبق یک تابع هدف، بهبود می‌بخشد.



شکل ۱: ساختار روش پایه تشخیص جامعه [9]

### ۱-۱-۳- ساخت یک شبکه‌ی دوستی

فرض کنید که  $D = \{d_1, d_2, \dots, d_3\}$  مجموعه‌ای از اسناد و اخبار هدف است و  $P = \{p_1, p_2, \dots, p_M\}$  یک مجموعه از افراد هدف می‌باشد که نام آن‌ها در مجموعه‌ی  $D$  آمده است. طی فرآیند ساخت شبکه‌ی دوستی، یک شبکه‌ی دوستی با عنوان  $G = \{P, E\}$  ساخته می‌شود که در آن، افراد هدف موجود در مجموعه  $P$ ، گره‌های شبکه را تشکیل می‌دهند؛ و  $E = (p_i, p_j)$  مجموعه‌ای از یال‌ها است که گرایش‌های دوستی افراد هدف را نشان می‌دهد (یعنی اینکه آیا رابطه‌ی میان افراد هدف، دوستانه است یا مخالفانه). به‌طور معمول، کشف گرایش‌های دوستی از روی متون دشوار است. باین‌حال، هریس<sup>۲</sup> [32] به این نتیجه رسیده است که واحدهای متنی با معانی متناقض و مغایر به‌ندرت در یک بافت مشابه ظاهر می‌شوند. علاوه بر آن، کانایاما و ناسوکاوا<sup>۳</sup> [33] نیز ثابت کرده‌اند که واحدهای متنی با گرایش‌های مشابه، در کنار یکدیگر ظاهر می‌شوند تا بافت را منسجم سازند. براین اساس، ضریب همبستگی<sup>۴</sup> [34] که وقوع هم‌زمان درجه‌ی افراد هدف در  $D$  را ارزیابی می‌کند، احتمالاً یک معیار یا شاخص مطلوب برای اکتشاف گرایش دوستی میان افراد هدف خواهد بود. باین‌حال، اسناد هدف گاهی مسائل جنجال‌برانگیز را تحت پوشش قرار می‌دهند. در این اسناد و اخبار، افراد با مواضع مختلف به‌شکلی یکدیگر را مورد انتقاد قرار می‌دهند. بنابراین تنها ضمن لحاظ کردن وقوع هم‌زمان درجه‌ی افراد هدف در  $D$ ، میزان دوستی رقبا بیش از حد واقعی و عملکرد تشخیص جامعه موضع، کمتر از

یا تضاد همسایگی) باعث افزایش قدرت دوستی می‌شود. و برعکس در حالت عدم دوستی یا تضاد یعنی  $SOCOR(p_i, p_j) < -\theta$  قدرت دوستی منفی خواهد بود که نشانه قدرت عدم دوستی یا تضاد است.

### ۲-۱-۳- توسعه جامعه موضع

شکل ۲ یک نمونه از توسعه جامعه موضع و شکل ۳، نمودار الگوریتم توسعه جامعه را نشان می‌دهد. در نمودار شکل ۳، نماد  $P_{unlabeled}$  نشان‌دهنده یک مجموعه از گره‌های بی‌نشان است (یعنی افراد هدف). در ابتدا داریم:  $P_{unlabeled} = P$ ، یعنی تمام گره‌ها فاقد نشان هستند. این الگوریتم، گره‌های  $K$  را به‌طور تصادفی به‌عنوان دانه‌های جوامع موضع انتخاب کرده و به‌صورت تکرار شونده، جوامع را ضمن ادغام گره‌های بی‌نشان توسعه می‌بخشد. در هر بار تکرار، یک مجموعه از گره‌های بی‌نشان  $U$  که مستقیماً به یک گروه متصل هستند، شناسایی می‌شود:  $(U = \{p_i \in P_{unlabeled} | (p_i, p_j) \in E, p_j \in c_k, 1 \leq k \leq k\})$  هر گرهی  $p_i$  در  $U$  متعاقباً باهدف شناسایی یک نشان مناسب برای گروه، مورد بررسی قرار می‌گیرد. فرض کنید  $Z_i$  نشان‌دهنده مجموعه گروه‌هایی باشد که گرهی بی‌نشان  $p_i$  به‌طور مستقیم با آن‌ها در ارتباط است؛ این بدان معناست که خواهیم داشت:

$$Z_i = \{c_k | (p_i, p_j) \in E, p_j \in c_k, 1 \leq k \leq k\}$$

به‌عنوان مثال  $Z_4$  که در شکل ۲ نشان داده شده است، گروه‌های  $c_1$  و  $c_2$  را تشکیل می‌دهد. چن و همکاران [16]، امتیاز ادغام برای هر گروه  $c_k$  در  $Z_i$  را به‌صورت رابطه ۴ محاسبه می‌نمایند:

$$ms_{i,k} = \sum_{p_j \in c_k, (p_i, p_j) \in E} \delta(p_i, p_j) \quad (۴)$$

در این رابطه،  $ms_{i,k}$  نشان‌دهنده امتیاز ادغام  $p_i$  با  $c_k$  است. اساساً امتیاز ادغام یا همجوشی حاصل جمع وزن یال‌های مرتبط با  $p_i$  با جامعه موضع  $c_k$  است. ادغام  $p_i$  با یک گروه که امتیاز ادغام مثبت دارد، باید یک گروه با مواضع یکپارچه و منسجم به دست دهد. زمانی که بیش از یک گروه دارای امتیاز ادغام مثبت باشند، الگوریتم اقدام به ادغام  $p_i$  با آن گروهی می‌کند که حداکثر امتیاز ادغام را به خود اختصاص داده است. این مورد قابل اثبات خواهد بود که این عمل بیشترین بازدهی را برای تابع هدف به همراه دارد. به این نکته توجه داشته باشید که اگر اکثر گره‌ها در  $c_k$  یک رابطه مخالف با  $p_i$  داشته باشند، امتیاز ادغام منفی

$$SOCOR(p_i, p_j) = \frac{A}{B * C} \quad (۲)$$

در این عبارت  $D_{friendly} \subseteq D$  نشان‌دهنده‌ی مجموعه‌ای از اسناد هدف است که مقدار یا ارزش موضع آن‌ها مثبت است؛  $D_{opposing} \subseteq D$  مجموعه‌ای از اسناد هدف را نشان می‌دهد که مقدار یا ارزش موضع آن‌ها منفی است؛ و  $\bar{p}_i.friendly$  و  $\bar{p}_i.opposing$  بسامدهای میانگین  $P_i$  را نشان می‌دهند که به ترتیب در  $D_{friendly}$  و  $D_{opposing}$  ظاهر شده‌اند. بازه‌ی  $SOCOR(p_i, p_j)$  نیز همانند ضریب همبستگی، معادل  $[-1, 1]$  است. اگر وقوع هم‌زمان  $p_i$  و  $p_j$  در  $D$  مستقل از یکدیگر باشد، مقدار برابر با صفر لحاظ می‌شود. اما اگر  $p_i$  و  $p_j$  تمایل داشته باشند و به‌طور هم‌زمان در اسناد حاکی از روابط دوستانه (یا مخالف) حضور یابند، آنگاه  $SOCOR(p_i, p_j)$  مثبت (یا منفی) خواهد بود. چن و همکاران [9] سپس، گرایش دوستی را با توجه به ضریب همبستگی موضع محور بررسی و تعریف نموده‌اند.

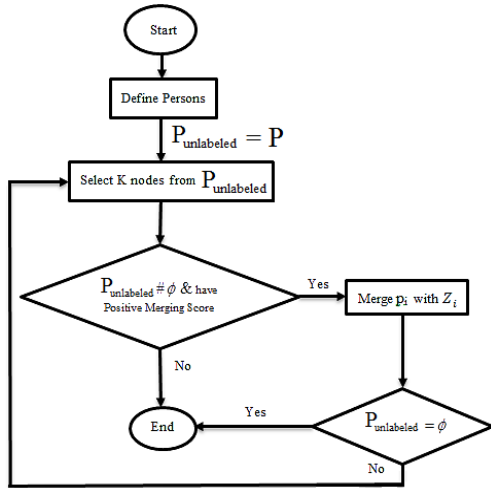
طبق تعریف اگر  $p_i \in P$  باشد، همسایه‌های دوست  $p_i$  با نماد  $\Gamma_{friendly}(p_i)$  نشان داده می‌شوند. همسایه‌های مخالف  $p_i$  با نماد  $\Gamma_{opposing}(p_i)$  نشان داده می‌شوند.

استحکام و قدرت همسایگی مشترک دوستانه میان  $p_i$  و  $p_j$  با نماد  $\gamma(p_i, p_j)$  نشان داده می‌شود و با استفاده از ضریب ژاکارد<sup>۷</sup> بدست می‌آید. قدرت و استحکام همسایگی مشترک مخالف میان  $p_i$  و  $p_j$  توسط  $\omega(p_i, p_j)$  نشان داده می‌شود که این نیز با استفاده از ضریب ژاکارد محاسبه می‌شود.

قدرت و استحکام دوستی که با نماد  $\delta(p_i, p_j)$  نشان داده می‌شود، نشان‌دهنده وزن یال  $(p_i, p_j)$  می‌باشد و براساس رابطه ۳ محاسبه می‌شود:

$$\delta(p_i, p_j) = \begin{cases} (SOCOR(p_i, p_j) + 1) \frac{\gamma(p_i, p_j) + \omega(p_i, p_j)}{2} & .if \ SOCOR(p_i, p_j) > \theta \\ -(|SOCOR(p_i, p_j)| + 1) \frac{\gamma(p_i, p_j) + \omega(p_i, p_j)}{2} & .if \ SOCOR(p_i, p_j) < -\theta \end{cases} \quad (۳)$$

در رابطه (۳) اگر رابطه بین  $p_i$  و  $p_j$  دوستانه باشد یعنی  $SOCOR(p_i, p_j) > \theta$  در این صورت افزایش مقدار توان  $\gamma$  قدرت دوستی همسایگی و  $\omega$  قدرت عدم دوستی



شکل ۳: نمودار الگوریتم توسعه جامعه موضع [16]

در الگوریتم پیشنهادی در هر بار تکرار، یک مجموعه از گره‌های مرزی با عنوان  $P_{boundary} \subseteq P$  تعریف می‌شوند. هر گره از  $P_{boundary}$  به یک گروه از مواضع تعلق داشته و به تعدادی گروه دیگر متصل است. به عبارت دیگر:

$$P_{boundary} = \{p_i | (p_i, p_j) \in E, p_i \in C_m, p_j \in C_n, m \neq n\}$$

صادق است. این الگوریتم، هر گره مرزی را مجدداً در آن جامعه موضع که بیشترین امتیاز ادغام را دارد، ادغام می‌کند. بدین طریق تازمانی که  $P_{boundary}$  خالی شده یا نتیجه‌ی شناسایی به ثبات برسد، گره‌های مرزی، شناسایی و خوشه‌بندی خواهند شد. اساساً، الگوریتم اصلاح جامعه پیشنهادی یک الگوریتم بالارونده است؛ چراکه نتیجه‌ی شناسایی را به‌طور تکرار شونده بهبود می‌بخشد. در الگوریتم پیشنهادی، با ایجاد هر جمعیت جدید از روی جمعیت قبلی، اعضای که مقدار تابع هدف آنها بیشترین مقدار است عیناً به جمعیت بعدی منتقل می‌شوند. به این ترتیب احتمال از دست دادن اعضای از جمعیت فعلی که ممکن است بتوانند مقادیر مطلوب‌تری برای تابع هدف در نسل‌های بعدی تولید کنند وجود نخواهد داشت. بنابراین انتظار می‌رود در جمعیت‌های بعدی تابع هدف غیر نزولی باشد.

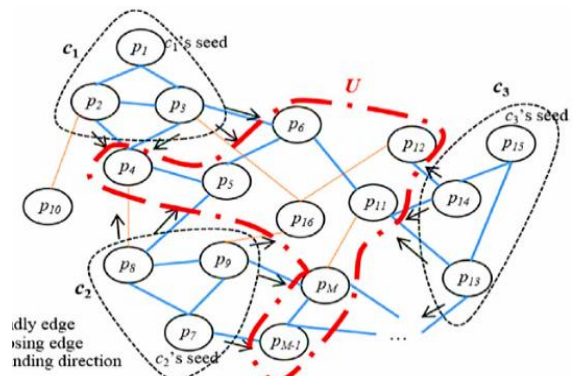
### ۲-۳- چارچوب الگوریتم پیشنهادی

بر طبق روش پیشنهادی، افرادی که نام آن‌ها در اسناد هدف آورده شده است، در جوامعی با مواضع یکسان طبقه‌بندی می‌شوند. شکل ۴، معماری روش پیشنهادی را نشان می‌دهد که از سه بخش تشکیل شده است: ساختار شبکه‌ی دوستی، ساختار الگوریتم ژنتیک (مقداردهی اولیه، ادغام، جهش و تابع برازندگی) و جامعه افرادی که تشخیص

خواهد بود. از آنجایی که ادغام  $p_i$  با یک گروه دارای موضع مخالف چندان مناسب نیست، این الگوریتم عملیات ادغام را مجدداً تکرار می‌کند؛ به شرطی که حداکثر امتیاز ادغام، منفی باشد. الگوریتم مذکور، گروه را به‌طور مکرر توسعه می‌بخشد؛ تا زمانی که تمام گره‌های بی‌نشان در شبکه‌ی دوستی، ادغام شده باشند یا هیچ گره‌ی بی‌نشانی دارای امتیاز ادغام مثبت با هیچ گروهی نباشد. سپس، یک نتیجه‌ی شناسایی گروه به دست خواهد آمد که با استفاده از الگوریتم اصلاح جامعه موضع (شکل ۳)، اصلاح می‌گردد.

### ۳-۱-۳ اصلاح جامعه موضع

الگوریتم توسعه جامعه موضع به‌طور تکرار شونده، گروه‌ها را از گره‌های دانه، توسعه و گسترش می‌دهد. در برخی از موارد، یک گره با یک گروه ادغام می‌شود، صرفاً به این دلیل که به دانه‌ای از آن گروه نزدیک باشد، حال آنکه بهتر است که با یک گروه دیگر ادغام شود. علاوه بر آن، نتیجه‌ی توسعه و گسترش، وابستگی بسیار زیادی به کیفیت دانه‌ها دارد. به‌منظور کمینه‌سازی اثر مسئله‌ی ادغام زودرس و کاهش تأثیر مقداردهی اولیه به دانه، الگوریتم پیشنهادی در این مقاله برای اصلاح یا پالایش جوامع موضع طراحی شده است. این الگوریتم، گروه‌ها را به‌طور تکرار شونده اصلاح می‌کند.

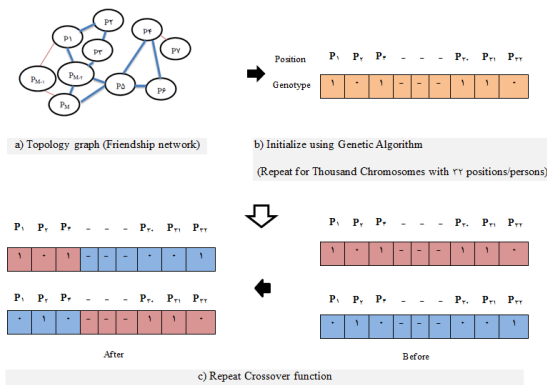


شکل ۴: نمونه از توسعه جامعه موضع [16]



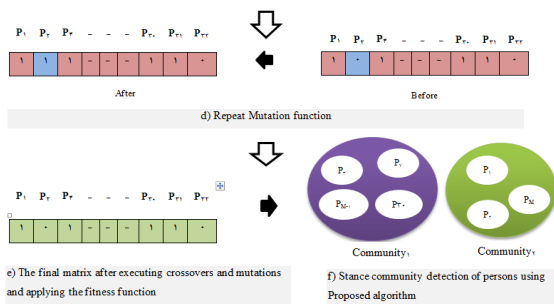
$$\text{Fitness function} = \sum_{i=1}^m \sum_{j=1}^m \delta(i,j) + \sum_{i=1}^m \sum_{j=1}^m \delta(i,j) - 2 \quad (5) \\ * \sum_{i=1}^m \sum_{j=1}^m \delta(i,j)$$

در این رابطه  $m$  تعداد افراد کلاس اول و  $n$  تعداد افراد کلاس دوم است. برای هر کروموزوم در هر نسل این رابطه تشکیل و خوشه‌بندی انجام می‌شود. لازم به ذکر است رابطه (۵) برای مسائلی با دو خوشه طراحی شده است. صورت نیاز می‌توان این رابطه را برای مسائلی با چند خوشه توسعه داد. با توجه به اینکه آزمایش‌ها در



شکل ۴: معماری الگوریتم پیشنهادی جهت تشخیص جامعه موضع افراد

در این رابطه اشخاصی که در کلاس اول هستند و با یکدیگر رابطه دارند (با موضع یکسان هستند) با سایر اشخاص همان کلاس مقدارشان جمع می‌شود. در عبارت دوم تابع مذکور نیز همین روال برای افراد کلاس دوم اتفاق



می‌افتد. در عبارت سوم چون هدف بهینه‌سازی است باید افرادی که در دو کلاس مختلف هستند مقدارشان براساس رابطه دلتا کاهش یابد پس یک مقدار منفی قبل از عبارت سوم درج می‌گردد و به دلیل اینکه دو کلاس وجود دارد در یک ضرب ۲ نیز ضرب می‌شوند.

داده شده‌اند. به بیان مشخص، روش مذکور ضمن استناد به یک مجموعه از اسناد و اخبار که گزارش‌هایی پیرامون یک مبحث در آن‌ها آمده است، افراد هدف نامبرده در اسناد و اخبار را تشخیص می‌دهد. یک شبکه‌ی دوستی برای افراد هدف ساخته می‌شود که مبتنی بر وقوع هم‌زمان نام افراد در اسناد و گرایش مواضع آن‌ها است. شبکه دوستی به‌عنوان جزئی از ساختار الگوریتم ژنتیک تعریف می‌شود. به بیان واضح‌تر، رابطه دلتا (رابطه ۳)، که نشان‌دهنده قدرت و استحکام دوستی است و به‌نوعی شبکه دوستی را ایجاد می‌نماید به‌عنوان جزئی از تابع برازندگی در ساختار الگوریتم ژنتیک مورد استفاده قرار می‌گیرد. سپس، ماتریس نام افراد به‌عنوان جمعیت اولیه در ساختار الگوریتم ژنتیک ایجاد و پس از اجرای توابع ادغام و جهش و ایجاد جمعیت‌های بعدی، این جمعیت‌ها توسط تابع برازندگی مورد ارزیابی قرار گرفته و بهترین ساختار کروموزوم (بردار)، به‌عنوان ماتریس نهایی انتخاب‌شده که در آن مشخص می‌شود افراد متعلق به چه جامعه‌ای هستند.

شکل ۴ روش ایجاد اعضای جمعیت و ایجاد جمعیت بعدی را نمایش می‌دهد و شکل ۵ الگوریتم روش پیشنهادی را نشان می‌دهد که بر طبق آن، افرادی که نام آن‌ها در اسناد هدف آورده شده است، در جوامعی با مواضع یکسان طبقه‌بندی می‌شوند. در این الگوریتم، ۳۲ شخص به‌عنوان یک ماتریس ۳۲ عضوی وارد ساختار الگوریتم ژنتیک می‌شوند. سپس تعداد بردارهای کروموزوم مشخص می‌گردد. اندازه جمعیت ۱۰۰۰ کروموزوم در نظر گرفته شده است که در تمام آزمایش‌ها ثابت است. هر کروموزوم ۳۲ عضو دارد (به تعداد اشخاص). در ادامه درصد ادغام‌ها و جهش‌ها مشخص می‌گردد. در الگوریتم شکل ۴ تعداد ۳۰۰ نسل تولید می‌شود که این تعداد در آزمایش‌ها ثابت در نظر گرفته شده است.

در مرحله بعد الگوریتم، رابطه بهینه‌سازی براساس رابطه دلتا تشکیل می‌گردد که همان تابع برازندگی است. رابطه (۵) که همان تابع برازندگی الگوریتم پیشنهادی است به جهت تشخیص یک جامعه منسجم طراحی گردیده است.

شکل ۵: الگوریتم پیشنهادی جهت تشخیص جامعه موضع افراد

#### ۴- آزمایش‌ها

در این بخش، مجموعه داده‌های مورد استفاده در آزمایش‌ها معرفی خواهد شد؛ عملکرد روش پیشنهادی با روش SCIFNET و برخی دیگر از روش‌های خوشه‌بندی در رابطه با تشخیص جوامع مقایسه خواهد شد. متعاقباً، یکی از نتایج تشخیص جامعه که توسط روش پیشنهادی تشخیص داده شده است، معرفی می‌شود.

۴-۱- مجموعه داده تشخیص جوامع موضع افراد، یک حوزه پژوهشی نسبتاً جدید به شمار می‌آید. در این پژوهش از یک مجموعه داده با عنوان Kroa که شامل ۷۴ سند و اخبار مرتبط با آن است که همه‌ی آن‌ها از رسانه خبری Google News دانلود شده‌اند، استفاده گردید. مباحث و موضوعات جمع‌آوری شده، حوزه‌ی مسائل سیاسی را پوشش می‌دهند. بعلاوه، از برخی متخصصان انسانی نیز درخواست گردیده است تا اسناد غیر مرتبط را به صورت دستی از مجموعه جدا کنند تا اطمینان حاصل گردد که اسناد تحت بررسی، کاملاً با موضوعات تناسب دارند. برای استخراج نام افراد هدف که در اسناد هدف آمده‌اند، قطعه کدی نوشته شده است. فراوانی نام افراد هدف ( $\lambda$ )، به میزان ۶۰ درصد از فراوانی کل اسامی هدف استخراج شده مشخص گردید. میانگین تعداد اسامی بررسی شده تحت هر وضعیت از ( $\lambda$ ) در جدول ۱ نشان داده شده است.

برای ارزیابی عملکرد، از شاخص رند<sup>۸</sup> که یک معیار ارزیابی خوشه‌بندی بسیار مهم است استفاده شده است [17]؛ زیرا در این روش، گروه‌های مختلف افراد هدف در خوشه‌های مختلف دسته‌بندی می‌شوند. شاخص رند، درصد تمام جفت‌های افراد را که به طور صحیح خوشه‌بندی شده‌اند نشان می‌دهد (یعنی اگر دو فرد با موضع یکسان در یک گروه واقع شده باشند یا دو فرد با مواضع مختلف در دو گروه مختلف قرار گرفته باشند). رابطه ۶ روش محاسبه این شاخص را نشان می‌دهد [17].

$$RI = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

در ادامه تابع ادغام انجام می‌شود که در آن، به طور تکرار شونده، دو کروموزوم (احتمال انتخاب کروموزوم‌ها با برابری بالاتر بیشتر است) انتخاب و ترکیب می‌شوند. در این تابع نوع ادغام، از نوع ادغام با یک نقطه قطع است و روش انتخاب کروموزوم‌ها به روش چرخ رولت می‌باشد. در ادامه الگوریتم پیشنهادی، جهش انجام می‌شود. در جهش به طور تکرار شونده، یک کروموزوم به صورت تصادفی انتخاب می‌شود و دو فیلد از آن به صورت تصادفی تعویض می‌شود. پس از انجام توابع ادغام و جهش و تولید نسل‌های جدید، بهترین کروموزوم بر اساس تابع برابری انتخاب می‌شود. در نهایت خروجی یک بردار است که به هر خانه مقدار ۰ یا ۱ نسبت داده که مشخص می‌کند هر شخص به کدام کلاس مربوط می‌شود (کلاس اول یا کلاس دوم).

جدول ۱: آمار نوشته‌های ارزیابی شده

topic	Number
# of topic documents	74
# of extracted topic persons	54
# of evaluated topic persons	32

**Proposed algorithm**

$P = P_{persons}$

**The input of the genetic algorithm is the number of persons based on their position**  
 $P_{persons} = \gamma\gamma$

$Maxiter = \gamma\gamma$  // Maximum number of replicas of the genetic algorithm

// Create a genetic algorithm structure

**While (Maxiter) do**

// According to  $\gamma\gamma$  Persons, the chromosome vectors (popsize) are  $\gamma\gamma$ -membered (nvar:  $P_{persons}$ )

// Determine percentage of crossover (pc) and mutations (pm) for the genetic algorithm

$popsize = 1000$      $nvar = P$      $pc = 0.7$

$ncross = \gamma * (\frac{popsize * pc}{\gamma})$

$pm = 1 - nvar$

$nmult = popsize * pm$

// Optimization relationship: Based on the delta matrix, the relation is as follows:

$$Fitness\ function = \sum_{i=1}^m \sum_{j=1}^m \delta(i,j) + \sum_{i=1}^n \sum_{j=1}^n \delta(i,j) - \gamma * \sum_{i=1}^m \sum_{j=1}^n \delta(i,j)$$

// m: is the number of first-class people and n: is the number of second-class people. For each chromosome in each generation, this relationship is formed and clustering is done.

// Carry out the crossover: the probable selection of  $\gamma$  chromosomes (the probability of chromosomes selection with higher fitness is higher) and their composition

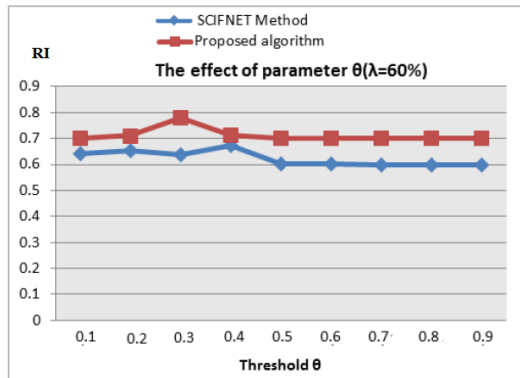
// Carry out the mutation: Choosing a completely random chromosome and replacing two houses together

// Select the best chromosomes for the next generation and perform crossovers and mutations to the specified number of repetitions

**End while**

**Return**  
 $C_1 = \{P_1, P_2, \dots, P_m\}$ ,  $C_2 = \{P_1, P_2, \dots, P_n\}$

// The output is a vector that assigns a value of 0 or 1 to each position, which specifies which Persons are related to which class.



شکل ۶: دقت الگوریتم پیشنهادی و الگوریتم SCIFNET

در روش پیشنهادی، امتیاز اندازه‌ها نیز به صورت یک‌روال خطی پیش می‌رود و فقط در  $\theta = 0.3$  افزایش می‌یابد که احتمالاً به دلیل ماهیت تصادفی بودن الگوریتم ژنتیک است. در روش پایه مقدار پایین  $\lambda$  نشان‌دهنده آن است که روال تشخیص جوامع موضع، دشوار است چرا که در این شرایط، فراوانی اسامی هدف در فرآیند شناسایی بسیار اندک است [16]. از آنجایی که روال ساخت یک شبکه دوستی با استناد به وقوع نام افراد در اسناد و اخبار هدف طی می‌شود، شمول نام افراد با فراوانی اندک می‌تواند کیفیت شبکه را کاهش داده و در نتیجه بر عملکرد تشخیص اثر بگذارد. اساساً، مقدار دو اندازه مذکور با افزایش مقدار  $\theta$  افزایش پیدا می‌کند زیرا  $\theta$  با مقدار بالا می‌تواند روابط نه‌چندان مهم میان افراد را پالایش و جداسازی کند تا کیفیت شبکه ارتقا یابد. زمانی که  $\theta$  بالاتر از ۰,۴ باشد، امتیاز اندازه‌ها به تدریج افت پیدا می‌کند. در این شرایط میان گره‌ها نیز ارتباط و پیوندی شکل نمی‌گیرد. در نتیجه شبکه‌ی دوستی، به قدری کم تراکم خواهد بود که ارتباطات اطلاعاتی میان افراد، بازنمایی نمی‌شود و عملکرد شناسایی گروه چندان با کیفیت نخواهد بود.

با توجه به اهمیت تعداد تکرار (شرط پایان) الگوریتم ژنتیک در روش پیشنهادی، شکل ۷ مقادیر شاخص رند را تحت شرایط  $\theta = 0.3$ ,  $\beta = 1$  با تعداد تکرارهای مختلف از الگوریتم ژنتیک نشان می‌دهد. با توجه به شکل مذکور و نتایج آزمایش‌ها تعداد تکرار سیصد مرتبه بهترین نتیجه را در بر خواهد داشت.

در این رابطه TP تعداد افراد با مواضع یکسان که به درستی خوشه بندی شده‌اند، TN تعداد افراد با مواضع متفاوت که به درستی در خوشه‌های مختلف قرار داده شده‌اند، FP تعداد افراد با مواضع یکسان که به اشتباه خوشه بندی شده‌اند و FN تعداد افراد با مواضع متفاوت که به اشتباه خوشه بندی شده‌اند را نشان می‌دهد. هرچه امتیاز شاخص رند نیز بیشتر باشد، کیفیت خوشه‌بندی بالاتر است. لازم به ذکر است رابطه (۶) برای مسائلی با چند کلاستر نیز قابل استفاده است. توضیح روش به کار گیری آن برای چنین مسائلی در این مرجع آمده است [17]. از آنجایی که الگوریتم توسعه جامعه مواضع، وابستگی بسیار زیادی به مقدار اولیه‌ی دانه‌ی گروه دارد و نیز روش پیشنهادی پژوهش وابسته به معیارهای تصادفی است، این روش با مقادیر تصادفی ده بار تکرار می‌شود. مقادیر شاخص رند برای موضوع Korea تحت بررسی، میانگین گیری می‌شوند تا عملکرد کلی روال تشخیص، مشخص گردد.

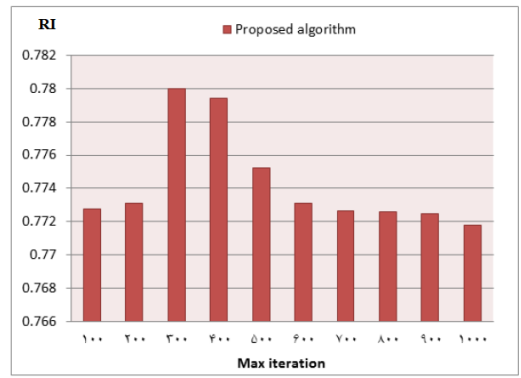
## ۲-۴- ارزیابی نتایج

در ابتدا، پارامتر  $\theta$  که نشان‌دهنده‌ی آستانه‌ی گرایش دوستی است و برای تعیین یال‌ها در یک شبکه‌ی دوستی کاربرد دارد در نظر گرفته می‌شود. در این آزمایش،  $\theta$  بین ۰,۱ تا ۰,۹ لحاظ می‌شود. جدول ۲، فهرست‌های Fwords و Owords را که توسط متخصص زبان‌شناسی تألیف و گردآوری شده‌اند، نشان می‌دهد [16] (یادآوری: Fwords و Owords به ترتیب مجموعه‌های لغات با مفاهیم دوستانه و مخالف را نشان می‌دهند که در رابطه‌ی ۱ به کار برده شده‌اند). از جدول مذکور در پیاده‌سازی‌ها و آزمایش‌های روش پیشنهادی و روش پایه استفاده شده است. فهرست لغات نشان‌دهنده مواضع، مورد استفاده‌ی ضریب همبستگی موضع محور بوده و برای محاسبه وزن یا ارزش موضع در یک سند هدف کاربرد دارد. پارامتر  $\beta$  که برای محاسبه قدرت و استحکام دوستی به کار می‌رود نیز برابر با ۱ در نظر گرفته می‌شود. شکل ۶، مقادیر شاخص رند را تحت شرایط مختلف  $\theta$  و  $\lambda$  نشان می‌دهد.

تمام این روش‌ها به یک شبکه ورودی احتیاج دارند. بعلاوه باید توجه داشت که SM و FEC برای شبکه‌های نشان‌دار<sup>۹</sup> طراحی شده‌اند. روش‌های FastModularity، CODA، SCAN، بر آن فرض استوار شده‌اند که شبکه‌های تحت بررسی، بدون نشان هستند و به همین علت، ساختارهای پیوند را به جهت شناسایی گروه‌ها بررسی می‌کنند. الگوریتم‌های K-means و HAC بازنمایانگر یک فرد هدف هستند و آن را در قالب یک بردار فراوانی N بعدی نشان می‌دهند که در آن، یک ورودی، فراوانی وقوع اسم افراد هدف در یک سند هدف را نمایش می‌دهد و برای ارزیابی رابطه میان افراد هدف، از تشابه کسینوسی استفاده می‌شود که برای تعیین تشابه بردارهای بسامد یا فراوانی کاربرد دارد. خوشه‌بندی در روش‌های CODA و K-means تا حد زیادی به مقاردهای اولیه آن‌ها وابسته است.

همان‌طور که در جدول ۳ ذکر گردیده است، روش پیشنهادی نسبت به روش پایه و سایر روش‌های ذکر شده در جدول، بهترین عملکرد را از خود نشان داده است. نتایج نشان می‌دهند که HAC و K-means افراد هدف مشهور را در کنار یکدیگر خوشه‌بندی می‌کنند. این بدان علت است که در تشابه کسینوسی، ضرب داخلی دو بردار فراوانی نرمال‌سازی شده است و اگر بردارهای محاسبه شده حاوی تعداد زیادی ورودی غیر صفر باشند، یک امتیاز تشابه بالا به دست می‌آید [16].

از آنجایی که اسامی افراد مشهور در بسیاری از اسناد و اخبار قید شده است، بردارهای فراوانی نرمال‌سازی شده حاوی تعداد زیادی ورودی غیر صفر خواهند بود. بنابراین، روش‌های خوشه‌بندی، رابطه میان افراد هدف مشهور را بالاتر از آنچه که هست برآورد می‌کنند و افراد مشهور دارای مواضع مختلف را در کنار یکدیگر خوشه‌بندی می‌نمایند. در نتیجه، عملکرد روش تضعیف می‌شود. عملکرد نه‌چندان قوی استراتژی پیوند منفرد در روش HAC، متأثر و ناشی از نقصان فوق است؛ چراکه طبق این استراتژی، تشابه دو خوشه از طریق بررسی شبیه‌ترین جفت افراد حاضر در خوشه انجام می‌شود. در نتیجه، این استراتژی خوشه‌های حاوی افراد مشهور را با یکدیگر ادغام می‌کند، حال آن‌که ممکن است این خوشه‌ها نمایانگر مواضع متفاوت باشند.



شکل ۷: دقت الگوریتم پیشنهادی تحت تاثیر تکرار الگوریتم ژنتیک

در ادامه، روش پیشنهادی با روش پایه و با شش روش دیگر تشخیص جامعه به نام روش‌های FastModularity [36]، [37]SCAN، [38]FEC، [21]CODA، [39]SM و دو روش شناخته شده خوشه‌بندی K-means و HAC [40] مقایسه خواهد شد. نتایج این مقایسه‌ها در جدول ۳ ذکر گردیده است.

جدول ۳: کارایی Rand index براساس  $\lambda=60\%$

Method	(Rand index)%
<b>proposed method</b>	<b>77.99</b>
SCIFNET	75.89
Fast Modularity	64.42
SCAN	65.23
CODA	68.87
SM	69.19
FEC	68.05
HAC (Single-Link)	53.23
K-means	75.1

	Good Team Work Partner Advocate Friend
Stance-opposing words (Owords)	Campaign Opposite Rival Fraud Accusation Contest Lost Beat Debate

آن، تابع هدف به‌سادگی می‌تواند مجموع وزن‌های یال‌ها در هر گروه را به حداکثر رسانده و رابطه میان گروه‌ها را نادیده انگارد. به همین علت، تعداد زیادی فرد مشهور با مواضع مختلف در یک گروه قرار می‌گیرند.

تابع هدف در روش پیشنهادی (که تابع برازندگی نامیده می‌شود) و همچنین روش پایه علاوه بر بیشینه‌سازی رابطه میان گره‌های موجود در گروه‌ها، رابطه بین گروه‌ها را به حداقل می‌رساند. بنابراین، روش پیشنهادی و روش پایه یک برتری در شناسایی گروه‌ها پیدا می‌کنند. روش FEC به‌طور معمول عملکرد بهتری نسبت به دیگر روش‌های تحت مقایسه دارد؛ این بدان علت است که روش مذکور برای شبکه‌های نشان‌دار طراحی شده و پیوندهای منفی در شبکه‌های دوستی ایجاد شده وجود ندارند. باین‌حال عملکرد روش پایه همچنان از FEC بهتر است. روش SM نیز برای شبکه‌های نشان‌دار طراحی شده است. اما SM گاهی نمی‌تواند برای یک مبحث تحت ارزیابی، K گروه بسازد؛ زیرا نشانه‌های ورودی‌ها در بردارهای ویژه‌ی اصلی همگی مثبت هستند. به همین علت، افرادی با مواضع متفاوت در یک گروه کنار هم قرار می‌گیرند. در نتیجه روش پایه عملکرد بهتری نسبت به روش SM خواهد داشت.

روش پیشنهادی عملکرد بهتری را نسبت به روش پایه و سایر روش‌های ذکر شده در جدول ۵ از خود نشان داده است. به‌طور خاص مزیت‌های روش پیشنهادی نسبت به روش پایه عبارت‌اند از: ۱- دقت خوشه‌بندی افزایش می‌یابد، ۲- عملیات خوشه‌بندی سراسری انجام می‌شود. در روش پایه عملیات خوشه‌بندی به‌صورت محلی انجام می‌گیرد، یعنی الگوریتم از یک نقطه شروع نموده و مقادری اولیه انجام می‌دهد (که ممکن است غلط باشد) سپس عضو جدید را با اعضای خوشه مقایسه می‌کند و بر اساس این‌که قدرت و استحکام دوستی (رابطه  $\delta(p_i, p_j)$ ) عضو جدید با کدام خوشه بیشتر است، عضو جدید به آن خوشه اختصاص می‌یابد و به‌تدریج خوشه‌ها تشکیل می‌شوند. در این حالت ممکن است به دلیل عدم شکل‌گیری کامل خوشه‌ها در این مرحله، یک گره (فرد) جدید به‌اشتباه به خوشه‌ای اختصاص یابد. اما در روش پیشنهادی از همان ابتدا با استفاده از روش ارائه‌شده، در ساختار الگوریتم ژنتیک همه حالت‌ها به‌صورت سراسری بررسی می‌شود تا به برازندگی بهتری برسد. ۳- در روش

همچنین روش پیشنهادی رابطه میان افراد هدف را بر اساس ساختار الگوریتم ژنتیک و یک تابع برازندگی تعیین می‌نماید. روش پایه رابطه میان افراد مشهور را برحسب ضریب همبستگی موضع محور و قدرت همسایگی مشترک ارزیابی می‌کند. ضریب همبستگی موضع محور برخلاف تشابه کسینوسی، تفاوت در وقوع اسامی دو فرد هدف در کنار هم در یک مجموعه از اسناد مرتبط با یک گرایش را بررسی می‌کند؛ و در نتیجه قادر است رابطه‌ی میان افراد هدف مشهور را به‌درستی برآورد نماید.

الگوریتم FastModularity، گره‌ها را در گروه‌ها و با توجه به معیار مدولاریته<sup>۱</sup> ادغام می‌کند؛ در این شرایط گروه‌هایی که با چندین یال در ارتباط هستند، ادغام می‌شوند. باین‌حال، این معیار وزن‌های یال در گره‌ها را نادیده می‌گیرد. بسیاری از یال‌های متصل، وزن‌های اندکی دارند که بر انسجام گروه ادغام‌شده اثر گذاشته و عملکرد الگوریتم را با اختلال مواجه می‌سازند. روش پیشنهادی گروه‌های اولیه (ساختار کروموزوم‌ها) را بر اساس جهش‌ها و تقاطع‌ها می‌سازد و بر اساس یک تابع برازندگی انتخاب می‌نماید. روش پایه، گروه‌ها را بر مبنای امتیاز ادغام آن‌ها در هم ادغام و ترکیب می‌کند (رابطه ۶). از آنجایی که این امتیاز برحسب وزن‌های یال‌ها (قدرت دوستی) تعیین می‌شود، گره‌ها در یک گروه، ارتباطی قوی و مستحکم با یکدیگر خواهند داشت.

متعاقباً، نتیجه تشخیص جوامع در روش پیشنهادی نسبت به نتیجه حاصل از الگوریتم پایه و روش FastModularity بهتر خواهد بود. در روش SCAN از یک تشابه شبه ژاکارد برای ارزیابی قدرت همسایگی مشترک میان گره‌ها استفاده شده و یک گره با یک گروه ترکیب می‌شود اگر قدرت همسایگی مشترک آن‌ها زیاد باشد. باین‌حال روش SCAN نیز همانند روش FastModularity، وزن‌های یال‌ها را نادیده می‌انگارد و عملکرد را دچار اختلال می‌سازد. علاوه بر قدرت همسایگی مشترک، قدرت دوستی در روش پایه نیز بر وقوع هم‌زمان گره‌ها در اسناد هدف با مواضع یکسان متمرکز است. در نهایت نتیجه گرفته می‌شود که روش پایه عملکرد بهتری نسبت به روش SCAN دارد.

به‌رغم آنکه در روش CODA، وزن‌های یال‌ها در تابع هدف خوشه‌بندی آن تلفیق می‌شود، اما این وزن‌ها مبتنی بر تشابه کسینوسی بردارهای فراوانی حاصل شده‌اند. علاوه بر



محل انجام می‌گردید یعنی الگوریتم از یک نقطه شروع می‌نمود و مقداردهی انجام می‌شد (که امکان اشتباه وجود داشت).

سپس اعضای جدید را با گروه مقایسه می‌نمود و بر اساس این که عضو جدید با کدام گروه رابطه دارد به آن گروهی که مقدار  $MS_{i,k}$  بیشتر است اختصاص می‌یافت (یادآوری: رابطه ۴). (در این عبارت،  $MS_{i,k}$  نشان‌دهنده امتیاز ادغام  $p_i$  با  $C_k$  است). اساساً امتیاز ادغام یا همجوشی حاصل جمع اوزان یال مرتبط با  $p_i$  با جامعه موضع  $C_k$  است. ادغام  $p_i$  با یک گروه که امتیاز ادغام مثبت دارد، باید یک گروه با مواضع یکپارچه و منسجم به دست دهد. زمانی که بیش از یک گروه دارای امتیاز ادغام مثبت باشند، الگوریتم اقدام به ادغام  $p_i$  با آن گروهی می‌کند که حداکثر امتیاز ادغام را به خود اختصاص داده است. به تدریج که گروه‌ها تشکیل می‌شوند، در این روش ممکن است چون گروه‌ها تکمیل نگردیده‌اند، یک گره جدید (شخص جدید) به یک گروه اختصاص یابد، که درست نباشد (به‌طور منطقی باید گروه‌ها کامل تشکیل شوند تا وضعیت مشخص شود). لذا در روش پیشنهادی در این پژوهش از همان ابتدا با استفاده از روش ارائه‌شده، با ساختار الگوریتم ژنتیک و کروموزوم‌ها همه حالت‌ها بررسی می‌شوند تا به برازندگی بهتری برسد. یعنی تمامی حالت‌ها به‌صورت سراسری بررسی می‌شود. همچنین در روش پیشنهادی، چالش مقداردهی اولیه موجود روش پایه با توجه به الگوریتم پیشنهادی و ساختار الگوریتم ژنتیک مرتفع گردید. به بیان واضح‌تر، در روش پایه گروه‌های اولیه به صورت تصادفی انتخاب می‌شدند. که باعث تغییرات زیادی در تشخیص جوامع می‌گردید.

در این پژوهش از یک مجموعه داده برای ارزیابی استفاده شد. این پیکره با عنوان Kroa و شامل ۷۴ سند و اخبار مرتبط با آن است که همه‌ی آن‌ها از رسانه خبری Google News بارگذاری شده‌اند. مباحث و موضوعات جمع‌آوری شده، حوزه‌ی مسائل سیاسی را پوشش می‌دهند؛ علاوه، از برخی متخصصان انسانی نیز درخواست گردیده است تا اسناد غیر مرتبط را به‌صورت دستی از مجموعه جدا کنند تا اطمینان حاصل گردد که اسناد تحت بررسی، کاملاً با موضوعات تناسب دارند.

در طی آزمایش‌های این پژوهش، حوزه‌های پژوهشی جالب توجهی برای آثار آتی شناسایی گردید. به‌عنوان مثال علیرغم آنکه ضریب همبستگی کاربرد فراوانی در شناسایی گرایش ارتباط میان افراد هدف دارد، اما توسط مسئله‌ی فراوانی کم نام افراد تحت تأثیر قرار می‌گیرد. از آنجایی که نام افرادی که فراوانی تکرار آن‌ها کمتر است در بسیاری از متون و اخبار غایب می‌باشد، ضریب همبستگی مبتنی بر موضع احتمالاً استحکام رابطه را بیش از حد واقعی برآورد می‌کند. کاهش وزن اسناد و متون در زمانی که نام افراد با فراوانی اندک از متون غایب هستند، می‌تواند مسئله‌ی برآورد فراتر از انتظار را حل نماید. علاوه بر این، لحاظ کردن متون و اخبار غیر مرتبط با مبحث موردنظر می‌تواند موجب وجود نام افراد غیر مرتبط در فرآیند شناسایی شود و عملکرد سیستم را با مشکل مواجه سازد. بنابراین، رویکردهای مبتنی بر حذف مباحث غیر مرتبط باید در جهت توسعه‌ی عملکرد شناسایی مورد استفاده قرار بگیرند. همچنین، اسناد و اخبار ورودی به‌صورت دستی جمع‌آوری شده بودند. برای آنکه به کاربران اینترنتی کمک شود مباحث و موضوعات نوظهور را به‌خوبی درک کنند، روش پایه می‌تواند با فن‌های مختلف شناسایی و مسیریابی موضوع تلفیق شود؛ بدین صورت اسناد و اخبار پیرامون یک مبحث به‌طور خودکار و از منابع اطلاعاتی مختلف (مثلاً آژانس‌های خبری)، شناسایی و رهگیری می‌شوند. در همین رابطه با توجه به روش بهینه ارائه شده با الهام از ساختار الگوریتم ژنتیک، در این پژوهش مشکل مقداردهی اولیه و نیز محلی بودن روش پایه حل گردید و باعث افزایش دقت خوشه‌بندی نیز گردید. لذا به دلیل ماهیت تصادفی بودن الگوریتم ژنتیک می‌توان این استنباط را داشت که بهبود در روش پایه قبل از بهره‌گیری از الگوریتم ژنتیک نیز می‌تواند مفید واقع گردد.

در روش پیشنهادی و در مقایسه با روش پایه، نتیجه‌های ذیل به دست آمد. اول در روش پیشنهادی دقت خوشه‌بندی نسبت به روش پایه و سایر روش‌های تشخیص جوامع، افزایش یافت. همچنین در روش پیشنهادی مسئله (چالش) حساس بودن به مقداردهی اولیه مرتفع گردید و در نهایت مسئله خوشه‌بندی محلی در روش پایه نیز حل شده به‌طوری که با روش ارائه شده، خوشه‌بندی (تشخیص جوامع) به‌صورت سراسری صورت گرفت. لازم به توضیح است که در روش پایه عمل خوشه‌بندی به‌صورت

Media:Architecture, Tools, and Approaches to Detect Criminal Activity,” در *Application of Big Data for Application of Big Data for National Security*, Elsevier, 2015, pp. 155-172.

11.G.-J. Qi , C. Aggarwal و T. Huang, “Community detection with edge content in social media networks,” در *Paper presented at the 2012 IEEE 28th international conference on data engineering* .۲۰۱۲ ,

12.S. Borgatti و M. Everett , “Graph colorings and power in experimental exchange networks,” *Soc Netw*, شماره ۱۴, جلد ۳, pp. 287–308, 1992.

13.S. Nijssen و J. Kok , “A quickstart in frequent structure mining can make a difference,” در *Paper presented at the Proceedings of the tenthACMSIGKDD international conference on knowledge discovery and data mining* .۲۰۰۴ ,

14.C. Troussas , M. Virvou, J. Caro و K. Espinosa , “Mining relationships among user clusters in Facebook for language learning,” در *Paper presented at the international conference on computer, information and telecommunication systems (CITS)* .۲۰۱۳ ,

15.C. Chen, Z.-Y. Chen و C.-Y. Wu , “An unsupervised approach for person name bipolarization using principal component analysis,” *IEEE Trans. Knowl. Data Eng.*, جلد ۲۴, pp. 1963-1976, 2012.

16.C. Chen و C.-Y. Wu , “Bipolar person name identification of topic documents using principal component analysis,” در *Proceeding of the 23rd International Conference on Computational Linguistics* , .۲۰۱۰

17.C. D. Manning, P. Raghavan and H. Schütze, "An Introduction to Information Retrieval," in *Cambridge University Press*, New York, 2009.

## منابع

1.A. M. Kaplan و M. Haenlein, “Users of the world, unite! The challenges and opportunities of social media,” *Business Horizons*, شماره ۱, جلد ۵۳, pp. 59-68, 2010.

۲. ادرون خطی. ۲۰۱۴

3.A. Mislove , “Online social networks: measurement, analysis, and applications to distributed information systems,” 2009.

4.M. Sachan , D. Contractor, T. Faruque و L. Subramaniam , “Using content and interactions for discovering communities in social networks,” در *Paper presented at the proceedings of the 21st international conference on world wide web* .۲۰۱۲ ,

5.D. Ganley و C. Lampe, “The ties that bind: social network principles in online communities,” *Decision Support Systems*, جلد ۴۷, شماره ۳, pp. 266-274, 2009.

6.M. Kuramochi و G. Karypis , “Finding frequent patterns in a large sparse graph,” *Data Min Knowl Discov* , شماره ۳, جلد ۱۱, p. 243–271.

7.X. Yan و J. Han, “gspan: graph-based substructure pattern mining,” در *Paper presented at the Proceedings of the IEEE international conference on data mining (ICDM 2002)* .۲۰۰۲ ,

8.H. Cai, V. W. Zheng و K. C.-C. Chang, “A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications,” *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, شماره ۹, جلد ۳۰, .۲۰۱۷

9.Z.-Y. Chen و C. C. Chen, “SCIFNET: Stance community identification of topic persons using friendship network analysis,” *Knowledge-Based Systems*, جلد ۱۱۰, p. 30–48, 2016.

10.B. Akhgar, G. B. Saathoff, H. R. Arabnia, R. Hill, A. Staniforth و P. Saskia Bayerl, “Mining Social کندگان,”



- 26.I. Messaoudi و N. Kamel, "A multi-objective bat algorithm for community detection on dynamic social networks," *Applied Intelligence*, pp. 1-18, 1 2019.
- 27.W. Luo, N. Lu, L. Ni, W. Zhu و W. Ding, "Local community detection by the nearest nodes with greater centrality," *Information Sciences*, جلد ۵۱۷، pp. ۳۹۲-۳۷۷، ۲۰۲۰.
- 28.W. Luo, . D. Zhang , . H. Jiang , L. Ni . و Y. Hu , "Local Community Detection With the Dynamic Membership Function," *IEEE Transactions on Fuzzy Systems*, جلد ۲۶، شماره ۵، ۲۰۱۸، ۵.
- 29.X. Wan, X. Zuo و F. Song, "Solving dynamic overlapping community detection problem by a multiobjective evolutionary algorithm based on decomposition," *Swarm and Evolutionary Computation*, ۲۰۲۰.
- 30.Y. Zhang, Y. Liu, J. Li, J. Zhu, C. Yang, W. Yang و C. Wenc, "WOCDA: A whale optimization based," *Physica A: Statistical Mechanics and its Applications*, جلد ۵۳۹، ۲۰۲۰.
- 31.X. Pan, . G. Xu , . B. Wang . و T. Zhang , "A Novel Community Detection Algorithm Based on Local Similarity of Clustering Coefficient in Social Networks," *IEEE Access* , جلد ۷، ۲۰۱۹، ۷.
- 32.Z. Harris, "Distributional structure," *Word* , جلد ۱۰، p. 146-162, 1954.
- 33.H. Kanayama و T. Nasukawa, "Fully automatic lexicon expansion for domain-oriented sentiment analysis," در *Proceedings of the Conference on Empirical Methods in Natural Language Processing* ۲۰۰۶،
- 34.G. Keller , "Statistics for Management and Economics," *Cengage Learning*, ۲۰۰۸
- 35.P. Turney و M. Littman , "Measuring praise and criticism: inference of semantic orientation from association," *ACM Trans.*
- 18.S. B. Yudhoatmojo و M. A. Samuar, "Community Detection On Citation Network Of DBLP Data Sample Set Using LinkRank Algorithm," *۴<sup>th</sup> Information Systems International Conference 2017, ISICO 2017* ,Bali, Indonesia, 2017.
- 19.B. Viswanath, M. Bashir , M. Crovella , S. Guha, K. Gummadi, B. Krishnamurthy و A. Mislove, "Towards detecting anomalous user behavior in online social networks.," در *Proceedings of the 23rd USENIX security symposium (USENIX Security ۲۰۱۴)* ,
- 20.M. Girvan و M. Newman, "Community structure in social and biological networks," *Proc Natl Acad Sci* , جلد ۹۹(۱۲)، p. 7821-6, 2002.
- 21.J. Gao, F. Liang , W. Fan, C. Wang , Y. Sun و J. Han, "On community outliers and their efficient detection in information networks," در *Proceedings of the 16th ACM SIGKDD international conference Proceedings of the 16th ACM SIGKDD international conference ۲۰۱۰* ,
- 22.X. Chen and J. Li, "Community detection in complex networks using edge-deleting with restrictions," *Physica A: Statistical Mechanics and its Applications*, vol. 519, pp. 181-194, 4 2019.
- 23.M. Rezvani, W. Liang, . C. Liu . و J. Xu Yu , "Efficient Detection of Overlapping Communities Using Asymmetric Triangle Cuts," *IEEE Transactions on Knowledge and Data Engineering*, شماره ۱۱، جلد ۳۰، ۲۰۱۸.
- 24.Z. Liu, B. Xiang, W. Guo , Y. Chen, K. Guo . و J. Zheng , "Overlapping Community Detection Algorithm Based on Coarsening and Local Overlapping Modularity," *IEEE Access*, جلد ۷، ۲۰۱۹، ۷.
25. T. Meng , L. Cai , . T. He , L. Chen . و Z. Deng, "Local Higher-Order Community Detection Based on Fuzzy Membership Functions," *IEEE Access*, جلد ۷، ۲۰۱۹، ۷.

38.B. Yang, W. Cheung و J. Liu, "Community mining from signed social networks," *IEEE Trans. Knowl. Data Eng.*, جلد ۱۹, p. 1333–1348, 2007.

39.P. Anchuri و M. Magdon-Ismai, "Communities and balance in signed networks: a spectral approach," در *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, ۲۰۱۲ ,

40.T. Mitchel, *Machine Learning*, McGrawHill, 1997.

*Inf. Syst.*, جلد ۲۱, p. 315–346, 2003.

36.M. Newman, "Fast algorithm for detecting community structure in networks," *Phys. Rev.*, جلد ۷۴, ۲۰۰۴ ,

37.X. Xu, N. Yuruk, Z. Feng و T. Schweiger, "SCAN: a structural clustering algorithm for networks," در *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ۲۰۰۷ ,

