
بسم الله الرحمن الرحيم

دانشگاه شهید بهشتی

دانشکده مهندسی برق و کامپیوتر

ارائه یک روش جدید برای بهبود الگوریتم‌های خوشه‌بندی اجتماعات وب با استفاده از وب‌کاوی

پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر

گرایش نرم‌افزار

رسول حسین‌زاده

استاد راهنما:

جناب آقای دکتر اسلام ناظمی

استاد مشاور:

جناب آقای مهندس حسین علیزاده

زمستان ۱۳۹۱

دانشگاه شهید بهشتی

دانشکده مهندسی برق و کامپیوتر

پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر

تحت عنوان:

ارائه یک روش جدید برای بهبود الگوریتم‌های خوشه‌بندی اجتماعات وب با استفاده از وب‌کاوی

در تاریخ پایان‌نامه دانشجو رسول حسین‌زاده، توسط کمیته تخصصی داوران مورد بررسی و تصویب نهایی قرار گرفت.

۱- استاد راهنما اول:	نام و نام خانوادگی: دکتر اسلام ناظمی	امضاء
۲- استاد مشاور:	نام و نام خانوادگی: مهندس حسین علیزاده	امضاء
۲- استاد داور (داخلی):	نام و نام خانوادگی: دکتر محسن ابراهیمی مقدم	امضاء
۳- استاد داور (خارجی):	نام و نام خانوادگی: دکتر بهروز مینایی	امضاء
۴- نماینده تحصیلات تکمیلی:	نام و نام خانوادگی: دکتر فرح ترکمنی آذر	امضاء

بـا تشکر از همه اساتید و دوستانی کـه مرا

در به انـجام رسانیدن این پایان‌نامه یـــاری نمودند.

بـه ویژه اسـاتیـد گـرامی دکتر ناظـمی و مهندس علـیـزاده

کـــه هـمـواره راهـنـمـایی آن‌ها راه‌گشـــای کـــار ایـنـجانب بود.

کلیه حقوق مادی مترتب بر نتایج مطالعات،

ابتکارات و نوآوری‌های ناشی از تحقیق موضوع

این پایان نامه متعلق به دانشگاه شهید بهشتی

می‌باشد.

به نام خدا

نام و نام خانوادگی: رسول حسین‌زاده

عنوان پایان‌نامه: ارائه یک روش جدید برای بهبود الگوریتم‌های خوشه‌بندی اجتماعات وب با استفاده از وب‌کاوی

استاد راهنما: جناب آقای دکتر اسلام ناظمی

استاد مشاور: جناب آقای مهندس حسین علیزاده

تقدیم به روح پاک پدرم که نیکنامی را به من آموخت.

و

تقدیم به مادر عزیزم که از سپاسگزاری زحمات بی‌دریغش قاصرم.

و

تقدیم به خواهــر عزیزم با آرزوی موفقیــت‌های هر چه بیشــتر...

فهرست مطالب

فهرست اشکال

فهرست جداول

چکیده

امروزه شبکه‌های اجتماعی دارای کاربردهای مختلف هستند و به خصوص جایگاه مهمی در بین کاربران اینترنت دارند. به همین دلیل تحلیل شبکه‌های اجتماعی حوزه پژوهشی مهم و تأثیرگذار در بین پژوهشگران است. در سال‌های اخیر، برای بهره‌برداری از حجم وسیع داده‌های وب روش‌های وب‌کاوی معرفی شده‌اند. وب‌کاوی، به کارگیری روش‌های داده‌کاوی برای کشف و استخراج خودکار اطلاعات از اسناد و سرویس‌های وب می‌باشد. از انواع روش‌های داده‌کاوی وب، داده‌کاوی ساختار وب است که با استفاده از پیوندها، اطلاعات جدیدی راجع به صفحات به دست می‌آورد. یکی از چالش‌های مهم در تحلیل شبکه‌های اجتماعی، تشخیص اجتماعات است. اجتماع مجموعه افراد یا سازمان‌هایی هستند که چگالی ارتباط آن‌ها با هم بیشتر از سایر موجودیت‌های شبکه است.

یکی از مهم‌ترین مشکلات در شبکه‌های پیچیده یا شبکه‌های اجتماعی تشخیص اجتماعات می‌باشد. خوشه‌بندی یا تشخیص اجتماعات، ساختار اجتماعات در شبکه‌های اجتماعی، ارتباطات پنهان بین مؤلفه‌های آن را آشکار خواهد نمود. با در نظر گرفتن افزایش پایگاه داده‌های مربوط به شبکه‌های اجتماعی، به الگوریتم‌های با دقت بالا و مقیاس-پذیری برای تجزیه تحلیل آن‌ها نیاز است.

اکثر روش‌های رایج تشخیص اجتماعات موجود قطعی نیستند، و نتایج آن‌ها به مقادیر اولیه‌ای که در اکثر مواقع به صورت تصادفی انتخاب می‌شود بستگی دارد. خوشه‌بندی ترکیبی در تحلیل داده‌ها برای رسیدن به نتایج پایدار بدون توجه به مقادیر اولیه تصادفی می‌باشد. در این پایان‌نامه یک روش برای پیدا نمودن اجتماعات صحیح و دقیق با استفاده از خوشه‌بندی ترکیبی با نام تشخیص اجتماعات ترکیبی ارائه خواهد شد و نشان خواهیم داد که خوشه‌بندی ترکیبی توانایی ترکیب با هر روش دیگری را خواهد داشت به گونه‌ای که دقت تقسیم‌بندی اجتماعات را افزایش می‌دهد. از نتایج این بررسی‌ها می‌تواند در مسائل بسیاری از جمله: بهبود موتورهای جستجو، درک ساختار شبکه، تشخیص دقیق-تر اجتماعات، بازاریابی، تبلیغات و... مورد استفاده قرار گیرد.

کلمات کلیدی: *خوشه‌بندی، تشخیص اجتماعات ترکیبی، خوشه‌بندی ترکیبی، تشخیص اجتماعات، پودمانی، شبکه-های اجتماعی، وب‌کاوی*

فصل اول: کلیات تحقیق

وب، محیطی وسیع، متنوع و پویا است که کاربران متعدد اسناد خود را در آن منتشر می‌کنند. وب طی یک فرآیند آشفته و غیرمتمرکز رشد می‌کند و این روند منجر به تولید حجم وسیعی از مستندات متصل به یکدیگر گشته است که از هیچ‌گونه سازماندهی منطقی برخوردار نیستند.

با توجه به حجم وسیع اطلاعات در وب، مدیریت آن با ابزارهای سنتی تقریباً غیرممکن است و ابزارها و روش‌هایی نو برای مدیریت آن مورد نیاز است. به طور کلی کاربران وب در استفاده از آن با مشکلات زیر روبرو هستند[۱]:

- **یافتن اطلاعات مرتبط:** یافتن اطلاعات مورد نیاز در وب دشوار است. روش‌های سنتی بازیابی اطلاعات که برای جستجوی اطلاعات در پایگاه داده‌ها به کار می‌روند، قابل استفاده در وب نمی‌باشند و کاربران معمولاً از موتورهای جستجو که مهم‌ترین و رایج‌ترین ابزار برای یافتن اطلاعات در وب هستند، استفاده می‌کنند. این موتورها، یک پرس‌وجوی[۱] مبتنی بر کلمات کلیدی از کاربر دریافت کرده و در پاسخ لیستی از اسناد مرتبط با پرس‌وجوی وی را که بر اساس میزان ارتباط با این پرس‌وجو مرتب شده‌اند، به وی ارائه می‌کنند. اما موتورهای جستجو دارای دو مشکل اصلی هستند. اولاً دقت[۲] موتورهای جستجو پایین است، چرا که این موتورها در پاسخ به یک پرس‌وجوی کاربر صدها یا هزاران سند را بازیابی می‌کنند، در حالی که بسیاری از اسناد بازیابی شده توسط آن‌ها با نیاز اطلاعاتی کاربر مرتبط نیستند. ثانیاً میزان فراخوان[۳] این موتورها کم می‌باشد، به آن معنی که قادر به بازیابی کلیه اسناد مرتبط با نیاز اطلاعاتی کاربر نیستند. چرا که حجم اسناد در وب بسیار زیاد است و موتورهای جستجو قادر به نگهداری اطلاعات کلیه اسناد وب، در پایگاه داده‌های خود نمی‌باشند. برای رفع این مشکلات در سال‌های اخیر از موتورهای جستجوی هوشمند، وب معنایی، روش‌های مختلف فشرده سازی و غیره استفاده شده است.
- **ایجاد دانش جدید با استفاده از اطلاعات موجود در وب:** این مشکل در واقع بخشی از مشکل مطرح شده در قسمت قبل است. در حال حاضر این سوال مطرح است که چگونه می‌توان داده‌های فراوان موجود در وب را به دانشی قابل استفاده تبدیل نمود، به طوری که یافتن اطلاعات مورد نیاز در آن به سادگی صورت بگیرد. همچنین چگونه می‌توان با استفاده از داده‌های وب به اطلاعات و دانشی جدید دست یافت.

---

[1] Query

[2] Precision

[3] Recall

- **خصوصی سازی[1] اطلاعات:** از آن جا که کاربران متفاوت هر یک درباره نوع و نحوه بازنمایی اطلاعات سلیقه خاصی دارند، این مسئله باید مورد توجه تأمین کنندگان اطلاعات در وب قرار بگیرد. برای این منظور با توجه به خواسته‌ها و تمایلات کاربران متفاوت، نحوه ارائه اطلاعات به آن‌ها باید سفارشی گردد.

برای بهره‌برداری از حجم وسیع داده و کاهش مشکلات فوق‌الذکر، در سال‌های اخیر روش‌های وب‌کاوی[2] معرفی شده‌اند. وب‌کاوی به کارگیری روش‌های داده‌کاوی[3] برای کشف و استخراج خودکار اطلاعات از اسناد و سرویس‌های وب می‌باشد. البته روش‌های وب‌کاوی تنها ابزار موجود برای حل این مشکلات نیستند. بلکه روش‌های مختلفی از سایر زمینه‌های تحقیقاتی همچون پایگاه داده‌ها، بازیابی اطلاعات، پردازش زبان طبیعی و ... قابل استفاده در این زمینه می-باشند. همچنین وب‌کاوی با زمینه‌های مختلف تحقیقاتی علوم کامپیوتر همچون داده‌کاوی، پایگاه داده، بازیابی اطلاعات، هوش مصنوعی، یادگیری ماشین، پردازش زبان طبیعی، استخراج اطلاعات، انبار داده‌ها[4]، تشخیص اجتماعات(خوشه‌بندی) و ... در ارتباط تنگاتنگ است.

امروزه با توجه به گسترش روزافزون اینترنت، خدمات وب و شبکه‌های اجتماعی مجازی نقش مهم آن‌ها در زندگی واقعی افراد بیشتر از همیشه احساس می‌گردد. در واقع شبکه‌های اجتماعی، شبکه‌های تعاملی هستند که از اینترنت به عنوان رسانه‌ای برای ایجاد ارتباط بین افراد استفاده می‌کنند. وبلاگ‌ها، پست‌های الکترونیکی و سایت‌های دوست‌یابی می‌توانند نمونه‌ای از شبکه‌های اجتماعی تلقی شوند.

شبکه‌های اجتماعی ساختاری اجتماعی است که افراد یا سازمان‌هایی تشکیل شده است که گره‌های شبکه را تشکیل می‌دهند. گره‌ها توسط یک یا چند نوع خاص از وابستگی به هم متصل هستند، برای مثال تبادلات مالی، دوستی‌ها، خویشاوندی، تجارت، پیوندهای وب، سرایت بیماری‌ها(اپیدمولوژی) یا مسیرهای هواپیمایی نمونه‌هایی از ارتباط هستند. اما ساختارهای حاصل اغلب بسیار پیچیده هستند. برای نمایش و تحلیل شبکه‌های اجتماعی معمولاً از تئوری گراف استفاده می‌شود.

---

[1] Personalization

[2] Web Mining

[3] Data Mining

[4] Data Warehouse

مؤلفه‌های موجود در تئوری گراف[۱] گره[۲] لبه[۳] است.گره‌ها در شبکه‌های اجتماعی، بازیگران فردی درون شبکه‌ها هستند و لبه‌ها نقش روابط میان این بازیگران را ایفا می‌کنند. تحلیل شبکه اجتماعی[۴] عبارت است از نگاشت و اندازه-گیری روابط و همکاری‌ها در بین افراد، گروه‌ها، سازمان‌ها، و هر موجودیتی که قابلیت پردازش اطلاعات و دانش داشته باشد. با افزایش سریع کاربران این شبکه‌ها، کاوش در مقیاس بالای داده‌ها می‌تواند کارایی بهتر و مؤثرتری از پتانسیل پنهان این شبکه‌ها فراهم کند.

بخش عمده‌ی فعالیت‌ها و تحقیقات انجام شده در وب‌کاوی به محتوای صفحات وب می‌پردازند. اما در سال‌های اخیر داده‌کاوی ساختار وب و داده‌کاوی استفاده از وب نیز مورد توجه قرار گرفته‌اند.

محور اصلی این پایان‌نامه داده‌کاوی ساختار وب است. همان طور که گفته شد، الگوریتم‌های داده کاوی ساختار وب با استفاده از پیوندها اطلاعات جدیدی راجع به صفحات به دست می‌آورند. در این نوع از وب‌کاوی، وب به صورت یک گراف مدل‌سازی می‌شود که در آن صفحات وب، گره‌های گراف و پیوندهای بین صفحات، یال‌های گراف هستند. الگوریتم‌های داده‌کاوی ساختار وب در کاربردهای متفاوتی همچون رتبه‌بندی[۵] صفحات وب، تشخیص اجتماعات وب[۶]، پیمایش صفحات وب[۷]، تحلیل گراف وب، مدل‌سازی و شبیه سازی فرآیند تولید گراف وب به کار می‌روند.

در شبکه‌های اجتماعی برخی گره‌ها (افراد یا سازمان‌ها) در مقایسه با کل گره‌های شبکه، ارتباط بیشتری با هم دارند که به آن‌ها اجتماع(گروه)[۸] گفته می‌شود. یکی از چالش‌های مهم در این تحلیل شبکه‌های اجتماعی، تشخیص اجتماعات می‌باشد.

یک راهکار مناسب برای بهبود نتایج روش‌های داده‌کاوی ساختار وب و در نهایت بهبود نتایج تشخیص اجتماعات که در این پایان‌نامه مورد بررسی قرار می‌گیرد، به کارگیری خوشه‌بندی ترکیبی است. اکثر روش‌هایی تشخیص اجتماعات روی جنبه‌های خاصی از داده‌های ورودی تاکید می‌کنند، در نتیجه روی مجموعه داده‌های خاصی کارآمد

---

[1] Graph theory
[2] Node
[3] Edge
[4] Social Network Analysis(SNA)
[5] Ranking
[6] Web Community Detection
[7] Crawling
[8] Community

هستند. بنابراین به روش‌هایی نیاز داریم که بر روی تمامی داده‌ها نتایج خوبی را به همراه داشته باشد. در این پایان نامه سعی داریم با استفاده از خوشه‌بندی ترکیبی و با ترکیب صحیح نتایج حاصل از الگوریتم‌های اولیه تشخیص اجتماعات، نتایج دقیق‌تر، کاراتر، پایدارتر، مستحکم‌تر، مطمئن‌تری را داشته باشیم.

۱-۱ انگیزه

با توجه به گسترش روزافزون حجم اطلاعات در وب و ارتباط وب‌کاوی با تجارت الکترونیکی، وب‌کاوی به یک زمینه تحقیقاتی وسیع مبدل گشته است. تکنیک‌ها و روش‌های وب‌کاوی از کاربرد وسیعی در حوزه‌های مختلف همچون موتورهای جستجو، تجارت الکترونیکی، دولت الکترونیکی، آموزش الکترونیکی، آموزش از راه دور، سازمان‌های مجازی، مدیریت دانش، کتابخانه‌های دیجیتال، ... برخوردارند.

روش‌های وب‌کاوی بر اساس آن که چه نوع داده‌ای را مورد کاوش قرار می‌دهند، به سه دسته داده‌کاوی محتوای وب[۱]، داده‌کاوی ساختار وب[۲] و داده‌کاوی استفاده از وب[۳] تقسیم می‌شوند. داده‌کاوی محتوای وب، فرآیند استخراج اطلاعات مفید از محتوای مستندات وب است. داده‌کاوی ساختار وب به کشف اطلاعات جدید با استفاده از پیوندهای[۴] بین صفحات وب می‌پردازد. داده‌کاوی استفاده از وب نیز داده‌های مربوط به استفاده کاربران از وب را مورد کاوش قرار می‌دهد و الگوهای استفاده از وب را به منظور درک و برآوردن بهتر نیازهای کاربران استخراج می‌کند.

هدف از بررسی پیوندها در وب و در نهایت تشخیص اجتماعات، جدا کردن گره‌هایی است که ارتباط بیشتری با هم دارند. اجتماع مجموعه‌ای از گره‌ها است که راجع به یک موضوع مشترک هستند و معمولاً توسط افراد یا سازمان‌های مختلف که علایق مشترک درباره یک موضوع خاص دارند، ایجاد می‌شوند. همچنین تعداد اتصالات صفحات یک اجتماع وب با یکدیگر بیش از تعداد اتصالاتشان با سایر صفحات وب است. در واقع تشخیص اجتماعات، تقسیم‌بندی‌های موجود

---

[1] Web Content Mining

[2] Web Structure Mining

[3] Web Usage Mining

[4] Hyperlink

در شبکه را نشان می‌دهد و گروه‌های یک گراف را از هم مجزا می‌کند. تشخیص اجتماعات به ما کمک می‌کند تا دید بهتری نسبت به ساختار شبکه پیدا کنیم. به طور کلی برخی از کاربردهای تشخیص اجتماعات عبارتند از[۲]:

- **درک ساختار شبکه:** در شبکه‌های پیچیده‌ای مانند وب و یا شبکه‌های اجتماعی که با حجم بالایی از گره‌ها روبرو هستیم، تشخیص اجتماعات اولین و مهم‌ترین قدم در راستای شناخت و تحلیل این شبکه‌ها محسوب می‌شود.
- **یافتن اجتماعاتی با ویژگی خاص:** گاهی در تشخیص اجتماعات به دنبال اجتماعاتی هستیم که ویژگی خاصی دارند. مثلاً اجتماع افرادی که در یک تیم خاص عضو هستند، یا اجتماع افراد تروریست، یا به طور کلی گروه افرادی که به محصول خاصی علاقه‌مندند هستند. یافتن اجتماعات خاص می‌تواند اهداف متفاوتی داشته باشد، که از آن جمله می‌توان به اهداف تجاری، بازاریابی، اهداف سیاسی و اجتماعی اشاره نمود.
- **پیمایش موضوعی:** پیمایشگرهای موضوعی کار خود را با یک مجموعه اولیه از صفحات آغاز می‌کنند و با استفاده از پیوندهای موجود در این صفحات، صفحات دیگر را پیمایش کرده و این روند را تا رسیدن تعداد صفحات پیمایش شده به حدی مشخص انجام می‌دهند. پیمایشگرهای موضوعی در روند پیمایش خود به صورت انتخابگر عمل می‌کنند و صفحاتی را برای پیمایش انتخاب می‌کنند که تا حد ممکن در ارتباط با موضوعی خاص باشند. بنابراین با پیدا نمودن یک اجتماع می‌توان اطلاعاتی در مورد موضوع اجتماع پیدا نمود.
- **قابل مشاهده کردن گراف:** با تشخیص اجتماعات می‌توانیم نمای کامل گراف یا بخشی از آن را مشاهده کنیم. با مشاهده اجتماعات موجود در یک گراف می‌توان اجزای آن را به طور مجزا از کل گراف مورد تحلیل قرار داد.
- **بهبود موتورهای جستجو:** موتورهای جستجو برای بازیابی اسناد مرتبط با پرس‌وجوی کاربران، اسناد را رتبه‌بندی می‌نمایند. روش‌های رتبه‌بندی میزان ارتباط پرس‌وجوی کاربر را با هر صفحه از یک مجموعه صفحات محاسبه می‌کنند و به هر صفحه رتبه‌ای اختصاص می‌دهند. برای این منظور پرس‌وجو و اسناد به صورتی که قابل مقایسه با یکدیگر باشند، بازنمایی می‌شوند. از تشخیص اجتماعات می‌توان در خوشه‌بندی موضوعی وب سایت‌ها نیز استفاده کرد. این روش به بهبود عملکرد موتورهای جستجو کمک شایان توجهی می‌کند.

تشخیص اجتماعات و خوشه‌بندی شباهت‌های بسیار زیادی با هم دارند و منابع زیادی این دو را یکسان می-دانند[۲][۳][۴][۵]. خوشه‌بندی[1] داده‌ها یکی از مراحل اصلی در داده‌کاوی است که وظیفه کاوش الگوهای پنهان در داده‌های بدون برچسب را بر عهده دارد. به خاطر پیچیدگی مسئله و ضعف روش‌های خوشه‌بندی پایه، امروزه اکثر مطالعات به سمت روش‌های خوشه‌بندی ترکیبی[2] هدایت شده است. از آنجایی که اکثر روش‌های خوشه‌بندی پایه روی جنبه‌های خاصی از داده‌ها تاکید می‌کنند، در نتیجه روی مجموعه داده‌های خاصی کارآمد می‌باشند. به همین دلیل، نیازمند روش‌هایی هستیم که بتواند با استفاده از ترکیب الگوریتم‌های تشخیص اجتماعات و گرفتن نقاط قوت هر یک، نتایج بهینه‌تری را تولید کند. در واقع هدف اصلی خوشه‌بندی ترکیبی جستجوی نتایج بهتر و مستحکم‌تر، با استفاده از ترکیب اطلاعات و نتایج حاصل از چندین خوشه‌بندی اولیه است.

در این پایان‌نامه یک رویکرد جدید برای بهبود روش‌های تشخیص اجتماعات با استفاده از خوشه‌بندی ترکیبی با نام تشخیص اجتماعات ترکیبی ارائه شده است. روش‌های تشخیص اجتماعات نیز همانند روش‌های خوشه‌بندی روی جنبه‌های خاصی از داده‌ها تاکید می‌کنند، بنابراین روی مجموعه داده‌های خاصی کارآمد هستند. بنابراین نیازمند به روش‌هایی هستیم که در همه حالات بتواند جواب نسبتاً بهینه را داشته باشد. به خاطر پیچیدگی مسئله تشخیص اجتماعات و ضعف روش‌های خوشه‌بندی پایه، امروزه اکثر مطالعات به سمت روش‌های خوشه بندی ترکیبی هدایت شده است. با استفاده از روش‌های خوشه‌بندی ترکیبی و ترکیب آن با روش‌های تشخیص اجتماعات می‌توان تا حدود زیادی این مشکل را مرتفع نمود.

۱-۲ چالش‌ها و مشکلات موجود

مسئله اصلی در تشخیص اجتماعات این است که بدانیم چگونه به بهترین حالت شبکه را به اجتماعات اصلی آن تقسیم کنیم. در شبکه‌های واقعی هیچ اطلاعاتی درباره تعداد اجتماعات وجود ندارد. این چالش، تشخیص اجتماعات را دچار مشکل می‌کند. در بعضی از روش‌های تشخیص اجتماعات فرض بر آن است که تعداد اجتماعات شبکه را از قبل

---

[1] Clustering

[2] Ensemble Clustering

می‌دانیم. در حالیکه در بسیاری از شبکه‌ها، هیچ دانش اولیه‌ای در مورد اجتماعات شبکه وجود ندارد. روش‌های جدید به دنبال بر طرف کردن این نقیصه هستند، در این پایان‌نامه نیز یک روش ترکیبی برای پیدا نمودن تعداد اجتماعات از روی گراف حاصله از شبکه مورد تحلیل، ارائه می‌شود. مسئله دیگر در تشخیص اجتماعات به خصوص در وب و یا شبکه‌های اجتماعی، مقیاس بالای گره‌های گراف است که چالش بزرگی در این زمینه به حساب می‌آید. روش‌های بسیاری در این زمینه وجود دارد که دارای کاربردهای مختلفی است. روش پیشنهادی مستقل از نوع روش استفاده شده برای تشخیص اجتماعات است، این روش تنها به نتایج نهایی هر الگوریتم نگاه می‌کند و با کنار هم قرار دادن نتایج متفاوت سعی در ارائه یک نتیجه دقیق‌تر و مستحکم‌تر برای تشخیص اجتماعات می‌باشد.

از مزایای مهم خوشه‌بندی ترکیبی در تشخیص اجتماعات می‌توان به استفاده از روش‌های قبلی به تعداد زیاد که نشان از مقیاس‌پذیری بالای این روش و نیز دقت بالای آن به نسبت روش‌های دیگر و در مجموع نتایج دقیق‌تر و پایدارتری را خواهد داشت. مزایای مهم دیگر در این روش این است که، نیازی به دانستن تعداد اجتماعات نیز وجود ندارد. در بسیاری از روش‌های تشخیص اجتماع، نیازمند دانستن تعداد اجتماعات هستیم. با توجه به اینکه تعداد اجتماعات در وب برخلاف روش‌های خوشه‌بندی معین نیست این روش می‌تواند کمک شایانی را در این خصوص داشته باشد. از معایب این روش به هزینه اجرایی آن که با بدترین حالت هزینه اجرایی در الگوریتم‌های استفاده شده برابر است و نیز حافظه زیادی را نیز مصرف می‌نماید، همچنین ممکن است در همه موارد جواب صددرصد بهینه را نداشته باشیم و الگوریتم‌هایی وجود داشته باشند که برای داده‌های خاصی جواب دقیق‌تری را محاسبه نمایند. به طور میانگین در تمامی روش‌ها نتیجه نهایی از میانگین کلی نتایج اولیه بهتر بوده است و در بعضی موارد از تمامی نتایج نیز بهتر مشاهده شده است.

۱-۳ محدوده‌ی این تحقیق

اگرچه وب‌کاوی با چالش‌ها و محدودیت‌های متنوعی رو به رو است که از آن جمله می‌توان به داده‌های ناصحیح و نادقیق، حجم وسیع و روبه گسترش داده‌های وب، تغییر داده‌های وب در فواصل زمانی کوتاه، کاربران گوناگون با نیازهای مختلف، عدم وجود ابزارها و الگوریتم‌های مناسب اشاره کرد. در میان کاربردهای داده‌کاوی ساختار وب تشخیص اجتماعات از این جهت که می‌تواند به کاربران در بازیابی اطلاعات از وب کمک کند، اهمیت ویژه‌ای دارد.

علاوه بر اجتماعاتی که صریحاً در وب تعریف شده‌اند(مانند گروه‌های خبری)، اجتماعات دیگری نیز به طور ضمنی در وب وجود دارند که حتی اعضای آن ممکن است از وجود آن بی‌اطلاع باشند. تحقیقات اخیر نشان می‌دهند که تعداد زیادی از اجتماعات وب در وب وجود دارد. با تشخیص یک اجتماع وب درباره یک موضوع خاص، کاربران می‌توانند با استفاده از صفحات اجتماع، اطلاعات مفیدی درباره آن موضوع به دست آورند. اجتماعات وب تا حدودی با اجتماعات واقعی متفاوت هستند. به عنوان مثال اجتماعات وب می‌توانند شامل رقبا یا نویسندگانی باشند که یکدیگر را نمی‌شناسند. از آنجا که امروزه حجم وب هر روزه در حال افزایش است، تشخیص اجتماعات وب و رابطه بین آن‌ها روز به روز دشوارتر می‌شود.

در این تحقیق به بهبود روش‌های تشخیص اجتماعات پرداخته شده است. در همین راستا در این پایان‌نامه ابتدا به بررسی الگوریتم‌های موجود خواهیم پرداخت. سپس با استفاده از خوشه‌بندی ترکیبی سعی در برطرف نمودن مشکل دقت و پایداری نتایج تشخیص اجتماعات می‌پردازیم. به عبارت دیگر در این پایان‌نامه قصد داریم با استفاده از روش-های خوشه‌بندی ترکیبی روشی با نام تشخیص اجتماعات ترکیبی را برای بهبود روش‌های تشخیص اجتماعات معرفی نماییم.

۱-۴ خلاصه‌ی این فصل

تشخیص اجتماعات در شبکه‌های اجتماعی نقش مهمی در پیدا نمودن اجتماعات مختلف دارد. همان‌طوری که در بخش ۱-۱ گفته شد، مسئله خوشه‌بندی و تشخیص اجتماعات بسیار شبیه به هم بوده و مراجع متعددی این دو را یکسان دانستند. بنابراین می‌توان از روش‌های مختلف خوشه‌بندی برای بهبود روش‌های تشخیص اجتماعات استفاده نمود. با توجه به پیچیدگی مسئله و ضعف روش‌های مختلف خوشه‌بندی پایه‌ای، امروزه اکثر مطالعات به سمت خوشه-بندی ترکیبی پرداخته‌اند. بنابراین با توجه به یکسان بودن خوشه‌بندی و تشخیص اجتماعات امکان استفاده از روش-های مختلف برای یکدیگر وجود دارد. در این پایان‌نامه قصد داریم با استفاده از خوشه‌بندی ترکیبی در تشخیص اجتماعات یک روش جدید با نام تشخیص اجتماعات ترکیبی را معرفی نماییم که در بهبود روش‌های مختلف تشخیص اجتماعات مورد استفاده قرار می‌گیرد.

۱-۵ ساختار پایان نامه

ساختار ادامه پایان نامه به این ترتیب می‌باشد که در فصل دوم به مفاهیمی چون معرفی وب‌کاوی و داده‌کاوی ساختار وب و همچنین تشخیص اجتماعات و خوشه‌بندی و تعریف مربوط به آن‌ها و خوشه‌بندی ترکیبی و چگونگی امکان استفاده آن در تشخیص اجتماعات پرداخته می‌شود. در فصل سوم به مرور و مقایسه کارهای انجام شده در زمینه تشخیص اجتماعات و خوشه‌بندی ترکیبی پرداخته می‌شود، در این فصل روش‌های خوشه‌بندی ترکیبی که در این پایان‌نامه مورد استفاده قرار می‌گیرد نیز مورد بررسی قرار گرفته‌اند. در فصل چهارم به بررسی روش پیشنهادی برای تشخیص اجتماعات ترکیبی پرداخته می‌شود، در این فصل روش‌های مختلف تشخیص اجتماعات ترکیبی را مورد بررسی قرار می‌دهیم. در فصل پنجم، تعدادی مجموعه داده استاندارد که در اکثر روش‌های تشخیص اجتماعات استفاده می‌شوند مورد بررسی قرار گرفته است و با استفاده از آن‌ها به ارزیابی روش پیشنهادی و مشاهده نتایج اجرای روش پیشنهادی و روش‌های مختلف تشخیص اجتماعات بر روی مجموعه داده‌های مختلف پرداخته شده است و در فصل ششم یک نتیجه گیری کلی و پیشنهاداتی برای تحقیقات آینده ارائه می‌شود. در انتها نیز فهرست مراجع مورد استفاده در پایان‌نامه آورده شده است.

فصل دوم: مفاهیم موجود در پایان‌نامه

۲-۱ مروری بر ادبیات موضوع

در این فصل به تعریف مفاهیم به کار رفته در این پژوهش از جمله وب‌کاوی، شبکه‌های اجتماعی، تشخیص اجتماعات و خوشه‌بندی ترکیبی می‌پردازیم. همان‌طور که در فصل اول اشاره شد، برای بهره‌برداری از حجم وسیع داده در وب، از روش‌های وب‌کاوی استفاده می‌شود. داده‌کاوی ساختار وب با استفاده از پیوندهای موجود بین صفحات وب اطلاعات زیادی راجع به این صفحات و ارتباطشان به دست می‌آورد. با تحلیل این پیوندها و ارتباطات بین آن‌ها می‌توان وب و یا شبکه اجتماعی را به صورت یک گراف مدل‌سازی کرد که در آن صفحات وب، گره‌های گراف و پیوندهای بین صفحات، یال‌های گراف هستند. حال می‌توان از روش‌های خوشه‌بندی و روش‌های تشخیص اجتماعات بر روی این گراف بهره جست و یکسری نتایج اولیه را از روی آن بدست آورد. بعد از بدست آمدن نتایج اولیه نوبت به خوشه‌بندی ترکیبی می‌رسد که با ترکیب نتایج حاصله می‌تواند نتایج دقیق‌تر، کاراتر، پایدارتر، مستحکم‌تر، مطمئن‌تری را از نتایج اولیه به ارمغان آورد. در این فصل سعی می‌کنیم به طور مختصر درباره این مفاهیم صحبت نماییم.

۲-۲ وب‌کاوی

در [۱] وب‌کاوی به صورت زیر تعریف شده است: وب‌کاوی به کارگیری تکنیک‌های داده‌کاوی برای کشف و استخراج خودکار اطلاعات از اسناد و سرویس‌های وب می‌باشد. وب‌کاوی شامل چهار مرحله اصلی است:

- **پیدا کردن منبع:** این مرحله شامل بازیابی صفحات وب مورد نظر می‌باشد.
- **انتخاب اطلاعات و پیش پردازش:** در این مرحله به صورت خودکار اطلاعات خاصی از صفحات بازیابی شده انتخاب و پیش پردازش می‌شوند.
- **تعمیم**[1]: در این مرحله به طور خودکار الگوهای عام در صفحات بازیابی شده کشف می‌شوند.
- **تحلیل:** در این مرحله الگوهای به دست آمده در مرحله قبل اعتبارسنجی[2] و تفسیر می‌شوند.

---

[1] Generalization

[2] Validation

در مرحله اول داده‌ها از منابع موجود در وب مانند خبرنامه‌های الکترونیکی، گروه‌های خبری، اسناد HTML، پایگاه داده‌های متنی و ... بازیابی می‌شوند. مرحله انتخاب و پیش پردازش شامل هرگونه فرآیند تبدیل داده‌های بازیابی شده در مرحله قبل می‌باشد. این پیش پردازش می‌تواند کاهش کلمات به ریشه آن‌ها[۱]، حذف کلمات زائد[۲]، پیدا کردن عبارات موجود در متن و تبدیل بازنمایی داده‌ها به قالب رابطه‌ای باشد. در مرحله سوم از تکنیک‌های داده‌کاوی و یادگیری ماشین برای تعمیم استفاده می‌شود. در مرحله آخر، الگوهای به دست آمده ارزیابی می‌گردند. در این پایان-نامه دو مرحله آخر وب‌کاوی را مورد بررسی قرار می‌دهیم.

به این ترتیب وب‌کاوی، فرآیند کشف اطلاعات و دانش ناشناخته و مفید از داده‌های وب می‌باشد. این فرآیند به طور ضمنی شامل فرآیند کشف دانش در پایگاه داده‌ها (KDD[۳]) نیز می‌شود. در واقع وب‌کاوی گونه توسعه یافته KDD است که بر روی داده‌های وب عمل می‌کند.

روش‌های وب‌کاوی بر اساس آنکه چه نوع داده‌ای استفاده می‌کنند، به ۳ دسته تقسیم می‌شوند[۴۶]:

- **کاوش محتوای وب:** کاوش محتوای وب فرآیند استخراج اطلاعات مفید از محتوای مستندات وب است. محتوای یک سند وب متناظر با مفاهیمی است که آن سند درصدد انتقال آن به کاربران است. این محتوا می‌تواند شامل متن، تصویر، ویدئو، صدا و یا رکوردهای ساخت یافته مانند لیست‌ها و جداول باشد. در این میان کاوش متن بیش از سایر زمینه‌ها مورد تحقیق قرار گرفته است. از جمله این تحقیقات می‌توان به تشخیص موضوع[۴]، استخراج قواعد انجمنی[۵]، خوشه‌بندی و طبقه‌بندی اسناد وب اشاره کرد. روش‌ها و تکنیک‌های موجود در این گروه، از تکنیک‌های بازیابی اطلاعات و پردازش زبان طبیعی نیز استفاده می‌کنند. هر چند در پردازش تصویر و بینایی ماشین، تحقیقات زیادی در زمینه استخراج دانش از تصاویر انجام شده است، اما به کارگیری این تکنیک‌ها در کاوش محتوای وب چندان چشمگیر نبوده است.
- **داده‌کاوی ساختار وب:** وب را می‌توان با گرافی که گره‌های آن اسناد و یال‌های آن پیوندهای بین اسناد است، بازنمایی کرد. داده‌کاوی ساختار وب، فرآیند استخراج اطلاعات با استفاده از پیوندها می‌باشد.

---

[1] Stemming

[2] Stop Words

[3] Knowledge Discovery in Data Base

[4] Topic Discovery

[5] Association Rule

- **داده‌کاوی استفاده از وب:** داده‌کاوی استفاده از وب، کاربرد تکنیک‌های داده‌کاوی برای کشف الگوهای استفاده از وب، به منظور درک و برآوردن بهتر نیازهای کاربران می‌باشد. این نوع از وب‌کاوی، داده‌های مربوط به استفاده کاربران از وب را مورد کاوش قرار می‌دهد.

باید توجه داشت که مرز مشخصی میان سه گروه وب‌کاوی وجود ندارد. به عنوان مثال تکنیک‌های کاوش محتوای وب می‌توانند علاوه بر به کارگیری متن مستندات، از اطلاعات کاربران هم استفاده کنند. همچنین می‌توان از ترکیب تکنیک‌های فوق برای حاصل شدن نتایج بهتر استفاده کرد[۶]. در این پایان‌نامه از داده‌کاوی ساختار وب برای بدست آوردن اجتماعات استفاده می‌نماییم، در ادامه به بررسی داده‌کاوی ساختار وب خواهیم رسید.

۲-۲-۱ داده‌کاوی ساختار وب

بخش عمده‌ی فعالیت‌ها و تحقیقات انجام شده در وب‌کاوی به محتوای صفحات وب می‌پردازند و کمتر پیوندهای میان صفحات را مد نظر قرار می‌دهند. داده‌کاوی ساختار وب، فرآیند کشف اطلاعات ساختاری از وب می‌باشد. در این نوع کاوش دو نوع داده ساختاری استفاده می‌شود[۶]:

- **پیوندها:** یک پیوند، یک واحد ساختاری است که یک صفحه وب را به صفحه دیگر یا به بخش دیگری از همان صفحه متصل می‌کند. به پیوند نوع اول، پیوند بین سند[1] و به پیوند نوع دوم، پیوند درون سند[2] گفته می‌شود. پیوندها می‌توانند به منظور سهولت در پیمایش صفحات وب استفاده شوند و یا برای ارجاع به صفحاتی که حاوی مطالب مرتبط با صفحه حاوی پیوند هستند، به کار روند.
- **ساختار سند:** محتوای یک صفحه وب می‌تواند بر اساس بر چسب‌های XML و HTML موجود در آن به صورت یک درخت بازنمایی شود.

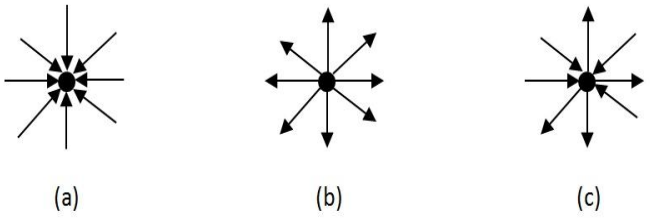
دو اصطلاح که در داده‌کاوی ساختار وب مورد استفاده قرار می‌گیرد Hub و Authorithy می‌باشد که برای مدل نمودن گراف وب استفاده می‌شوند، در ادامه به تعریف این دو می‌پردازیم.

---

[1] Intra Document Hyperlink

[2] Inter Document Hyperlink

می‌توان کل گراف وب را با استفاده از یک گراف بزرگ مدل نمود، برای رسم این گراف از گره‌ها و لبه‌ها در تئوری گراف استفاده می‌شود. گره‌ها در این گراف در شرایطی نام‌های مختلفی می‌گیرند که در اینجا به بررسی نام‌های Hub و Authorithy برای گره‌ها می‌پردازیم. مدل (a) در شکل 2-۱ یک نوع صفحه‌ی وب را بازنمایی می‌کند که به آن Authorithy گفته می‌شود. یک صفحه Authorithy، صفحه‌ای است که صفحات زیاد دیگری به آن اشاره کرده‌اند. مدل (b) نوع دیگری از صفحات وب را بازنمایی می‌کند که به آن Hub گفته می‌شود. یک صفحه Hub، صفحه‌ای است که به صفحات زیاد دیگری اشاره می‌کند. مدل (c) نیز ترکیبی از دو مدل قبل است. از این اصطلاحات در الگوریتم‌ها و روش‌های مختلف وب‌کاوی و تشخیص اجتماع استفاده می‌شود.



شکل 2-۱ Hub و Authority جهت مدل نمودن گراف وب[۶]

داده‌کاوی ساختار وب بر اساس آنکه چه نوع داده ساختاری استفاده می‌کند، به ۲ دسته تقسیم می‌شود:

- **تحلیل پیوند**[1]: به نوعی از داده‌کاوی ساختار وب، که در آن از پیوندها استفاده می‌شود، تحلیل پیوند گفته می‌شود. در تحلیل پیوند تنها پیوندهای میان صفحات به کار گرفته می‌شود. در این پایان‌نامه این نوع تحلیل را مورد بررسی قرار داده‌ایم.
- **تحلیل ساختار سند**[2]: به نوعی از داده‌کاوی ساختار وب که در آن از ساختار داخلی سند استفاده می‌شود، تحلیل ساختار سند گفته می‌شود. کاوش در اینجا بر استخراج مدل شیئ سند[3] متمرکز می‌شود.

البته در داده‌کاوی ساختار وب بیشتر از نوع اول استفاده می‌شود و بسیاری از مراجع داده‌کاوی ساختار وب و تحلیل پیوند را معادل دانسته‌اند[۶].

---

[1] Link Analysis

[2] Document Structure Analysis

[3] Document Object Model

۲-۳ شبکه‌های اجتماعی

شبکه در علوم مختلف دارای کاربردهای مختلف است. شبکه‌های کامپیوتری، شبکه جهانی وب، شبکه‌های بیولوژیکی، شبکه‌های حمل و نقل، شبکه‌های اجتماعی و ... از جمله نمونه‌های شبکه به شمار می‌آیند. مفهوم شبکه، در انواع شبکه‌ها یکسان است ولی مفاهیم گره و لبه در هر حوزه متناسب با همان کاربرد تعریف می‌شود. به عنوان مثال در شبکه‌های کامپیوتری، کامپیوترها و مسیریاب‌ها گره‌های شبکه را تشکیل می‌دهند در حالیکه در شبکه‌های اجتماعی افراد یا سازمان‌ها نقش گره‌ها را ایفا می‌کنند. در اینجا به تعریف مفاهیم موجود در شبکه اجتماعی می-پردازیم.

۲-۳-۱ تعریف شبکه

به طور کلی شبکه شامل دو مؤلفه اصلی است: گره‌ها و لبه‌ها. شبکه‌هایی که به صورت گراف هستند، به شکل $G =< V, E >$ تعریف می‌شوند که در آن گراف G، مؤلفه V شامل مجموعه گره‌های گراف و E مجموعه لبه‌های آن است.

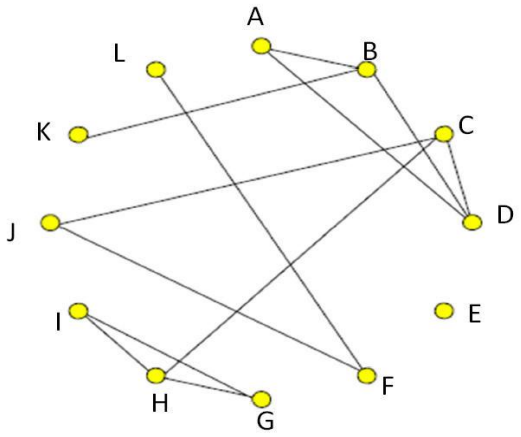
۲-۳-۲ تعریف درجه گره

یکی از مفاهیم پر کاربرد در گراف‌ها مفهوم درجه گره است. درجه هر گره مجموعه تمام لبه‌هایی است که به آن گره متصل است. مجموع درجات تمام گره‌ها دو برابر تعداد لبه‌های گراف است. درجه گره i با k_i نشان داده می‌شود. اگر ماتریس A را ماتریس مجاورتی گراف G گویند، این ماتریس یک ماتریس مربعی به ابعاد تعداد گره‌های گراف متناظر با آن بوده و در صورت وجود ارتباط بین دو گره به آن ماتریس عدد یک و در غیر این صورت عدد صفر منظور می‌گردد.

۲-۳-۲ نحوه نمایش و ذخیره شبکه

شبکه‌ها به دو صورت نمایش داده می‌شوند[۳]:

به صورت گراف: در نمایش شبکه به صورت گراف، گره‌ها و لبه‌ها به وسیله ساختارهای گرافی به هم متصل می‌شوند. شکل 2-۲ نمایی از نحوه نمایش گرافی یک گراف بدون جهت و بدون حلقه را نشان می‌دهد.



شکل 2-۲ نمایش گرافی شبکه

به صورت ماتریس: ماتریس‌ها همان‌طوری که در شکل 2-۳ نشان داده شده است، ابزار دیگری برای نمایش گراف هستند. در گراف‌ها از ماتریس مجاورتی استفاده می‌شود. بدین صورت که هر سطر و هر ستون نمایانگر یک گره است و چنان‌چه لبه‌ای بین این دو گره وجود داشته باشد، درایه متناظر با این دو گره در ماتریس مجاورتی مقدار یک خواهد داشت.

	A	B	C	D	E	G	H	I	J	K	L
A	0	1	0	1	0	0	0	0	0	0	0
B	1	0	0	1	0	0	0	0	0	1	0
C	0	0	0	1	0	0	1	0	1	0	0
D	1	1	1	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	1	0	1
G	0	0	0	0	0	0	1	1	0	0	0
H	0	0	1	0	0	1	0	1	0	0	0
I	0	0	0	0	0	1	1	0	0	0	0
J	0	0	1	0	0	0	0	0	0	0	0
K	0	1	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0

شکل 2-۳ نمایش ماتریس مجاورتی گراف ترسیم شده در شکل 2-۲

در یک شبکه اجتماعی افراد یا سازمان‌ها نقش گره‌های را بازی می‌کنند و در صورت وجود رابطه دوستی یا همکاری، لبه‌ها با اتصال گره‌ها به هم رابطه بین آن‌ها را نشان می‌دهند.

۲-۳-۴ انواع گره در شبکه اجتماعی

در شبکه اجتماعی بعضی از گره‌ها ویژگی متفاوتی دارند. این گره‌ها را بر اساس نوع ارتباطشان با سایر گره‌های شبکه به چند دسته تقسیم می‌کنند[۳]. انواع گره‌ها بر اساس نوع ارتباط بین آن‌ها عبارتند از:

- **گره ستاره**[1]: گره‌هایی که ارتباط زیادی در شبکه اجتماعی دارند، را ستاره می‌نامند. در واقع در شبکه‌های بدون جهت یک معیار شناسایی این گره‌ها درجه آن‌ها است.
- **گره مجزا**[2]: گره‌هایی که ارتباطی با سایر گره‌ها ندارند و یا میزان رابطه آن‌ها خیلی کم است. این گره‌ها معمولاً نقش مؤثری در تشخیص اجتماع ندارند.
- **گره پل**[3]: این گره‌ها نقش واسطه را بین دو گره دیگر بازی می‌کنند. در واقع رابط میان دو گره هستند. چنانچه ممکن است دو گره رابطه مستقیمی با هم نداشته باشند ولی از طریق گره پل به صورت غیر مستقیم به هم متصل شوند. در بعضی موارد این گره‌ها در مرز بین گره‌ها قرار دارد.

۲-۳-۵ نحوه پیوستگی گره‌ها

همچنین می‌توان مجموعه‌ای از گره‌ها را بر اساس میزان پیوستگی بین آن‌ها به چند دسته تقسیم کرد. این تقسیم‌بندی میزان در هم تنیدگی گره‌های یک گراف را نشان می‌دهد. در بعضی روش‌های تشخیص اجتماع از روابطی استفاده می‌شود که به اندازه‌گیری میزان این پیوستگی می‌پردازد[۲].

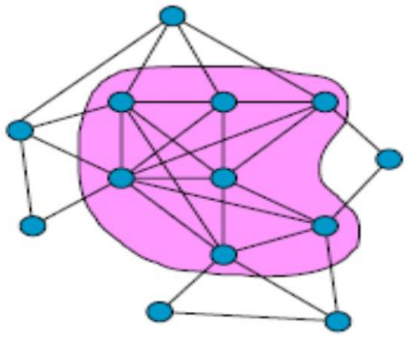
---

[1] Star

[2] Isolate

[3] Bridge

- **گره‌های کاملاً پیوسته**[1]: مجموعه‌ای از گره‌ها که شامل حداقل سه گره هستند و همه آن‌ها به طور کامل با هم ارتباط دارند.
- **گره‌های شبه پیوسته**[2]: گره‌هایی که چگالی اتصال آن‌ها زیاد است ولی کاملاً پیوسته نیستند.
- **گره‌های کاملاً پیوسته و دارای همپوشانی**[3]: در بعضی از گراف‌ها، زیر مجموعه‌ای از گره‌های گراف دارای چند پیوستگی کامل هستند که با هم همپوشانی دارند. در بخش هاشور زده شکل 2-۴ یک زیر گراف دارای همپوشانی نشان داده شده است.



شکل 2-۴ نمونه‌ای از گراف کاملاً پیوسته و دارای همپوشانی[۲]

پس از بررسی مفاهیم پایه در شبکه‌ها به خصوص در شبکه‌های اجتماعی به تعریف اجتماع می‌پردازیم.

۲-۴ تعریف اجتماع

از آنجا که امروزه حجم وب هر روزه در حال افزایش است، تشخیص اجتماعات وب و رابطه بین آن‌ها روز به روز دشوارتر می‌شود. در مراجع مختلف تعاریف متفاوتی از اجتماع ارائه شده است که برخی از آن‌ها در ادامه ذکر می‌شود:

- یک اجتماع وب مجموعه‌ای از صفحات وب است که توسط افراد یا سازمان‌های مختلف که علایق مشترک درباره یک موضوع خاص دارند، ایجاد شده‌اند. مانند صفحات علاقه‌مندان یک تیم بیس بال و یا صفحات تولید کنندگان تجهیزات کامپیوتری[۴].

---

[1] Clique

[2] Quasi-clique

[3] Overlapping clique

- در تعریف دیگری از اجتماع آمد است در یک اجتماع چگالی لبه‌ها داخل آن بیشتر از چگالی لبه‌های بین آن با سایر گره‌های شبکه است[۵].
- یک اجتماع مجموعه‌ای از گره‌های مرتبط به یکدیگر است که پیوندهای بین آن‌ها متراکم است[۷].
- در [۸] تعریفی ارائه شده که بر پایه مقایسه چگالی لبه‌ها در بین گره‌هاست. اجتماعات در دو حالت قوی و ضعیف تعریف شده‌اند. در حالت قوی، یک زیر گراف اجتماع است که در بین گره‌های آن اتصالات بیشتری نسبت به سایر گره‌های گراف باشد. تعریف دقیق‌تر این است که مجموعه گره‌های V در حالت قوی در یک اجتماع هستند اگر $k_i^{in} > k_i^{out}, \forall i \in V$ این بدان معنی است که هر گره i در اجتماع V تعداد لبه‌های بیشتری با گره‌های اجتماع خود نسبت به سایر گره‌های شبکه داشته باشد. در حالت ضعیف، یک زیر گراف اجتماع است اگر $\sum_{i \in V} k_i^{in} > \sum_{i \in V} k_i^{out}$ باشد. به عبارت دیگر مجموع تعداد لبه‌های گره‌های داخل اجتماع V بزرگتر از مجموع تعداد لبه‌های گره‌های داخل اجتماع با گره‌های خارج اجتماع است.
- در تعاریف دیگری رابطه شباهت بین گره‌های گراف تعیین می‌شود. در گره‌هایی که شباهت بیشتری دارند در یک اجتماع قرار می‌گیرند. برای تعیین میزان شباهت نیز از معیارهایی نظیر فاصله اقلیدسی و همبستگی پیرسن استفاده می‌شود.

همان‌طوری که مشاهده می‌شود تعاریف متفاوتی برای اجتماع ارائه شده است، در مقالات مختلف تعاریف دیگری هم وجود دارد. لازم به ذکر است که با وجود اهمیت مفهوم اجتماع، توافق آرایی از تعریف اجتماع وجود ندارد.

۲-۴-۱ انواع اجتماعات

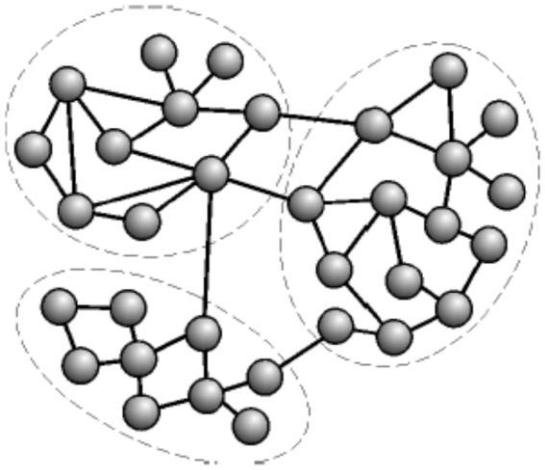
اجتماع را می‌توان از نظر چگالی گره‌های موجود در اجتماع به دو دسته تقسیم کرد. اجتماعات در شبکه به دو دسته اصلی متجانس و نامتجانس تقسیم می‌شوند[۳]:

- **اجتماع نامتجانس**[1]: بسیاری از شبکه‌های واقعی ساختار گرافی نامتجانس را نشان می‌دهند. این بدان معناست که تقسیم‌بندی گره‌ها در بین اجتماعات مساوی نیست. برخی از اجتماعات دارای گره‌های فراوانی

[1] Incongruence

هستند در حالیکه اجتماعاتی هستند که چند گره بیشتر ندارند. این موضوع در روش‌های تشخیص اجتماع دارای اهمیت است. چرا که روش تشخیص اجتماع باید قابلیت تشخیص اجتماعات نامتجانس را داشته باشد. این ساختار نامتجانس در بسیاری انواع شبکه‌ها مانند شبکه‌های اجتماعی، شبکه‌های متابولیکی، شبکه‌های وب و... یافت می‌شوند.

- **اجتماع متجانس**[1]: اجتماعاتی که توزیع گره‌ها در بین آن‌ها مساوی است. در واقع می‌توان گفت به طور تقریبی تعداد گره‌ها به طور یکسان بین اجتماعات تقسیم شده‌اند. شکل 2-۵ شبکه با اجتماعات متجانس را نشان می‌دهد که تقسیم‌بندی گره‌ها در بین اجتماعات به طور مساوی است.



شکل 2-۵ نمونه‌ای از شبکه متجانس، اجتماعات توسط خط چین از هم جدا شده‌اند[۳]

گره‌هایی که در یک اجتماع قرار دارند ویژگی‌های مشترکی دارند. این ویژگی‌های اجتماعات موجود در شبکه باعث درک بهتری از ساختار گراف می‌شود. برای مثال در شبکه وب، صفحاتی که موضوع مشابهی دارند پس از تشخیص اجتماع ، در یک اجتماع قرار می‌گیرند، بنابراین تشخیص این اجتماعات می‌تواند به جستجوی اطلاعات کمک بسزایی کند.

حال که با مفهوم شبکه‌های اجتماعی و اجتماع آشنا شدیم، به بررسی تشخیص اجتماع می‌پردازیم.

[1] Congruence

۲-۵ تشخیص اجتماع

در شبکه‌های اجتماعی معمولاً افراد به دنبال حلقه‌های دوستی می‌گردند. در این شرایط یافتن یک دوست آن‌ها را به یک گروه دوستی پیوند می‌زند. در این میان نقش تشخیص اجتماع جستجو برای پیدا کردن این اجتماعات است. اجتماع مجموعه‌ای از گره‌هاست که اطلاعاتی را با هم به اشتراک می‌گذارند، مثلاً افرادی که علاقه‌مندی‌های مشابه دارند یا وب سایت‌های که دارای محتوای مشابهی هستند. در شبکه‌های اجتماعی افرادی که رابطه دوستی یا همکاری به هم دارند، معمولاً تشکیل یک اجتماع می‌دهند. اتصال بین گره‌های اجتماع می‌تواند کاملاً پیوسته، شبه پیوسته یا پیوستگی بین اجتماعات دارای همپوشانی باشد. در بیشتر شبکه‌های اجتماعی افراد یک اجتماع پیوستگی کامل ندارند. زیرا ویژگی شبکه‌های اجتماعی این است که گراف آن‌ها خلوت است که در آن تعداد گره‌ها و لبه‌ها با هم تقریباً برابر است[۵].

در [۲] تشخیص اجتماعات را بدین‌صورت تعریف می‌نماید که، اگر A را ماتریس مجاورتی یک گراف G در نظر بگیریم، درجه یک گره در گراف را با k_i نشان می‌دهند که برابر است با: $k_i = \sum_j A_{ij}$ می‌باشد. با در نظر گرفتن یک زیر گراف S به طوری که $S \subseteq G$ است.

$$S \subseteq G \Rightarrow k_i(S) = k_i^{in}(S) + k_i^{out}(S) \qquad (۱-۲)$$

در رابطه (۲-۱) نیز درجه هر گره را برای اجتماعات داخلی و خارجی به صورت $k_i^{in}(S) = \sum_{j \in S} A_{ij}$ و $k_i^{out}(S) = \sum_{j \notin S} A_{ij}$ تعریف می‌کنیم.

حال به مسئله اصلی که همان تشخیص اجتماع است، می‌پردازیم.

$$\sum_{i \in S} k_i^{in}(S) >> \sum_{i \notin S} k_i^{out}(S) \qquad (۲-۲)$$

با توجه به رابطه (۲-۲)، $\sum_{i \in S} k_i^{in}(S)$ مجموع درجه‌های تمامی گره‌هایی که در اجتماع یا زیر گراف S وجود دارد می‌باشد و $\sum_{i \notin S} k_i^{out}(S)$ مجموع درجه‌های تمامی گره‌هایی که در اجتماع یا زیر گراف S وجود ندارد می‌باشد.

روش‌های مختلفی برای بدست آوردن اجتماعات در شبکه وجود دارد. در این فصل به بررسی برخی از این روش‌ها می‌پردازیم. بررسی تک‌تک گره‌ها و قرار دادن آن‌ها در اجتماعات مختلف و در نهایت ارزیابی آن دارای هزینه زمانی و محاسباتی بسیار بالایی است و این رویکرد به طور عملی امکان‌پذیر نیست. برای غلبه بر این مشکل روش‌های مبتنی بر تجربه[۱] به کمک آمده‌اند. روش‌های طیفی[۲]، تقسیم‌کننده[۳]، تجمعی[۴]، روش‌های مبتنی بر بهینه‌سازی پودمانی[۵]، تکنیک‌های حریصانه[۶]، الگوریتم‌های تکاملی[۷] [۹] از جمله آن‌هاست. هر کدام از این روش‌ها به نوعی در بهبود عملکرد تشخیص اجتماع موثر است. در بخش بعدی به تشریح هر کدام از روش‌ها می‌پردازیم.

۲-۵-۱ انواع تشخیص اجتماع

روش‌های تشخیص اجتماع به دو نوع کلی تقسیم می‌شوند که عبارتند از: تشخیص اجتماعات ریزدانه[۸] و درشت دانه[۹][۳].

ریزدانه: در این روش، تقسیم‌بندی در تعداد زیادی اجتماع انجام می‌شود. در واقع با افزایش مقیاس شبکه، تعداد اجتماعات نیز افزایش می‌یابد.

درشت‌دانه: در این روش، تقسیم‌بندی در تعداد مشخصی اجتماع انجام می‌شود. به عبارتی با افزایش تعداد گره-های شبکه، تعداد اجتماعات تغییر نمی‌کنند.

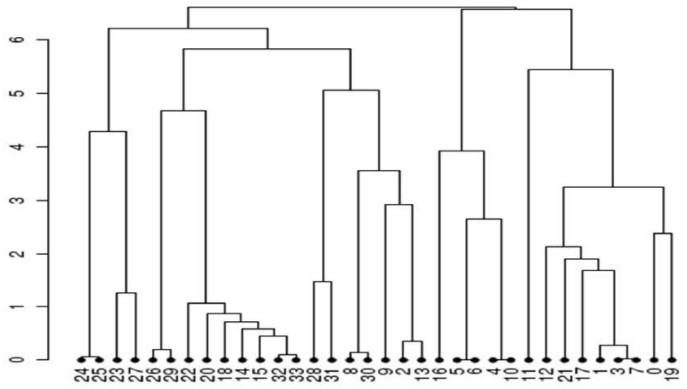
در روش ریزدانه به نوعی پویایی در تعداد اجتماعات وجود دارد، در حالیکه درشت‌دانه بر روی افزایش اجتماعات تمرکز ندارند و سعی می‌کند گره‌ها را در اجتماعات موجود قرار دهد. هنگامی که تعداد اجتماعات حدوداً مشخص است روش درشت‌دانه پاسخگوی مناسبی است.

---

[1] Heuristic
[2] Spectral methods
[3] Divisive
[4] Agglomerative
[5] Modularity optimization
[6] Greedy techniques
[7] Evolutionary
[8] Fine-grained
[9] Coarse-grained

۲-۵-۲ نمودار بایگان درختواره

در بعضی روش‌های تشخیص اجتماع می‌توان نموداری به نام درختواره[1] ترسیم کرد. درختواره در روش‌های بایگان[2] نمایان گر نحوه عملکرد روش است. در این نمودار مراحل به هم پیوستن گره‌ها و تشکیل اجتماعات نشان داده می‌شود. شکل 2-۶ نمونه‌ای از یک درختواره را نشان می‌دهد. اعداد، گره‌های گراف هستند که به وسیله یک معیار مشابهت به هم متصل شده‌اند، تا در نهایت کل شبکه را تشکیل دهند. در روش‌های تشخیص اجتماع بعد از یافتن درختواره ، برای دست‌یابی به اجتماعات، درختواره را برش می‌زنند. با هر برشی که بر روی این درختواره داده می‌شود، یکسری اجتماعات ایجاد خواهد شد، نکته قابل توجه این است که درختواره از کدام ناحیه برش داده شود، که اجتماعات صحیح را برای ما داشته باشد. در [۱۰] اشاره به این مطلب دارد که هرچه فاصله میان خوشه‌ها از هم بیشتر باشد آن محل بهترین حالت برای برش دادن درختواره است.



شکل 2-۶ نمونه‌ای از درختواره

در ادامه‌ی این فصل به معیار ارزیابی در تشخیص اجتماعات می‌پردازیم.

[1] Dendogram

[2] Hierarchical

۲-۶ پودمانی

پودمانی[1] یکی از معیارهای ارزیابی و تشخیص اجتماع می‌باشد. پودمانی از پرکاربردترین معیارهای مورد استفاده در روش‌های مختلف است. این معیار کمیتی از گروه‌بندی برای کل گراف بدست آمده، ارائه می‌کند و نقش بسزایی در تعیین صحت گروه‌بندی دارد. معمولاً برای ارزیابی هر روش تشخیص اجتماع در شبکه، مقدار پودمانی گروه‌بندی پیشنهادی آن روش را برای شبکه‌ها و گراف‌های مختلف محاسبه می‌کنند، هر چه پودمانی به دست آمده بیشتر باشد، دقت روش مورد نظر بهتر بوده است، معیار پودمانی مهم‌ترین محک در ارزیابی روش‌های تشخیص اجتماع در شبکه-های اجتماعی است.

پودمانی به صورت زیر تعریف می‌شود[۱۱]:

$$Q = \sum_i (e_{ii} - a_i^2) \qquad (۳-۲)$$

$$a_i = \sum_j e_{ij} \qquad (۴-۲)$$

$$e_{ij} = \begin{cases} \frac{1}{2}\left(\frac{No.edges\ connect\ community\ i\ to\ community\ j}{No.all\ edges}\right) & i \neq j \\ \left(\frac{No.edges\ inside\ community\ i}{No.all\ edges}\right) & i \equiv j \end{cases} \qquad (۵-۲)$$

در این روابط i و j اندیس‌های اجتماع است و در رابطه (۲-۳) e_{ii} نسبت تعداد لبه‌هایی که گره‌های داخل اجتماع i را به هم متصل می‌کند به کل لبه‌های گراف است. به عبارتی دیگر کسری از تعداد لبه‌ها که دو سر آن در اجتماع i قرار دارد به کل لبه‌های گراف و در رابطه (۲-۴) a_i نسبت تعداد لبه‌هایی که حداقل یک گره آن در اجتماع i است به کل لبه‌های گراف می‌باشد. برای محاسبه a_i از e_{ij} استفاده می‌شود. در رابطه (۲-۵) e_{ij} بدین صورت تعریف می‌شود که اگر $i = j$ «گره‌های داخل یک اجتماع» باشد، نسبت تعداد لبه‌های داخل اجتماع i به کل لبه‌های گراف در نظر گرفته می‌شود و اگر $i \neq j$ باشد، نصف تعداد لبه‌های داخل اجتماع i و j را مدنظر قرار می‌گیرد. مجموع این دو حالت برای کل اجتماعات شبکه که با اجتماع i در ارتباط هستند رابطه a_i را می‌سازد.

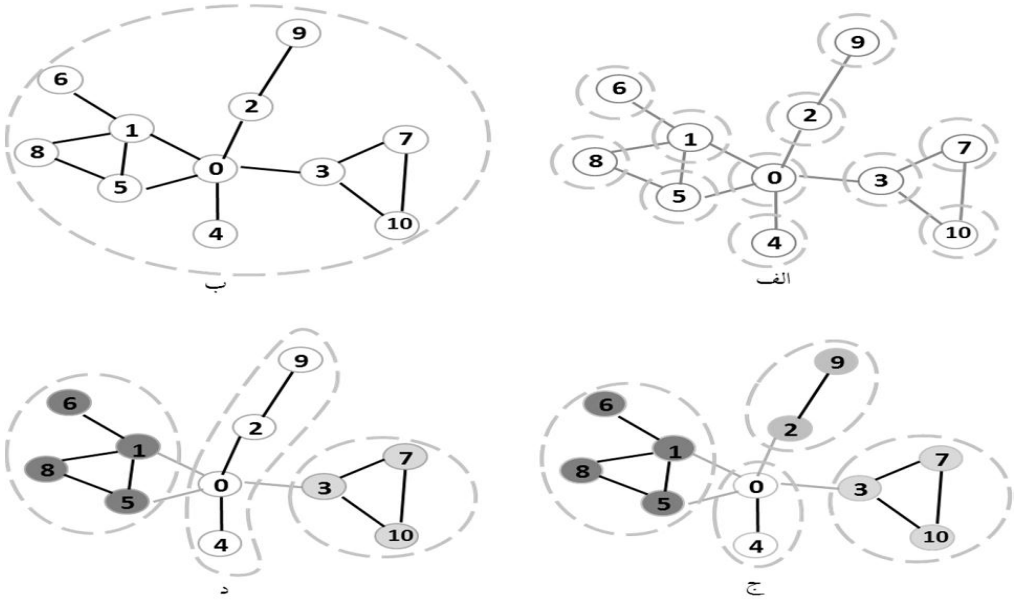
---

[1] Modularity

در رابطه پودمانی مقدار $[e_{ii} - a_i^2]$ سهم هر اجتماع i را در کل شبکه بیان می‌کند. هر چه این عدد بزرگتر باشد، این اجتماع سهم بیشتری در شبکه دارد، در واقع اجتماع قوی‌تری است. بدین معنی که ارتباطات داخل اجتماع در آن بیشتر از ارتباط آن با سایر اجتماعات شبکه است.

اگر کل گراف شامل تنها یک اجتماع باشد یا گره‌ها به صورت تصادفی در بین اجتماعات قرار گرفته باشند $Q = 0$ خواهد بود. هر چه مقدار Q به 1 نزدیک‌تر باشد اجتماعات بهتر جدا شده‌اند ولی هیچ گاه مقدار آن یک نمی‌شود. در عمل مقدار $Q > 0$ ساختار گروهی مناسبی را نشان می‌دهد. پودمانی می‌تواند مقادیر منفی را هم بپذیرد.

مقدار پودمانی ممکن است برای یک شبکه در بهترین حالت گروه‌بندی از عدد خاص بیشتر نشود. این بدان معناست که لزوماً نمی‌توان گفت اگر پودمانی مقدار 0.4 دارد، تشخیص اجتماع به درستی انجام نشده است. ممکن است 0.4 بیشترین مقداری باشد که پودمانی در آن شبکه دارد. در شکل 2-۷ پودمانی را برای یک گراف آزمایشی در اجتماعات مختلف نشان داده‌ایم. این گراف دارای 11 گره و 13 لبه است. در شکل 2-۷-الف هر گره در یک اجتماع قرار داده شده است. این نوع گروه‌بندی دارای خطای زیادی است و در نتیجه پودمانی این گروه‌بندی منفی و برابر $Q = -0.115$است. در شکل 2-۷-ب کل گره‌های گراف در یک اجتماع قرار داده شده است. رابطه پودمانی به گونه‌ای است که در این حالت انگار گروه‌بندی انجام نشده است و پودمانی گراف برابر صفر است. دو شکل ج و د دو گروه‌بندی دیگر از گراف آزمایشی را نشان می‌دهد. بهترین گروه‌بندی مربوط به شکل 2-۷-د است که بیشترین پودمانی را بدست می‌دهد.



شکل 2-۷ نمایی از یک گراف آزمایشی در گروه‌بندی‌های مختلف

پودمانی بیشینه که برای گراف آزمایشی محاسبه شده برابر با $Q = 0.451$ است. همان‌طور که مشاهده می‌کنید، بیشترین مقدار پودمانی برای این گراف به 0.5 نمی‌رسد. این بدان معناست که لزوماً بهترین گروه‌بندی برای یک گراف پودمانی نزدیک به 1 ندارد.

به طور کلی ویژگی پودمانی را می‌توان به صورت زیر خلاصه کرد:

- ✓ مقدار آن در بازه $(-1,+1)$ قرار دارد.
- ✓ مقادیر نزدیک به 1 نشان می‌دهد شبکه ساختار گروهی منسجمی دارد.
- ✓ اگر همه‌ی گره‌ها در یک اجتماع باشند مقدار پودمانی صفر است.
- ✓ اگر پودمانی بزرگتر از 0.3 داشته باشد، گروه‌بندی گراف مناسب است.

در [۱۲] فرم جدیدی از رابطه پودمانی ارائه شده است که با درجه گراف کار می‌کند. برای توضیح آن از ماتریس مجاورتی شروع می‌کنیم. A_{vw} در رابطه (۲-۶) ماتریس مجاورتی را نشان می‌دهد.

$$A_{vw} = \begin{cases} 1 & if\ vertices\ v\ and\ w\ is\ connected \\ 0 & otherwise \end{cases} \qquad (۲-۶)$$

فرض کنید گراف به چند اجتماع تقسیم شده است به‌طوریکه گره v متعلق به اجتماع C_v است.

لبه‌هایی که گره‌های داخل اجتماع را به هم متصل می‌کند به صورت رابطه (۲-۷) تعریف می‌شود:

$$\frac{\sum_{vw} A_{vw}\delta(C_v, C_w)}{\sum_{vw} A_{vw}} = \frac{1}{2m} \sum_{vw} A_{vw}\delta(C_v, C_w) \qquad (۲-۷)$$

که در تابع $\delta(i, j)$ برای $i = j$ مقدار 1 دارد و در غیر این صورت مقدار آن صفر است و $m = \frac{1}{2} \sum_{vw} A_{vw}$ تعداد کل لبه‌های گراف است.

درجه k_v تعداد لبه‌های متصل به گره v را در رابطه (۲-۸) محاسبه می‌کند.

$$k_v = \sum_{w} A_{vw} \qquad (۲-۸)$$

فرم جدید پودمانی برای ماتریس مجاورتی A_{vw} به صورت رابطه (۲-۹) تعریف می‌شود:

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(C_v, C_w) \qquad (۲-۹)$$

این رابطه کمک می‌کند تا مستقیماً از ماتریس مجاورتی و درجه گره‌ها برای محاسبه پودمانی استفاده کنیم. هر دو رابطه پودمانی، نتیجه یکسانی را بدست می‌آورد.

۲-۷ خوشه‌بندی ترکیبی

روش‌های خوشه‌بندی گروهی از عناصر مشابه را در میان حجم وسیعی از داده‌ها جستجو می‌کنند. ایده اصلی خوشه‌بندی اطلاعات، جدا کردن نمونه‌ها از یکدیگر و قرار دادن آن‌ها در گروه‌ها شبیه به هم می‌باشد. به این معنی که نمونه‌های شبیه به هم باید در یک اجتماع قرار بگیرند و با نمونه‌های اجتماعات دیگر حداکثر تفاوت را دارا باشند[۱۳]. در واقع خوشه‌بندی داده‌ها یک ابزار ضروری برای یافتن اجتماعات در داده‌های بدون برچسب است. دلایل اصلی اهمیت خوشه‌بندی عبارتند[۱۴]:

- جمع‌آوری و برچسب‌گذاری یک مجموعه بزرگ از الگوهای نمونه می‌تواند بسیار با ارزش باشد.
- با خوشه‌بندی می‌توانیم یک دید و بینشی از طبیعت و ساختار داده به دست آوریم. کشف زیر رده‌های مجزا یا شباهت‌های بین الگوها ممکن است به طور چشمگیری در روش‌های مختلف به ما پیشنهاد ارایه کند.

خوشه‌بندی داده‌ها یکی از مراحل اصلی در داده‌کاوی است که وظیفه کاوش الگوهای پنهان در داده‌های بدون برچسب را بر عهده دارد. به خاطر پیچیدگی مسئله و ضعف روش‌های خوشه‌بندی پایه، امروزه اکثر مطالعات به سمت روش‌های خوشه بندی ترکیبی هدایت شده است. از آنجایی که اکثر روش‌های خوشه‌بندی پایه روی جنبه‌های خاصی از داده‌ها تاکید می‌کنند، در نتیجه روی مجموعه داده‌های خاصی کارآمد می‌باشند. به همین دلیل، نیازمند روش‌هایی هستیم که بتواند با استفاده از ترکیب این الگوریتم‌ها و گرفتن نقاط قوت هر یک، نتایج بهینه‌تری را تولید کند. در واقع هدف اصلی خوشه‌بندی ترکیبی جستجوی نتایج بهتر و مستحکم‌تر، با استفاده از ترکیب، اطلاعات و نتایج حاصل از چندین خوشه‌بندی اولیه است. پراکندگی در نتایج اولیه یکی از مهم‌ترین عواملی است که می‌تواند در کیفیت نتایج نهایی خوشه‌بندی ترکیبی اثر گذار باشد. همچنین، کیفیت نتایج اولیه نیز عامل دیگری است که در کیفیت نتایج حاصل از ترکیب موثر است و در آن به طور خلاصه خوشه‌بندی ترکیبی شامل دو مرحله اصلی زیر می‌باشد[۸]:

- تولید نتایج متفاوت از خوشه‌بندی‌ها، به عنوان نتایج خوشه‌بندی اولیه بر اساس اعمال روش‌های مختلف که این مرحله را، مرحله ایجاد تنوع یا پراکندگی[1] می‌نامند.
- ترکیب نتایج به دست آمده از خوشه‌بندی‌های متفاوت اولیه برای تولید خوشه نهایی؛ که این کار توسط تابع توافقی[2] (الگوریتم ترکیب کننده) انجام می‌شود.

بنابراین به صورت کلی این دو مرحله در اصل یک چارچوب برای هر روش ترکیبی خواهد بود. در ابتدا نتایج مختلف را از اجرای روش‌های پایه بدست می‌آوریم و سپس با اجرای یک تابع توافقی که در بخش ۴-۳ به آن پرداخته می‌شود نتایج مختلف را با هم ترکیب می‌کنیم تا یک نتیجه دقیق و مستحکم و از همه مهم‌تر پایدار را ایجاد نماییم.

۲-۸ خلاصه‌ی این فصل

در این فصل به مفاهیم موجود در این پایان‌نامه از قبیل: وب‌کاوی، تعریف اجتماع، تشخیص اجتماعات، پودمانی، خوشه‌بندی و خوشه‌بندی ترکیبی پرداختیم. تشخیص اجتماعات در وب مسئله مهمی می‌باشد که دارای جای کار زیاد به علت جدید بودن این موضوع می‌باشد. در دو فصل آتی با استفاده از این تعاریف و بررسی کارهای انجام شده در این حوزه‌ها به ترکیب روش‌های خوشه‌بندی در تشخیص اجتماعات می‌پردازیم و روشی را با نام تشخیص اجتماعات ترکیبی را معرفی می‌نماییم. این روش پیشنهادی نتایج دقیق‌تر و مستحکم‌تری را ارائه خواهد نمود که در فصل‌های بعدی به بررسی و ارزیابی این روش با معیار تعریف شده پودمانی خواهیم پرداخت.

---

[1] Diversity

[2] Consensus Function

فصل سوم : مرور و مقایسه کارهای انجام شده

در این فصل ابتدا به بررسی روابط و روش‌های در تشخیص اجتماعات می‌پردازیم و سپس خوشه‌بندی ترکیبی را مورد بررسی قرار می‌دهیم. در نهایت امکان استفاده از خوشه‌بندی ترکیبی در تشخیص اجتماعات مورد بررسی قرار خواهد گرفت.

۳-۱ مقدمه

در این فصل به بررسی مقالات ارائه شده در زمینه وب‌کاوی، تشخیص اجتماعات و نیز خوشه‌بندی ترکیبی می‌پردازیم. بحث تشخیص اجتماعات در چند سال اخیر در کنفرانس‌ها و مجلات مورد توجه واقع شده است. نکته جالب این است که پیشتازان جدید این عرصه از فیزیکدانان هستند و این موضوع در حوزه فیزیک از محبوبیت خاصی برخوردار است و مقالات بسیاری را شامل می‌شود. از جمله این فیزیکدانان گیروان[۱] و نیومن[۲] هستند، که رابطه پودمانی را در تشخیص اجتماع پایه‌گذاری کرده‌اند. پس از آن نیز مقالات زیادی در این حوزه ارائه شده است. همچنین کارهای انجام شده و مقالاتی در زمینه خوشه‌بندی ترکیبی را مورد بررسی قرار خواهیم داد. ابتدا به توضیح روش‌هایی که هر کدام در مقالات متعددی استفاده شدند، می‌پردازیم و سپس آن‌ها را دسته‌بندی و مقایسه می‌کنیم.

۳-۲ الگوریتم‌های داده‌کاوی ساختار وب

در این بخش به بررسی الگوریتم‌هایی که در داده‌کاوی ساختار وب به کار می‌روند، پرداخته می‌شود. کل وب را می‌توان به صورت یک گراف جهت‌دار که شامل مجموعه‌ای از گره‌ها و یال‌های جهت‌دار است، مدل‌سازی کرد. گره‌ها، صفحات وب و یال‌ها پیوندهای میان آن‌ها هستند.

الگوریتم‌هایی که در این بخش بررسی خواهند شد، الگوریتم‌های HITS[۳] و Page Rank، برای بازیابی صفحات وب و رتبه‌بندی آن‌ها بر اساس میزان ارتباط با پرس‌وجوی کاربر استفاده می‌شوند. این الگوریتم‌ها از دسته الگوریتم-

---

[1] Girvan

[2] Newman

[3] Hyperlink-Induced Topic Search

هایی به شمار می‌آیند که برای یک صفحه وب به کار می‌روند. اما الگوریتم‌هایی که برای تشخیص اجتماعات معرفی خواهند شد، از دسته الگوریتم‌هایی به شمار می‌آیند که برای چندین صفحه وب مورد استفاده قرار می‌گیرند.

HITS ۱-۲-۳

الگوریتم HITS یکی از الگوریتم‌های رایج برای رتبه‌بندی صفحات وب بر اساس میزان ارتباط آن‌ها با پرس‌وجوی کاربر است که در سال ۱۹۹۹ توسط Kleinberg ارائه شد[۱۵]. این الگوریتم یک روش وابسته به پرس‌وجو است. در این نوع روش‌ها برای هر پرس‌وجو تحلیل پیوندها انجام می‌شود. برای انجام تحلیل پیوند، ابتدا گراف خاص پرس‌وجو به نام گراف همسایگی[1] ساخته می‌شود که در حالت ایده آل تنها شامل صفحات مرتبط با موضوع پرس‌وجو است. برای ساخت این گراف، ابتدا یک مجموعه از اسناد مرتبط با پرس‌وجو، به وسیله موتور جست‌وجو واکشی می‌شوند. به این مجموعه، مجموعه ریشه[2] گفته می‌شود. سپس مجموعه ریشه به وسیله همسایگانش تکمیل می‌گردد. همسایه‌ها، مجموعه‌ای از اسناد هستند که یا از اسناد موجود در مجموعه ریشه به آن‌ها پیوند داده شده است و یا به اسناد موجود در مجموعه ریشه پیوند داده‌اند. از آنجا که تعداد اسنادی که به اسناد موجود در مجموعه ریشه پیوند داده‌اند ممکن است عدد بزرگی شود، این عدد محدود و برای تعداد این اسناد حدی در نظر گرفته می‌شود. به این مجموعه جدید، مجموعه پایه[3] یا گراف همسایگی گفته می‌شود. سپس الگوریتم HITS برای هر گره در گراف همسایگی، به طور تناوبی دو امتیاز Authority و Hub را محاسبه و گره‌ها را با توجه به این امتیازات رتبه‌بندی می‌نماید. گره‌های با امتیاز بالای Authority، Authority خوب و گره‌های با امتیاز بالای Hub، Hub خوبی هستند. این الگوریتم فرض می‌کند سندی که به اسناد دیگر بیشتری اشاره می‌کند، Hub خوبی است، و سندی که اسناد بیشتری به آن اشاره می‌کنند، Authority خوبی می‌باشد. به طور بازگشتی می‌توان نتیجه گرفت سندی که به تعداد Authority های خوب بیشتری اشاره می‌کند، Hub بهتری است و سندی که Hub های خوب بیشتری به آن اشاره می‌کنند، Authority بهتری می‌باشد. الگوریتم بازگشتی برای محاسبه امتیاز Hub و Authority به صورت زیر بیان می‌شود:

۱. V، مجموعه گره‌ها در گراف همسایگی در نظر گرفته می‌شود.

---

[1] Neighborhood Graph

[2] Root Set

[3] Base Set

۲. مقدار اولیه Hub[A] برای همه گره‌ها ۱ می‌باشد.

۳. تا وقتی که دو بردار Aut و Hub در رابطه (۳-۱) همگرا نشده‌اند:

- ✓ برای همه A های موجود در $Aut[A] = \sum_{(B,A)\in V} Hub[B]$
- ✓ برای همه A های موجود در $Hub[A] = \sum_{(A,B)\in V} Aut[B]$
- ✓ بردارهای Hub و Aut به‌هنجار می‌شوند.

$$Aut[A] = \sum_{(B,A)\in V} Hub[B] \quad \text{و} \quad Hub[A] = \sum_{(A,B)\in V} Aut[B] \qquad (۱-۳)$$

جبر خطی نشان می‌دهد که بردارهای Hub و Aut در نهایت همگرا خواهند شد. اما تعداد دفعات تکرار در حلقه بالا مشخص نیست.

از مزایای این روش، ارائه دو لیست مرتب شده، یکی بر اساس امتیاز Hub و دیگری بر اساس امتیاز Authority و در نظر گرفتن تعداد محدودی از صفحات وب و در نتیجه کوچک‌سازی مسئله است. از مشکلات HITS وابسته بودن آن به پرس‌وجو است، به طوری که برای هر پرس‌وجو باید یک گراف همسایگی ساخته شود و امتیاز Hub و Authority محاسبه شود. مشکل دیگر، فریب خوردن HITS از کاربران است. کاربران با ایجاد پیوندهایی در/به صفحات، امتیاز Hub را تغییر می‌دهند و از آنجا که مقدار Authority از روی Hub به دست می‌آید، آن هم تغییر می‌کند. همچنین اگر موضوع بیشتر صفحات گراف همسایگی با موضوع پرس‌وجو متفاوت باشد، صفحات موجود در لیست Authority وHub متفاوت خواهند بود. این الگوریتم برای پرس‌وجوهایی که تعداد پیوندهای میان صفحات مرتبط با موضوع، در آن‌ها زیاد است، موفق‌تر عمل می‌کند. اگر پرس‌وجو برای موضوعاتی به کار رود که خاص هستند و حجم پیوندها اندک است، نتایج تولید شده معمولاً مرتبط با یک موضوع عام‌تر خواهد بود.

با توجه به مشکلات الگوریتم HITS، محققین به روش‌های متفاوت آن را بهبود داده‌اند[۱۶][۱۷].

Page Rank ۲-۲-۳

الگوریتم Page Rank که اولین بار در سال ۱۹۹۸ توسط Larry Page و Sergey Brin ارائه شده است[۱۸]، یک روش مستقل از پرس‌وجو[1] می‌باشد. این روش یک بار به هر سند وب امتیاز اختصاص می‌دهد و از این امتیاز، با یا بدون در نظر گرفتن معیاری با توجه به پرس‌وجوی کاربر جهت رتبه‌بندی اسناد استفاده می‌کند.

این الگوریتم رتبه هر صفحه را با اختصاص وزن به پیوندی که به آن صفحه داده شده است به دست می‌آورد. مقدار این وزن به کیفیت صفحه‌ای که پیوند در آن قرار گرفته، بستگی دارد. در این صورت پیوندهای صفحات مهم‌تر وزن بیشتری می‌گیرند. جهت مشخص کردن کیفیت صفحه‌های رجوع کننده، در Page Rank از رتبه آن صفحه که به صورت بازگشتی تعیین و مقدار اولیه آن اختیاری است، استفاده می‌شود. اگر n سند در دسترس باشد، مقدار اولیه رتبه سند را می‌توان برابر 1/n در نظر گرفت.

رتبه هر صفحه مانند P طبق رابطه (۲-۳) محاسبه می‌شود که B_P مجموعه همه صفحات اشاره کننده به P می‌باشد. در این رابطه ε مقدار ثابتی بین ۰.۱ و ۰.۲، n تعداد گره‌ها در گراف G (تعداد صفحات وب در مجموعه) و Outdegree(Q) تعداد پیوندهای خروجی موجود در صفحه Q است. رتبه مرحله j صفحه P_i طبق رابطه زیر محاسبه می‌شود:

$$r_j(Pi) = \frac{\varepsilon}{n} + (1 \quad \varepsilon) \sum_{Q \in B_{P_i}} \frac{r_{j\ 1}(Q)}{Outdgree(Q)} \qquad j = 1,2,3,... \qquad (۲-۳)$$

در رابطه (۲-۳)، رتبه صفحه P به رتبه صفحه Q که به آن اشاره می‌کند، بستگی دارد. این الگوریتم در واقع یک تحلیل احتمالی از قدم زدن تصادفی در گراف وب است و به خوبی صفحه‌های باکیفیت را از صفحه‌های فاقد کیفیت متمایز می‌سازد. Page Rank فرض می‌کند صفحه خوب به صفحه خوب ارجاع می‌دهد. بنابراین صفحاتی که توسط صفحه‌ای خوب مورد ارجاع قرار گرفته‌اند رتبه بالاتری دارند.

رمز موفقیت این الگوریتم به کارگیری اهمیت اسناد به جای در نظر گرفتن مرتبط بودن آن‌ها است. این روش مهم‌ترین مشکل الگوریتم HITS که وابسته بودن به پرس‌وجو است را بر طرف کرده است. بنابراین تعیین لیست مرتب شده اسناد در زمان پرس‌وجو به سرعت انجام می‌شود. از آنجا که ایجاد پیوند از صفحات با اهمیت به صفحه‌ای خاص مشکل است، بر خلاف HITS، در این روش کاربر نمی‌تواند آن را فریب دهد. دلیل دیگر این امر سراسری بودن گراف

[1] Query Independent Schemes

مورد استفاده در Page Rank جهت محاسبه رتبه صفحه می‌باشد. سراسری بودن گراف باعث می‌شود، تغییرات درجه ورودی و خروجی هر گره، تغییر محسوسی در رتبه اسناد ایجاد نکند. اما از آنجا که Page Rank، مستقل از پرس‌وجو می‌باشد، نمی‌تواند بین صفحاتی که در حالت کلی معتبر هستند با صفحاتی که با توجه به موضوع پرس‌وجو، معتبر هستند، تمایز قائل شود. بنابراین لیست نتیجه ممکن است شامل صفحاتی نا مرتبط با پرس‌وجوی کاربر باشد و یا صفحات کم اهمیتی که مرتبط با پرس‌وجو هستند را در بر نگیرد.

با توجه به برخی مشکلات الگوریتم PageRank و به منظور افزایش دقت نتایج تولید شده توسط آن، برخی محققین آن را به روش‌های مختلف بهبود داده‌اند[۱۹][۲۰][۲۱].

۳-۳ روابط تشخیص اجتماع

۳-۳-۱ مرکزیت بینابینی[1]

در شبکه‌ها، هر چه تعداد مسیرهایی که از گره یا لبه خاصی عبور می‌کنند بیشتر باشد، آن گره یا لبه مهم‌تر است. بنابراین با فرض اینکه کوتاه‌ترین مسیر بین دو گره محاسبه پذیر باشد، اهمیت یک گره یا لبه را می‌توان اندازه‌گیری کرد که به آن مرکزیت بینابینی گفته می‌شود[۲۲]. مرکزیت بینابینی از روابطی است که برای تشخیص اجتماع از آن استفاده می‌شود.

مرکزیت بینابینی به صورت زیر تعریف می‌شود:

$$B_u = \sum_{ij} \frac{\sigma(i,u,j)}{\sigma(i,j)} \qquad (۳-۳)$$

که در رابطه (۳-۳)، $\sigma(i,u,j)$ تعداد کوتاه‌ترین مسیرها بین دو اجتماع i و j است که از گره یا لبه u عبور می‌کند. $\sigma(i,j)$ تعداد کل مسیرها بین دو گره i و j است.

[1] Betweenness centrality

اگر تعداد کوتاه‌ترین مسیرها قابل اندازه‌گیری نباشد، به جای آن از الگوریتم جستجو استفاده می‌شود. در این صورت B_u برای یک گره یا لبه تعداد ملاقات‌هایی است که توسط آن الگوریتم جستجو انجام می‌شود. استفاده از الگوریتم جستجو که نوعی تعمیم دادن روش قبلی است، که در [۲۳] ارائه شده است.

۳-۳-۲ برش به‌هنجار شده[1]

این معیار بر این اساس است که یک گروه‌بندی خوب تعداد لبه‌هایی بین اجتماعات را کمینه می‌کند، در عین حال لبه‌های داخل اجتماع را نگه می‌دارد[۲۴]. این معیار به صورت زیر تعریف می‌شود:

گراف $G(V,E)$ را در نظر بگیرید که در آن V مجموعه گره‌های گراف و E مجموعه تمام لبه‌های گراف است. گره‌های گراف را به دو مجموعه مجزای A,B تقسیم می‌کنیم به طوری که $B=V-A$ باشد. مقدار این رابطه کسری از اتصالات بین A,B با توجه به اتصالات جداگانه A,B است.

مقدار cut بین A,B به صورت رابطه (۳-۴) است:

$$cut(A,B)=\sum_{i\in A,\, j\in B} W(i,j) \qquad (۴-۳)$$

و مقدار association به صورت رابطه (۳-۵) است:

$$assoc(A,V)=\sum_{i\in A, v\in V} W(i,v) \qquad (۵-۳)$$

مقدار $W(i,v)$ وزن بین گره v و i است. رابطه آخر برای به‌هنجار کردن اندازه اجتماعات بکار می‌رود. این روش را معمولاً برای گراف‌های وزن‌دار استفاده می‌کنند که البته قابل تعمیم به گراف‌های بدون‌وزن نیز هست(با در نظر گرفتن وزن 1 در صورت وجود لبه و صفر در صورت عدم وجود لبه).

حال از دو رابطه اخیر برای تعریف برش به‌هنجار شده استفاده می‌کنیم.

$$N_{cut}(A,B)=\frac{cut(A,B)}{assoc(A,V)}+\frac{cut(A,B)}{assoc(B,V)} \qquad (۶-۳)$$

همان‌طور در رابطه (۳-۶) می‌بینیم این رابطه فقط شامل دو اجتماع است که برای استفاده در روش‌های تشخیص اجتماع نیاز به تعمیم دارد.

---

[1] Normalized cut

۳-۴ روش‌های تشخیص اجتماع

۳-۴-۱ روش‌های تقسیم کننده

این روش از جمله روش‌هایی است که توپولوژی شبکه را تغییر می‌دهد. در واقع در این روش به دنبال لبه‌هایی هستیم که اجتماعات مختلف را به هم وصل می‌کند و این لبه‌ها را در یک فرم تکرار شونده حذف می‌کنیم تا شبکه به اجتماعات مجزایی از گره‌ها تقسیم شود.

معروف‌ترین روش تقسیم‌کننده روش گیروان-نیومن[1][۲۵] است. چون اجتماعات مختلف توسط تعداد کمی لبه به هم متصل شده، این روش گلوگاه‌هایی که در لبه‌ها قرار دارند و دو اجتماع را به هم متصل می‌کند مورد توجه قرار می-دهد. از طریق این لبه‌ها همه مسیرهای کوتاه عبور می‌کنند. این روش از مرکزیت بینابینی استفاده کرده است.

روش کار بدین صورت است که مرکزیت بینابینی برای تک‌تک لبه‌ها محاسبه می‌شود و لبه‌هایی که بیشترین مقدار را دارند به عنوان لبه‌های بین اجتماعات شناسایی شده و حذف می‌شوند و این کار تا تشخیص کل اجتماعات ادامه می‌یابد.

اگر چه این روش قدرتمندی در تشخیص اجتماع است ولی معایبی دارد که مهم‌ترین آن هزینه بالای محاسباتی آن است. برای بهبود این روش، تایلر[2][۲۶] روشی ارائه کرده که در آن یک عنصر آشوبی (بی نظمی) به این روش اضافه شده است که محاسبه B را به مجموعه جزئی از لبه‌ها محدود می‌کند و از محاسبات آماری برای تخمین B واقعی استفاده می‌کند. پس از این روش، الگوریتم‌های دیگری ارائه شده که از دقت و سرعت بیشتری برخوردارند.

در رویکرد تقسیم‌کننده در بسیاری از روش‌های ارائه شده، تعداد قابل توجهی از لبه‌ها مورد پردازش قرار می‌گیرند. تعداد لبه‌ها در شبکه‌ها زیاد است و محاسباتی که روی آن‌ها قرار می‌گیرد، به طبع آن بالاست. بنابراین روش‌های تقسیم‌کننده از پیچیدگی زمانی و محاسباتی زیادی برخوردار است.

---

[1] Girvan-newman(GN)

[2] Tyler

۳-۴-۲ روش‌های تجمعی

این روش‌ها بر این حقیقت بنا شده که گره‌های یک اجتماع ویژگی‌های مشترکی دارند و می‌توان از این ویژگی‌های مشترک برای گروه‌بندی استفاده کرد. در برابر روش‌های تقسیم‌کننده، روش تجمعی در ابتدا همه گره‌ها را جدا از هم و غیر متصل در نظر می‌گیرد و آن‌ها را بر اساس ویژگی‌های مشترک به هم متصل می‌کند تا به اجتماعات برسد. در واقع نحوه عملکرد این رویکرد بدین صورت است که لبه‌های بین اجتماعات خود به خود حذف شده و تنها لبه‌های داخل اجتماع باقی می‌مانند.

یک گروه مهم از روش‌های تجمعی، خوشه‌بندی بایگان[1] است، که با N گره غیر متصل و بدون لبه شروع می‌شود. لبه‌ها در طول روش بر اساس کاهش شباهت به شبکه اضافه می‌شوند. در ابتدا لبه‌هایی که بیشترین شباهت را دارد، اضافه می‌شود. سپس به مرور سایر گره‌ها مشابه به اجتماع اضافه می‌شوند یکی از روش‌های اندازه‌گیری شباهت، فاصله اقلیدسی است که برای گراف‌های وزن‌دار قابل استفاده است و به صورت رابطه (۳-۷) تعریف می‌شود (در تمام روابط زیر a_{ij} ماتریس مجاورتی است)، شباهت دو گره که با لبه (i, j) در ارتباط است برابر است با:

$$\sqrt{\sum_{k\neq i,j}(a_{ik}-a_{jk})^2} \qquad (۳-۷)$$

و یا از همبستگی پیرسن[2] بین گره‌ها در ماتریس مجاورتی که به صورت رابطه (۳-۸) تعریف شده استفاده می‌شود.

$$\frac{\frac{1}{N}\sum_k (a_{ik}-\mu_i)(a_{jk}-\mu_j)}{\sigma_i \sigma_j} \qquad (۳-۸)$$

که در رابطه (۳-۸) μ_i برابر است با:

$$\mu_i = \frac{1}{N}\sum_j a_{ij} \qquad (۳-۹)$$

$$\sigma_i^2 = \frac{1}{N-1}\sum_j (a_{ij}-\mu_i)^2 \qquad (۳-۱۰)$$

هر چند این روش سریع است ولی گروه‌بندی آن برای شبکه‌های واقعی راضی کننده نیست.

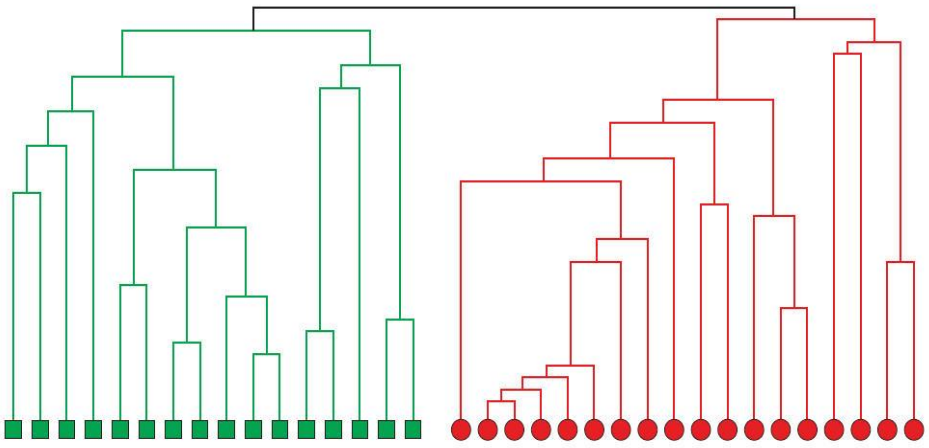
---

[1] Hierarchical clustering

[2] Pearson

در [۲۷] که توسط نیومن ارائه شده روش تجمعی برای تشخیص اجتماع بکار گرفته شده است. در این مقاله از اصول روش گیروان-نیومن استفاده شده، با این تفاوت که زمان اجرای آن کاهش پیدا کرده است. این روش با معیار پودمانی ارزیابی می‌شود که در بخش ۲-۶ توضیح داده شد.

روش کار بدین صورت است که در ابتدا هر گره یک اجتماع را تشکیل می‌دهد. اجتماعات به شرطی که بیشترین افزایش در پودمانی یا Q(یا کم‌ترین کاهش در Q) را داشته باشند به هم می‌پیوندند این فرایند را می‌توان با درختواره نشان داد. شکل ۳-۱ نمایی از درختواره بدست آمده توسط این روش برای باشگاه کاراته است که در بخش ۵-۲-۲ این داده استاندارد مورد بررسی قرار گرفته است.



شکل ۳-۱ درختواره اجتماعات بدست آمده توسط مقاله [۷] برای گراف باشگاه کاراته

تغییرات در Q بر اثر اتصال دو اجتماع به صورت زیر است:

$$\Delta Q = e_{ij} + e_{ji} - 2a_i a_j = 2(e_{ij} - a_i a_j) \qquad (۱۱-۳)$$

این روش بجای محاسبه مکرر و زمان‌بر Q از ΔQ استفاده می‌کند که به طور محلی تغییرات Q را محاسبه می‌کند و باعث کاهش زمان روش است. روش بکار رفته قابل تبدیل به گراف وزن‌دار است و از سری روش‌های حریصانه محسوب می‌شود.

مقاله [۱۲] روش تجمعی بایگان دیگری ارائه کرده است. این روش به نوعی تکمیل کننده این روش می‌باشد. در [۲۷] از ماتریس مجاورتی برای ذخیره کردن گراف استفاده شده در حالیکه [۱۲] به جای ذخیره ماتریس مجاورتی، ماتریس مقادیر ΔQ ذخیره می‌شود. اجتماعاتی که ارتباطی بینشان نیست تغییری در ΔQ آن‌ها ایجاد نمی‌شود. در نتیجه ذخیره نمی‌شوند و این باعث ایجاد یک ماتریس اسپارس می‌شود.

مقاله [۲۸] از ساختارهای تجمعی و بایگان برای تشخیص اجتماع استفاده کرده‌اند. از آنجا که در رویکرد تجمعی به لبه‌ها توجه نداریم، پیچیدگی محاسباتی آن زیاد نیست. با داشتن یک معیار شباهت مناسب می‌توان با هزینه زمانی پایین، اجتماعات را در شبکه تشخیص داد.

۳-۴-۳ بیشینه سازی پودمانی[1]

جستجو برای یافتن پودمانی بهینه (بیشینه) از نوع مسائل بسیار سخت به نظر می‌آید. چرا که فضای تقسیم‌بندی-هایی که از گراف امکان‌پذیر است با بزرگ شدن اندازه گراف به سرعت در حال افزایش است. به همین دلیل رویکردهای جستجوی مکاشفه‌ای برای محدود کردن فضای جستجو الزامی است[۲۹].

نیومن روشی بر پایه ادغام کردن اجتماعات ارائه کرده است به طوری که پودمانی بیشینه شود[۱۲]. در این روش دو اجتماع i و j بر اساس میزان وابستگی‌شان به هم که تغییرات Q در شبکه آن را تعیین می‌کند، می‌پیوندد. این تغییر هنگامی رخ می‌دهد که دو اجتماع با هم یکی می‌شوند.

$$\Delta Q_{ij} = 2(e_{ij} - \frac{\sum_j e_{ij} \sum_i e_{ij}}{2M}) \qquad (۱۲-۳)$$

بنابراین با شروع از هر گره غیر متصل و با در نظر گرفتن آن‌ها به عنوان یک اجتماع، در یک فرآیند تکرارشونده اجتماعات را یکی می‌کنیم، تا جایی که کل گره‌ها در یک اجتماع قرار بگیرند. سپس درختواره حاصل را از مرحله‌ای خاص برش می‌زنیم تا به اجتماعات برسیم.

روش‌هایی که هدف آن بیشینه سازی پودمانی است، معمولاً با رویکردهای تقسیم‌کننده، تجمعی و یا طیفی ترکیب می‌شود و یا از روش‌های مکاشفه‌ای استفاده می‌کند تا به پودمانی بیشینه دست یابد.

---

[1] Modularity maximization

۳-۵ روش‌های طیفی

این روش‌ها بر مبنای تحلیل ماتریس‌های بردار ویژه‌ای است که از شبکه بدست آمده است. در این روش بر روی مقادیر ویژه ماتریس‌های مرتبط با ماتریس مجاورتی کار می‌شود، که می‌تواند ماتریس لاپلاسین (که به آن Kirchhoff نیز گفته می‌شود) باشد[۳۰].

$$L = D - A \qquad (۱۳-۳)$$

در رابطه (۳-۱۳) A ماتریس مجاورتی و D ماتریس قطری از درجه گره‌ها با المان‌های زیر است:

$$d_{ii} = \sum_j a_{ij}, \qquad d_{ij} = 0 \ \ \forall i \neq j \qquad (۱۴-۳)$$

روش خاصی از نوع طیفی به نام روش طیفی دوبخشی[1] وجود دارد، که بر مبنای قطری سازی ماتریس لاپلاسین است. اگر شبکه به c اجتماع مجزا تقسیم شده باشد، L به صورت قطری است و دارای c بردار ویژه هم ارز است که مقادیر ویژه صفر دارند، ولی اگر کاملاً مجزا نباشند، قطری سازی L، یک بردار ویژه با مقدار ویژه صفر و $c-1$ مقدار ویژه غیر صفر دارد.

وقتی در روش طیفی دو بخشی دو اجتماع داریم $(C=2)$، تقسیم‌بندی شبکه با نسبت دادن مقادیر مثبت بردار ویژه به یک اجتماع و مقادیر منفی به اجتماع دیگر بدست می‌آید[۳۱].

روش‌های طیفی، ساده پیاده‌سازی می‌شوند ولی در شبکه‌های واقعی ضعیف عمل می‌کنند و به درستی تشخیص اجتماع انجام نمی‌شود. مقالات مشابه روش‌های مذکور به روش طیفی عمل می‌کنند.

در [۳۲] روشی ارائه شده است که مفهوم پودمانی را به شکل بردار ویژه فرمول‌بندی مجدد می‌کند و آن را ماتریس پودمانی می‌نامد. برای هر زیر گراف g ماتریس پودمانی $B^{(g)}$ درایه‌های زیر را دارد:

$$b_{ij}^{(g)} = a_{ij} - \frac{k_i k_j}{2M} - \delta_{ij} \sum_{u \in N(g)} \left[a_{iu} - \frac{k_i k_u}{2m} \right] \qquad (۱۵-۳)$$

که برای گره‌های j و i در زیر گراف g تعریف شده است. δ_{ij} در صورت وجود لبه بین گره‌های j و i، یک و در غیر این صورت صفر است. در رابطه (۳-۱۵)، M تعداد کل لبه‌های گراف و $N(g)$ مجموعه گره‌های موجود در زیر گراف g است. k_u, k_j, k_i درجات گره‌های i، j و u است.

---

[1] Spectral bisection

بنابراین برای جداسازی شبکه به اجتماعات تشکیل دهنده آن ابتدا باید ماتریس پودمانی ساخته شود و مثبت‌ترین مقدار ویژه و بردار ویژه متناظر با آن تعیین شود. بر اساس علامت درایه‌های این بردار، شبکه به دو قسمت تقسیم می-شود (گره‌های با مقادیر مثبت در یک اجتماع و مقادیر منفی در اجتماع دیگر). این فرایند به طور بازگشتی برای هر اجتماع تکرار می‌شود تا پودمانی کل به مقدار صفر یا منفی برسد. در ادامه این ایده تعریف جدیدی از اجتماع به نام زیر گراف‌های نامرئی ارائه شده یا زیر گراف‌هایی که تقسیم‌بندی‌شان باعث افزایش پودمانی می‌شود. این روش در بین روش‌های ارائه شده دیگر دقیق‌تر است و قادر به یافتن تقسیم‌بندی از گراف است که پودمانی مناسبی را در بسیاری از شبکه‌ها دارد روش مشابهی نیز در [۳۳] ارائه شده است.

در مجموع روش‌های طیفی از روش‌های اولیه است که در تشخیص اجتماع ارائه شده است و محاسبات برداری آن در حجم بالایی است.

۳-۵-۱ بهینه‌سازی اکسترمم

این روش که توسط داچ و ارناز[1] [۳۴] پیشنهاد شده است، جستجوی مبتنی بر تجربه برای بهینه‌سازی مقدار پودمانی است. به طور پیش فرض، شبکه به دو بخش تصادفی تقسیم می‌شود که تعداد گره‌ها در آن‌ها مساویند. در هر مرحله، سیستم با حرکت دادن گره‌هایی با کم‌ترین برازش از یک بخش به بخش دیگر خودسازماندهی می‌کند. این فرآیند هنگامی که پودمانی بیشینه شود متوقف می‌شود. اکنون شبکه دارای دو اجتماع است. حال لبه‌ها بین اجتماعات را حذف می‌کنیم. در اینجا یک اجتماع را برداشته و روند پیشین را برای آن تکرار می‌کنیم تا به دو اجتماع جدید برسیم. فرایند بالا را آن قدر تکرار می‌کنیم تا به پودمانی بیشینه برسیم. اگرچه این روش سریع نیست، ولی می‌تواند مقادیر پودمانی بالایی بدست آورد.

[1] Duch and Arenas

۳-۵-۲ روش‌های مبتنی بر الگوریتم‌های تکاملی

روش‌های مختلفی در حوزه الگوریتم‌های تکاملی بر روی تشخیص اجتماع انجام شده است. تارگین و بینگل[1] [۳۵] روشی ارائه کردند که از پودمانی برای سنجش استفاده می‌کند. در این روش هر کروموزوم کل گره‌های موجود در شبکه را شامل می‌شود. در ابتدا به هر گره به صورت تصادفی یک اجتماع نسبت داده می‌شود و سپس crossover که تغییراتی روی آن صورت گرفته، انجام می‌شود. گاهی نیز mutation که گره‌های دو اجتماع را جابجا می‌کند اعمال می‌شود و در نهایت هر کروموزوم توسط معیار پودمانی سنجیده می‌شود. سپس پودمانی مقادیر همه اجتماعات را محاسبه می‌کند که با به‌کارگیری ساختار بایگان بدست آمده‌اند و به عنوان نتیجه، اجتماعی که بیشترین مقدار پودمانی را دارد، بر می‌گرداند.

مشکل این روش آن است که در گراف‌های بزرگ نمی‌توان از آن استفاده کرد زیرا نیاز به حافظه بسیار بزرگی دارد که بتواند در هر کروموزوم آن کل گره‌های شبکه را جا دهد.

پیزوتی[2] [۳۶] در روش خود که آن را GA-Net نامیده است، از ماتریس مجاورتی استفاده کرده است. در این روش کل گراف با ماتریس نشان داده می‌شود و اجتماعات زیر ماتریس‌هایی از ماتریس اصلی هستند که در ابتدا به صورت تصادفی انتخاب می‌شوند. برای سنجیدن درستی کار خود پیزوتی معیار جدیدی را ارائه کرده است که در آن به اجتماعات با چگالی بیشتر امتیاز بالاتری داده شده است.

روش دیگری که از الگوریتم ژنتیک استفاده کرده، روش ACGA[3] است که توسط لیپزاک و میلیوس[4] [۳۷] ارائه شده است. این روش به حافظه زیادی نیاز ندارد زیرا هر کروموزوم آن تنها بخشی از پاسخ مسئله است و نه تمام آن و کل کروموزوم‌ها به هم تشکیل نتیجه را می‌دهند. این روش را می‌توان به راحتی برای گراف‌های بزرگ نیز استفاده کرد. سایز بزرگ شبکه‌های اجتماعی، روش‌های قدیمی تشخیص اجتماع را با استفاده از الگوریتم‌های تکاملی که در آن کل شبکه در یک کروموزوم قرار می‌گرفت ناکارآمد کرده است. در اینجا روشی با عنوان الگوریتم ژنتیک خوشه‌بندی تجمعی (ACGA) ارائه شده است که در آن هر مرحله تنها بخشی از گراف که شامل دو اجتماع و همسایه‌های آنان

---

[1] Targin and Bingol

[2] Pizzuti

[3] Agglomerative Clustering Genetic Algorithm

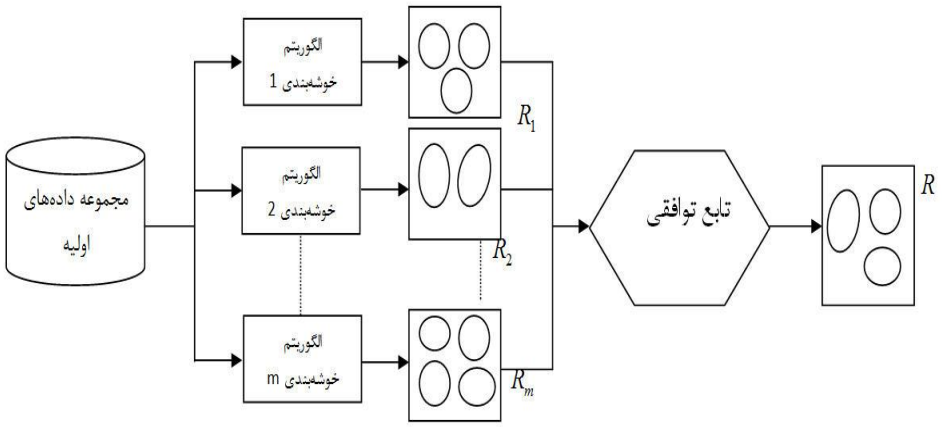
[4] Lipszak and Millios

است، مورد بررسی قرار می‌گیرد. در واقع بررسی گراف به صورت محلی انجام می‌شود. عملکرد برنامه بدین صورت است که ابتدا همه گره‌های گراف را به صورت تصادفی در چند اجتماع قرار می‌دهد و هر اجتماع یک کروموزوم را تشکیل می‌دهد. با این کار کروموزوم‌های نسل اول ساخته می‌شوند. در این روش هر کروموزوم تنها بخشی از راه حل است و همه کروموزوم‌های یک نسل با هم راه حل نهایی را ارائه می‌دهند. مرحله بعد انتخاب والد است. ابتدا یک گره به طور تصادفی انتخاب می‌شود و همسایه‌های آن بدست می‌آیند. همسایه‌هایی که در اجتماع آن گره نیستند، در لیستی قرار داده می‌شوند و از بین آن‌ها یک گره انتخاب می‌شود. حال کروموزوم مربوط به دومین گره که در اجتماع نیست با کروموزوم مربوط به گره اول ترکیب می‌شود. در crossover ابتدا دو اجتماع در کنار هم در یک اجتماع قرار می‌گیرند و سپس گره‌ها به طور تصادفی با هم جابجا می‌شوند. اکنون اجتماع جدید به دو قسمت تقسیم می‌شود و دو فرزند جدید بدست می‌آید. در این مرحله پودمانی کل گراف قبل و بعد از قرار دادن فرزندان محاسبه می‌شود. اگر پودمانی فرزندان بیشتر بود جایگزین والدین می‌شود، در غیر این صورت بدون تغییر باقی می‌ماند و این روند بارها تکرار می‌شود تا به تقسیم‌بندی مناسبی‌تر اجتماعات برسیم. معیار ارزیابی درستی روش، پودمانی است.

بعضی دیگر از روش‌های تشخیص اجتماعات در [۳۸] آمده است.

۳-۶ روش‌های خوشه‌بندی ترکیبی

ایده اصلی خوشه‌بندی اطلاعات، جدا کردن نمونه‌ها از یکدیگر و قرار دادن آن‌ها در اجتماعات شبیه به هم می‌باشد. به این معنی که نمونه‌های شبیه به هم باید در یک اجتماع قرار بگیرند و با نمونه‌های اجتماعات دیگر حداکثر تفاوت را دارا باشند. از آنجا که اکثر روش‌های خوشه‌بندی پایه روی جنبه‌های خاصی از داده‌ها تاکید می‌کنند، در نتیجه روی مجموعه داده‌های خاصی کارآمد می‌باشند[۱۴]. به همین دلیل، نیازمند روش‌هایی هستیم که بتواند با استفاده از ترکیب این الگوریتم‌ها و گرفتن نقاط قوت هر یک، نتایج بهینه‌تری را تولید کند. در واقع هدف اصلی خوشه‌بندی ترکیبی همان‌طوری که در شکل ۳-۲ می‌بینیم، جستجوی نتایج بهتر و مستحکم‌تر، با استفاده از ترکیب اطلاعات و نتایج حاصل از چندین خوشه‌بندی اولیه است.



شکل ۳-۲ فرآیند پایه خوشه‌بندی ترکیبی[۳۹]

تحقیقات اخیر در این زمینه نشان داده‌اند که خوشه‌بندی داده‌ها می‌تواند به طور چشمگیری از ترکیب چندین افراز داده سود ببرد. خوشه‌بندی ترکیبی می‌تواند جواب‌های بهتری از نظر استحکام، نو بودن، پایداری و انعطاف‌پذیری نسبت به روش‌های پایه ارایه دهد. همان‌طوری که در فصل قبلی گفتیم به طور خلاصه خوشه‌بندی ترکیبی شامل دو مرحله اصلی زیر می‌باشد:

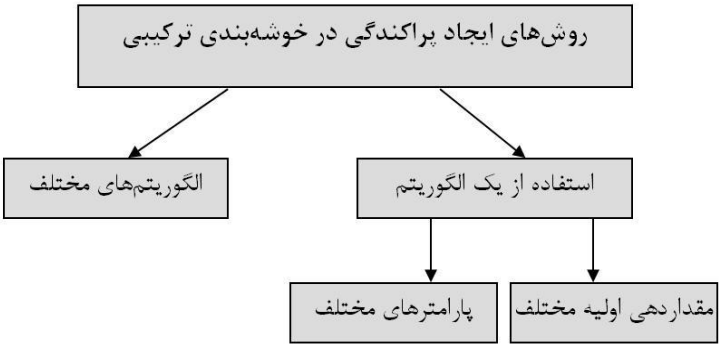
- تولید نتایج متفاوت از خوشه‌بندی‌ها، به عنوان نتایج خوشه‌بندی اولیه بر اساس اعمال روش‌های مختلف که این مرحله را، مرحله ایجاد تنوع یا پراکندگی می‌نامند.
- ترکیب نتایج به دست آمده از خوشه‌بندی‌های متفاوت اولیه برای تولید خوشه نهایی؛ که این کار توسط تابع توافقی (الگوریتم ترکیب کننده) انجام می‌شود.

همان‌طوری که در شکل ۳-۲ نشان داده می‌شود، با اعمال روش‌های مختلف بر روی مجموعه داده‌های اولیه و بدست آوردن نتایج مختلف و نهایت با اعمال یک تابع توافقی بر روی نتایج بدست آمده می‌توان به یک نتیجه بهتر و مستحکم‌تر رسید.

۳-۶-۱ ایجاد پراکندگی در خوشه‌بندی ترکیبی

در مرحله اول تعدادی خوشه‌بندی‌های اولیه که هر کدام بر ویژگی خاصی از داده‌ها تاکید دارند، ایجاد می‌شود. اولین و ساده‌ترین روش برای ایجاد نتایج مختلف و پراکنده از یک مجموعه داده، استفاده از الگوریتم‌های مختلف

خوشه‌بندی است. هر الگوریتم خوشه‌بندی از یک جنبه خاصی به مسئله نگاه می‌کند. بنابراین خطاهای موجود در روش‌های مختلف، می‌تواند با هم متفاوت باشد. این امر می‌تواند موجب ایجاد پراکندگی در نتایج الگوریتم‌های پایه خوشه‌بندی گردد. مهم‌ترین الگوریتم‌های خوشه‌بندی پایه که معمولاً در خوشه‌بندی ترکیبی استفاده می‌شوند شامل الگوریتم‌های خوشه‌بندی بایگان و الگوریتم‌های خوشه‌بندی افرازبندی[1] می‌باشند[۴۸].



شکل ۳-۳ طبقه‌بندی روش‌های ایجاد پراکندگی در خوشه‌بندی ترکیبی[۴۸]

همان‌طوری که در شکل ۳-۳ مشاهده می‌شود، روش‌های موجود به طور کلی از دو روش برای تولید نتایج متفاوت استفاده می‌نمایند که عبارتند از:

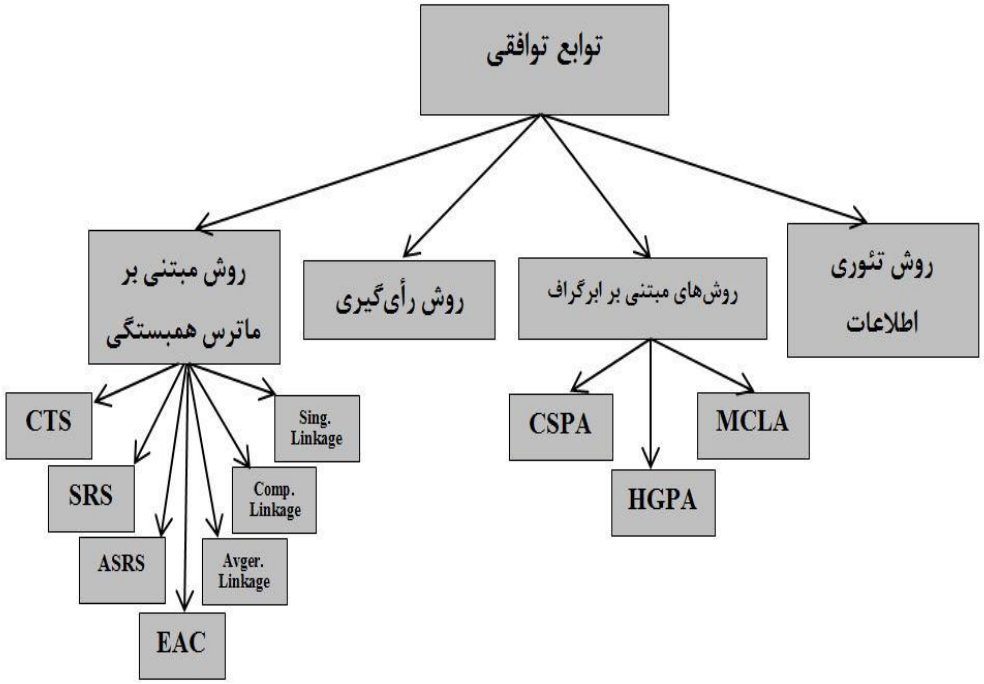
- استفاده از یک الگوریتم: در این روش همان‌طوری که در شکل سعی می‌کنیم با مقداردهی اولیه متفاوت و یا با تغییر پارامترهای مختلف در الگوریتم مثلاً پارامتر اتمام الگوریتم و مواردی از این قبیل نتایج متفاوتی را ایجاد نماییم.
- استفاده از الگوریتم‌های مختلف: در این روش نیز همان‌طوری که از نام آن مشخص است روش‌های گوناگون را بر روی داده خود اعمال می‌نماییم تا نتایج متفاوتی را ایجاد نماییم.

۳-۶-۲ تابع توافقی

پس از اینکه نتایج اولیه (تا حد ممکن پراکنده) تولید شد، معمولاً با استفاده از یک تابع ترکیب کننده این نتایج ترکیب می‌شوند. یکی از متداول‌ترین روش‌های ترکیب نتایج استفاده از ماتریس همبستگی می‌باشد. این روش که

[1] Partitional

مبتنی بر ماتریس همبستگی است، اولین بار توسط فرد و جین[1] [۱۰] مطرح شد و خیلی زود به صورت یک روش متداول درآمده است. امروزه روش‌های دیگری نیز مبتنی بر ماتریس همبستگی ارائه شده است[۴۰]. در [۴۱] یک طبقه‌بندی کلی برای روش‌های مختلف خوشه‌بندی ترکیبی به صورت شکل ۳-۴ برای توابع توافقی ارائه شده است.



شکل ۳-۴طبقه‌بندی توابع توافقی در خوشه‌بندی ترکیبی[۴۱]

در [۴۲] نیز یک روش بر مبنای تابع توافقی برای خوشه‌بندی ارائه شده است. روش‌های مختلفی برای ترکیب خوشه‌بندی‌های اولیه و به دست آوردن خوشه‌های نهایی وجود دارد. در این پایان‌نامه روش‌های مبتنی بر ماتریس همبستگی و روش‌های مبتنی بر ابرگراف را بررسی می‌کنیم.

۳-۶-۲-۱ روش مبتنی بر انباشت مدارک

در [۱۰] یک روش برای ترکیب خوشه‌های ایجاد شده از مرحله یکم برای بدست آوردن خوشه‌های نهایی ارائه شده است. ایده اصلی در این روش که انباشت مدارک[2] نام دارد، بدین صورت می‌باشد که شباهت بین نتایج مختلف بدست

[1] Fred & Jain

[2] Evidence Accumulation Clustering

آمده در مرحله اول از خوشه‌بندی ترکیبی را مورد بررسی قرار می‌دهد. روش کار بدین صورت است که نتایج m الگوریتم خوشه‌بندی توسط تابع توافقی بررسی و در یک ماتریس همبستگی $n \times n$ ذخیره می‌شود. تابع توافقی بدین صورت تعریف می‌گردد و ماتریس همبستگی را می‌سازد که:

$$C(i, j) = \frac{n_{i,j}}{m} \qquad (۱۶-۳)$$

که $n_{i,j}$ تعداد دفعاتی است که جفت نمونه‌های i و j با هم در یک خوشه گروه‌بندی شده‌اند و m تعداد الگوریتم‌های استفاده شده یا نتایج m الگوریتم خوشه‌بندی می‌باشد. پس از ساخت ماتریس همبستگی، می‌توان با استفاده از یکی از الگوریتم‌های بایگان نظیر اتصال منفرد یا اتصال میانگین[۱]، خوشه‌های نهایی را استخراج کرد. این نکته قابل یادآوری است که در اینجا تنها یک ماتریس استخراج می‌شود و خوشه‌های نهایی نیز به سادگی با اعمال یکی از الگوریتم‌های بایگان از این ماتریس به دست می‌آیند.

۳-۶-۲-۲ روش‌های مبتنی بر ابرگراف

در [۱۴] نیز یک روش‌های دیگری برای ترکیب نتایج خوشه‌های ارائه شده است که مبتنی بر ابرگراف می‌باشد که از این الگوریتم‌ها می‌توان به MCLA[۲]، HGPA[۳] و CSPA[۴] اشاره نمود. در هر سه الگوریتم یک ابرگراف ایجاد می‌شود که ارتباطات بین خوشه‌ها در روش‌های مختلف خوشه‌بندی ترکیبی را نشان می‌دهد. در این ابرگراف ایجاد شده هر گره با یک خوشه در خوشه‌های ترکیبی(بدست آمده از خوشه‌بندی پایه) رابطه دارد و هر لبه وزن بین هر دو گره‌ی خوشه‌هایی که با هم در ارتباط هستند را با استفاده از معیار خاصی حساب می‌نماید. در نهایت هر گره در ابرگراف ایجاد شده، درجه وابستگی مشخصی را با هم خواهند داشت. مقدار آن نیز از تعداد روش‌های خوشه‌بندی پایه بدست می‌آید. خوشه‌های نهایی با توجه به وزن موجود بین لبه‌های ابرگراف تولید می‌شود. کدهای این روش‌ها نیز در سایت نویسندگان مقاله وجود دارد و برای تشخیص اجتماعات ترکیبی نیز از همان کدها استفاده شده است. البته یکسری تغییر در جهت قابل استفاده شدن این روش‌های برای تشخیص اجتماعات نیاز بوده است.

---

[1] Single Linkage(SL) or Average Linkage(AL)

[2] Meta CLustering Algorithm

[3] Hyper Graph Partitioning Algorithm

[4] Cluster based Similarity Partitioning Algorithm

CSPA: در این روش برای هر نتیجه خوشه‌بندی پایه می‌بایست یک ماتریس مشابهت $n \times n$ در نظر گرفته شود بدین صورت که اگر دو گره در یک خوشه باشند مقدار یک و در غیر این صورت مقدار صفر را قرار می‌دهیم. در نهایت مجموع مقادیر تمام آن‌ها یک ماتریس $n \times n$ می‌باشد که این ماتریس در اصل همان ابرگراف ما است که n تعداد گره‌های این گراف و وزن هر لبه آن نیز برابر با مقادیر متناظر با ماتریس آن خواهد بود و با اعمال روش‌های خوشه-بندی بر روی گراف می‌توان خوشه‌های نهایی را بدست آورد.

HGPA: این روش یک روش مستقیم برای بدست آوردن خوشه‌بندی ترکیبی با دوباره تقسیم‌بندی نمودن گره‌ها با استفاده از خوشه‌بندی‌های اولیه می‌باشد. در اینجا مسئله خوشه‌بندی ترکیبی به مسئله تقسیم‌بندی ابرگراف به برش لبه‌ها به کمترین مقدار فاصله یا بیشترین مقدار شباهت خود تبدیل می‌گردد. در این روش همه گره‌ها و لبه‌ها وزن برابری دارند و فقط شباهت بین خوشه‌ها مد نظر می‌باشد. بدین صورت که هر روش خوشه‌بندی پایه بر روی گراف اعمال می‌گردد و در نهایت هر گره‌ای که شباهت بیشتری داشته باشد و در خوشه‌بندی‌های مختلف با گره‌های دیگر هم خوشه باشد به عنوان خوشه‌بندی نهایی معرفی می‌گردد.

MCLA: این روش بر پایه خوشه‌بندی خوشه‌ها است. ایده اصلی در این روش گروه‌بندی لبه‌ها در یک خوشه به طوری که بیشترین ارتباط را با هم داشته باشند و در غیر این صورت خارج نمودن آن از گروه‌بندی است. در ابتدا یک گراف وزن‌دار از تمام خوشه‌بندی‌های اولیه بدست می‌آید. این گراف وزن‌دار فاصله بین دو گره را محاسبه می‌نماید. حال با اجرای یک الگوریتم خوشه‌بندی گراف خوشه‌های مختلف جدا می‌شود. در این مرحله اگر خوشه‌های جدا شده از میانگین وزن‌های گره‌های در ارتباط با هم کمتر باشد از خوشه‌بندی حذف شده و به خوشه دیگری که شباهت بیشتری دارد وصل می‌گردد. در نهایت خوشه‌بندی نهایی برای ما به ارمغان خواهد آمد.

۳-۶-۲-۳ روش‌های مبتنی بر شباهت پیوند

این روش در اصل همان روش مبتنی بر ماتریس همبستگی می‌باشد، ولی در جهت افزایش کارایی آن روش‌ها در [۳۹] سه روش مختلف برای بدست آوردن ماتریس همبستگی ارائه شده است، که از این الگوریتم‌ها می‌توان به CTS[۱]، SRS[۲] و ASRS[۳] اشاره نمود. در هر سه الگوریتم ارائه شده یک ماتریس همبستگی متفاوت ایجاد می‌شود که هر یک از این ماتریس‌ها شباهت‌های میان روش‌های خوشه‌بندی اولیه را نشان می‌دهد. کدهای این روش‌ها نیز در سایت نویسندگان مقاله وجود دارد و برای تشخیص اجتماعات ترکیبی نیز از همان کدها استفاده شده است. البته یکسری تغییر در جهت قابل استفاده شدن این روش‌های برای تشخیص اجتماعات نیاز بوده است.

CTS: ایده این روش بر پایه تخمین تعداد مثلث‌های مرتبط با هم می‌باشد. در این روش با محاسبه تعداد مثلث-های ایجاد شده بین لبه‌های دو خوشه که با هم در ارتباط هستند به ایجاد ماتریس همبستگی می‌پردازد. بدین صورت که تعداد مثلث‌های مرتبط با هم در بین دو خوشه به مجموع تعداد مثلث‌های ممکن، ماتریس مشابهت را تشکیل می-دهد.

$$SimCT(i,j) = \frac{CT_{ij}}{CT_{\max}} \qquad (۱۷-۳)$$

همان‌طوری که در رابطه (۳-۱۷) مشاهده می‌شود CT_{ij} تعداد مثلث‌های مرتبط با هم بین دو خوشه i و j می-باشد و $CT_{\max}$ تعداد کل مثلث‌های مرتبط با هم است.

SRS: ایده این روش بر این است که همسایه‌های در یک ماتریس شبیه به هم خواهند بود اگر همسایه‌های آن‌ها به اندازه کافی شبیه به هم باشند. به عبارت دیگر شباهت بین دو گره با استفاده از لبه‌هایی که به طور مستقیم به این دو گره متصل هستند بدست می‌آید. با محاسبه تعداد لبه‌های متصل به دو گره مختلف می‌توان شباهت بین آن‌ها در خوشه‌بندی‌های مختلف در یک ماتریس همبستگی بدست آورد.

---

[1] Connected Triple based Similarity

[2] SimRank based Similarity

[3] Approximate SimRank based Similarity

ASRS: این روش نیز همانند روش SRS می‌باشد با این تفاوت که در SRS شباهت بین دو گره در تمام روش-های خوشه‌بندی مورد بررسی قرار می‌گیرد ولی در این روش تقریبی از تمامی روش‌های خوشه‌بندی برای بدست آوردن ماتریس همبستگی نهایی استفاده می‌شود.

در [۴۲] روشی ترکیبی برای تشخیص اجتماعات ارائه شده است که با استفاده از ماتریس همبستگی نتایج دقیق و پایداری را برای تشخیص اجتماعات در شبکه‌های پویا بدست آورده است. این روش برای شبکه‌های پویا مورد استفاده قرار می‌گیرد و با اعمال یک الگوریتم بر روی شبکه‌های پویا که در هر لحظه در حال تغییر هستند نتایج متفاوتی را ایجاد می‌نماید. با ترکیب این نتایج تنها با توجه به هم خوشه بودن نتایج مختلف بدست آمده درصدد بدست آوردن نتایج دقیق و پایدار هستند، در این روش ما نیازمند دانستن تعداد اجتماعات هستیم. معیار ارزیابی نیز در این روش اطلاعات متقابل نرمال شده[1] می‌باشد که نشان‌دهنده دقت نسب به روش‌های دیگر است.

در این پایان‌نامه سعی داریم که با استفاده از هفت روش مختلف خوشه‌بندی ترکیبی که در بخش ۳-۶-۲ به آن‌ها اشاره شد، استفاده نماییم تا تشخیص اجتماعات را بدون دانستن تعداد اجتماعات با دقتی نزدیک به نتیجه بهینه بدست آوریم. معیار ارزیابی در روش پیشنهادی پودمانی است که در بخش ۲-۶ به آن اشاره شده است.

۳-۷ خلاصه‌ی این فصل

در این فصل کارهای انجام شده قبلی را مورد بررسی قرار دادیم. در این فصل به کارهای انجام شده بر روی تشخیص اجتماعات، وب‌کاوی، خوشه‌بندی، خوشه‌بندی ترکیبی و مواردی از این قبیل پرداختیم. همان‌طوری که مشاهده شد، روش‌های مختلفی برای تشخیص اجتماعات ارائه شده است. با توجه به شباهت‌های خوشه‌بندی و تشخیص اجتماعات در این پایان‌نامه قصد داریم که از روش‌های خوشه‌بندی ترکیبی برای تشخیص اجتماعات استفاده نماییم. در فصل بعدی چگونگی ترکیب روش‌های تشخیص اجتماعات را بیان خواهیم نمود و در نهایت یک روش تشخیص اجتماعات ترکیبی را ارائه می‌دهیم. این روش امکان تشخیص دقیق‌تر، مطمئن‌تر و مستحکم‌تر اجتماعات را به همراه می‌آورد. در فصل پنجم نیز با استفاده از این معیار پودمانی به ارزیابی روش پیشنهادی خواهیم پرداخت.

---

[1] Normalized Mutual Information (NMI)

فصل چهارم : روش پیشنهادی تشخیص اجتماعات

۴-۱ مقدمه

در میان کاربردهای داده‌کاوی ساختار وب تشخیص اجتماعات در وب از این جهت که می‌تواند به کاربران در بازیابی اطلاعات از وب کمک کند، اهمیت ویژه‌ای دارد. علاوه بر اجتماعاتی که صریحاً در وب تعریف شده‌اند، مانند گروه‌های خبری، اجتماعات دیگری نیز به طور ضمنی در وب وجود دارند که حتی اعضای آن ممکن است از وجود آن بی اطلاع باشند. تحقیقات اخیر نشان می‌دهند که تعداد زیادی اجتماعات در وب وجود دارد. با تشخیص یک اجتماع در وب درباره یک موضوع خاص، کاربران می‌توانند با استفاده از صفحات اجتماع، اطلاعات مفیدی درباره آن موضوع به دست آورند. اجتماعات که در وب وجود دارد تا حدودی با اجتماعات واقعی متفاوت هستند. هدف نیز در این پایان‌نامه ارائه یک روش ترکیبی جهت پیدا نمودن اجتماعات دقیق در وب است.

در فصل پیش به بررسی انواع روش‌های تشخیص اجتماع و روش‌های خوشه‌بندی ترکیبی پرداختیم. همان‌طوری که در فصل قبلی بررسی شد، روش‌های خوشه‌بندی ترکیبی با ترکیب روش‌های خوشه‌بندی پایه‌ای نتایج بهتر و مستحکم‌تری را ایجاد می‌نمایند. با الهام از روش‌های خوشه‌بندی ترکیبی و استفاده آن در روش‌های تشخیص اجتماعات می‌توان به نتایج دقیق‌تر و مطمئن‌تری در این زمینه نیز دست پیدا نمود.

همان‌طور که در بخش ۱-۱ اشاره شد، مراجع مختلفی تشخیص اجتماعات را نوعی خوشه‌بندی می‌داند که بر روی گراف‌ها قابل اعمال می‌باشد و در این مراجع خوشه‌بندی و تشخیص اجتماعات را به صورت جایگزین برای هم مورد استفاده قرار می‌دهند. اکثر روش‌های خوشه‌بندی یا تشخیص اجتماعات پایه روی جنبه‌های خاصی از داده‌ها تاکید می‌کنند، برای مثال یک الگوریتم روی داده‌ها کوچک جواب خوبی را دارد ولی این الگوریتم برای حجم داده‌های بسیار بالا کارآمد نخواهد بود، و یا برعکس. داده‌های موجود در شبکه‌های اجتماعی نیز بسیار متنوع می‌باشند، از حجم بسیار پایین شروع می‌شوند و در بسیاری موارد به حجم داده‌ای برابر با کل شبکه جهانی می‌رسند. این تنوع در حجم داده یک الگوریتم پیشنهادی برای استفاده در حجم متنوع را ناکارآمد خواهد نمود. در نتیجه اکثر الگوریتم‌ها در تشخیص اجتماعات روی مجموعه داده‌های خاصی کارآمد می‌باشد بنابراین نیازمند به روش‌هایی هستیم که بر روی تمام داده‌ها نتایج خوبی را داشته باشند. چرا که اجتماعات بدست آمده می‌تواند در اهداف بلند مدت سازمانی نیز مورد استفاده قرار گیرد و این مسئله در دراز مدت می‌تواند مشکلات عدیده‌ای را برای سازمان ایجاد نماید. به عنوان مثال از تشخیص اجتماعات می‌توان در مسائلی از قبیل جاسوسی، تبلیغات، بهبود موتورهای جستجو، سیستم‌های پیشنهاد دهنده و ... استفاده نمود. بنابراین نوعی حساسیت خاص بر حسب شرایط استفاده از این روش وجود دارد، که در یک شرایط دقت

تشخیص اجتماعات برای ما از اهمیت خاصی برخوردار است. با ترکیب روش‌های مختلف تشخیص اجتماعات و با بررسی تک‌تک نتایج بدست آمده از روش‌های مختلف درصدد پیدا نمودن بهترین گروه‌بندی برای داده‌های ورودی هستیم. خوشه‌بندی ترکیبی یک روش جدید و کارا می‌باشد که می‌توان مشکل تشخیص اجتماعات را نیز توسط آن تا حدود زیادی برطرف نمود. در این فصل یک روش جدید برای پیدا نمودن اجتماعات دقیق‌تر پیشنهاد شده است که با نام خوشه‌بندی ترکیبی به آن اشاره شده است، به طوری که این روش برای تمام مجموعه داده‌ها به صورت گراف قابل اعمال می‌باشد. البته در این پایان‌نامه صرفاً گراف بدون جهت مورد بررسی قرار می‌گیرد، که به تشریح روش پیشنهادی می‌پردازیم.

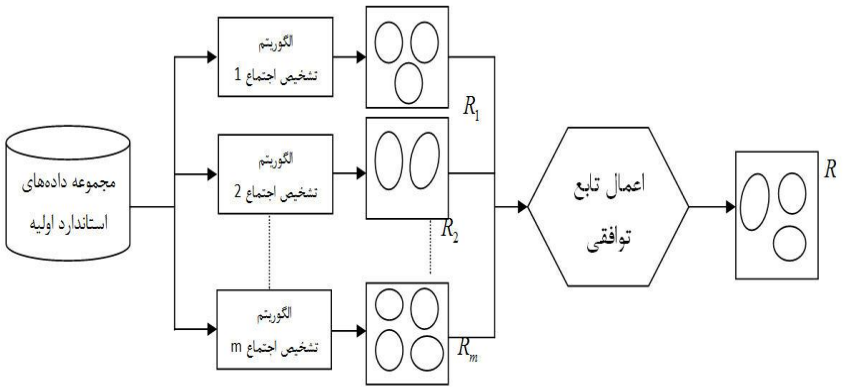
۴-۲ روش پیشنهادی برای تشخیص اجتماعات

روش‌های تشخیص اجتماعات که تاکنون گزارش شده است، تنها از یک روش و یا با ترکیب دو روش مختلف با هم به دنبال پیدا نمودن اجتماعات می‌باشد. ولی استفاده از یک یا ترکیب دو روش، روی مجموعه خاصی از داده‌های کارآمد می‌باشد بنابراین نیازمند به روش‌هایی هستیم که بر روی تمام داده‌ها نتایج خوبی را به همراه باشد. در این بخش روشی مبتنی بر خوشه‌بندی ترکیبی با استفاده از ترکیب نتایج روش‌های تشخیص اجتماعات پایه پیشنهاد می‌گردد. از آن جا که این روش پیشنهادی با ترکیب نتایج روش‌های مختلف بدست می‌آید، نتایج مثبت هر الگوریتم را می‌تواند با هم ترکیب نماید و در نهایت نتایج دقیق‌تر و مؤثرتری را ارائه نماید. همان‌طوری که در شکل ۴-۱ مشاهده می‌شود، در این پایان‌نامه از خوشه‌بندی ترکیبی برای بدست آوردن اجتماعات دقیق‌تر برای الگوریتم‌های تشخیص اجتماع استفاده می‌شود. نام روش پیشنهادی تشخیص اجتماعات ترکیبی[1] است، مراحل این روش دقیقاً همانند روش خوشه‌بندی ترکیبی است، با این تفاوت که در اینجا از روش‌های تشخیص اجتماع بجای خوشه‌بندی برای این کار استفاده شده است.

تشخیص اجتماعات ترکیبی با بهره‌گیری از روش‌های خوشه‌بندی ترکیبی به ترکیب روش‌های مختلف می‌پردازد، البته محدودیتی برای استفاده از الگوریتم‌ها و روش‌های خاص در خوشه‌بندی ترکیبی وجود ندارد و امکان استفاده از هر روشی در آن وجود دارد. حتی اگر یک یا چند روش نتایج نسبتاً بد یا خیلی بدی را به همراه داشته باشند در نتیجه

[1] Ensemble Community Detection(ECD)

نهایی تفاوت زیادی ایجاد نمی‌شود و باز هم خوشه‌بندی ترکیبی توانایی پیدا نمودن اجتماعات مناسب را خواهد داشت. هر چه تعداد الگوریتم‌های استفاده شده در تشخیص اجتماعات ترکیبی بیش‌تر و پراکنده‌تر باشد، نتیجه ترکیبی بهتر خواهد بود[۱۴]. پراکندگی به معنای وجود نتایج متفاوت است، به عبارت دیگر هر روش نتایج مختلفی را به همراه داشته باشد و این نتایج با یکدیگر فرق داشته باشند. چرا که تابع توافقی در تشخیص اجتماعات ترکیبی بهتر می‌تواند ایفای نقش نماید و با بررسی تک‌تک نتایج یک ماتریس همبستگی مناسبی را ایجاد می‌نماید که امکان تشخیص اجتماعات در آن با استفاده از روش‌های گوناگون دقیق‌تر خواهد بود. همان‌طوری که در شکل ۴-۱ مشاهده می‌شود امکان استفاده از هر نوع داده به صورت گراف برای روش پیشنهادی وجود دارد. بعد از تبدیل آن به ماتریس مجاورتی متناظر با آن گراف، ماتریس مجاورتی مربوطه را به روش‌های تشخیص اجتماعات می‌دهیم تا هر یک نظر خود را در مورد تشخیص اجتماعات ارائه نمایند. حال نوبت تابع توافقی روش تشخیص اجتماعات ترکیبی است که با بررسی نتایج که در قسمت بعدی توضیح داده می‌شود، یکسری اجتماعات نهایی را به عنوان جواب نهایی نشان دهد.



شکل ۴-۱ روش پیشنهادی تشخیص اجتماعات ترکیبی

بنابراین روش پیشنهادی همانند شکل ۴-۱ برای تشخیص اجتماعات، از دو مرحله کلی تشکیل شده است. در مرحله اول، با استفاده از روش‌های پایه و یا هر روش دیگر در زمینه تشخیص اجتماعات نتایج اولیه روش‌های مختلف را ایجاد می‌نماییم. برای بدست آوردن این نتایج نیز همان‌طوری که در بخش ۳-۶ به آن اشاره شد، از دو روش استفاده می‌نماییم. روش اول که با استفاده از تغییر مقادیر اولیه متفاوت و یا با تغییر پارامترهای روش تشخیص اجتماع نتایج گوناگونی را از یک الگوریتم دریافت می‌نماییم و روش دیگر با استفاده از الگوریتم‌های مختلف تشخیص اجتماعات نتایج لازم را برای انجام تحلیل‌ها و بررسی‌ها توسط تابع توافقی را ایجاد می‌کنیم. در این پایان‌نامه نیز چند روش مختلف از توابع توافقی را برای تشخیص دقیق‌تر اجتماعات مورد بررسی قرار می‌دهیم.

یکی از ویژگی‌های روش پیشنهادی آن است که وابسته به ساختار خاصی از گراف برای تشخیص اجتماع نمی‌باشد و تقریباً بر روی تمام ساختارها به صورت گرافی بدون جهت در این پایان‌نامه و جهت‌دار جواب می‌دهد، امکان استفاده از گراف‌های جهت‌دار نیز برای این روش پیشنهادی وجود دارد که موضوع بحث ما نخواهد بود. در حالی که برخی از روش‌های تشخیص اجتماع ساختار خاصی از گراف مانند گراف کامل دوبخشی را جستجو می‌کنند و بخش‌هایی از گراف وب را که شامل این ساختار می‌باشند به عنوان اجتماع معرفی می‌کنند. ویژگی دیگر مقیاس‌پذیری بالای این روش است که امکان استفاده از هر روش تشخیص اجتماعات وجود دارد، هر چه تعداد روش‌ها زیاد باشد تشخیص اجتماعات ترکیبی نتیجه مناسب‌تری را ارائه می‌کند. ویژگی دیگر پراکندگی در نتایج اولیه تشخیص اجتماع و نتایج نسبتاً بد در روش‌های اولیه است که هر چه پراکندگی نتایج اولیه بدست آمده بیشتر باشد، در نهایت دقت نتایج در خوشه‌بندی ترکیبی نهایی بهتر خواهد بود.

ویژگی دیگر روش پیشنهادی آن است که در این روش نیازی به دانستن تعداد اجتماعات در گراف اولیه نمی‌باشد. بلکه در بسیاری از روش‌های تشخیص اجتماعات و یا خوشه‌بندی نیازمند به تعداد خوشه‌ها و یا اجتماعات هستیم. این ضعف اصلی روش‌های خوشه‌بندی با روش‌های تشخیص اجتماعات می‌باشد، در روش‌های خوشه‌بندی و یا خوشه‌بندی ترکیبی معمولاً تعداد خوشه‌ها می‌بایست مشخص باشد ولی در شبکه‌های اجتماعی تعداد اجتماعات را نداریم و هیچ‌گونه ذهنیت خاصی راجع به آن وجود ندارد. در اینجا یک روش نیز برای پیدا نمودن تعداد اجتماعات در شبکه‌های اجتماعی و تشخیص اجتماعات ارائه شده است که می‌توان از آن برای روش‌های خوشه‌بندی ترکیبی بر روی ماتریس همبستگی و مواردی از این قبیل استفاده نمود.

در شبه کد شکل ۴-۲ روش پیشنهادی تشخیص اجتماعات ترکیبی را نشان می‌دهد.

```
Input: A(Adjacency matrix from Graph)
N(The Number of Community)= ComputeN
   for Number of community detection method do
            while (Can change initialization parameter) do
                    ECD+:=run(community detection method)
            end while
            if(Can change final state community detection method) then
                    ECD+:=run(community detection method)
            end if
    end for
ECD: Ensemble members of Community Detection methods
/*Each column show one result of community detection method*/
    Co.assosiation matrix:=run(Consensus function)
    C:=run(Hierarchical clustering method)
Output:C(Final community detection result)
```

شکل ۴-۲ شبه کد روش تشخیص اجتماعات ترکیبی

در این شبه کد ماتریس مجاورتی حاصل از گراف مربوط به شبکه‌های اجتماعی مورد بررسی را به عنوان ورودی گرفته و تعدادی اجتماع را به عنوان نتیجه نهایی نشان می‌دهد.

برای اجرای شبه کد فوق می‌بایست ماتریس مجاورت گرافی را که می‌خواهیم مورد بررسی قرار دهیم را بدست آوریم. از آنجایی که در داده‌های استاندارد تست شده برای ارزیابی این روش ماتریس مجاورت وجود دارد، نحوه بدست آمده آن در کد بالا نوشته نشده است. ولی به سادگی می‌توان از روی هر گراف دلخواه ماتریس مجاورت متناظر با آن را بدست آورد. با داشتن ماتریس مجاورت و با اعمال یک تابع با نام ComputeN می‌توان تعداد اجتماعات را بدست آورد. حال به تعداد روش‌هایی که برای تشخیص اجتماعات با تغییر مقادیر اولیه روش‌های تشخیص اجتماع و یا با تغییر پارامترهای روش تشخیص اجتماع به دنبال ایجاد یک ماتریس از نتایج روش‌های مختلف تشخیص اجتماعات هستیم. هرچه نتایج بیشتر باشد روش ما بهتر جواب می‌دهد، بعد از ایجاد این ماتریس که سطر آن به تعداد گره‌های شبکه مورد بررسی می‌باشد و ستون‌های آن به تعداد نتایج حاصله از روش‌های تشخیص اجتماعات مختلف است. حال نوبت اجرای تابع توافقی و بدست آمدن ماتریس همبستگی می‌باشد که در بخش ۴-۳ به بررسی تابع توافقی می‌پردازیم. بعد از بدست آمدن ماتریس همبستگی نهایی حال روش‌های خوشه‌بندی بایگان را بروی این ماتریس اعمال می‌نماییم تا نتایج نهایی که همان اجتماعات مختلف می‌باشد بدست آید.

در شکل ۴-۳ شبه کد روش پیشنهادی برای تشخیص تعداد اجتماعات را نشان می‌دهد، روش کار تابع ComputeN نیز بدین شرح می‌باشد که:

```
Input: A(Adjacency matrix from Graph)
for (Number of base community detection method) do
        NECD+:=NumberOfCommunity(run(community detection method))
end for
NECD: Maximum number of Community in Ensemble community Detection methods
   NECD:=Refine(NECD)
   N:=Avg(NECD)
Output:N(Final community number)
```

شکل ۴-۳ شبه کد تابع تشخیص تعداد اجتماعات

تابع فوق ماتریس مجاورتی را گرفته و تعدادی روش تشخیص اجتماعات که در داده‌های مختلف دارای نتایج نسبتاً خوب می‌باشند را اجرا می‌کند، بر خلاف روش پیشنهادی در تشخیص اجتماعات ترکیبی در این تابع می‌بایست روش-های خاصی از تشخیص اجتماعات اجرا گردد و بعد از اجرای روش‌های مختلف میانگینی از نتایج بدست آمده به عنوان تعداد اجتماعات ارائه می‌شود. بعد از بدست آوردن تعداد اجتماعات می‌توان در صورت نیاز به N از آن به عنوان ورودی استفاده نمود. به عنوان مثال مقدار N برای اجرای روش‌های خوشه‌بندی بایگان الزامی است. از مزایای این روش عدم

نیاز به تعداد اجتماعات است، چرا که در داده‌های وب برخلاف روش‌های خوشه‌بندی هیچ‌گونه اطلاعی در مورد تعداد اجتماعات وجود ندارد، بنابراین به روش‌هایی نیازمند خواهیم بود که بتوان بدون داشتن تعداد اجتماعات به بررسی آن-ها بپردازیم.

در روش تشخیص اجتماعات ترکیبی امکان استفاده از هر روش تشخیص اجتماع امکان‌پذیر خواهد بود، حتی اگر الگوریتم که مورد استفاده قرار می‌گیرد نتیجه بدی را به همراه داشته باشد. چرا که در نهایت بعد از ایجاد ماتریس همبستگی از بررسی تک‌تک نتایج به هر لبه گراف یک عددی نسبت داده می‌شود، که این عدد بیانگر مقدار وابستگی آن گره‌ها در لبه مد نظر است. هر چه تعداد بیشتری الگوریتم بر روی آن لبه اتفاق نظر(اجماع) داشته باشند، احتمال صحیح بودن آن لبه در اجتماع مدنظر بیشتر خواهد بود و بعد از بررسی همه نتایج و ایجاد ماتریس همبستگی و اجرای الگوریتم‌های بایگان گوناگون بر روی ماتریس همبستگی می‌توان نتایج دقیق‌تر و مستحکم‌تری را ایجاد نمود.

ولی در روش پیشنهادی برای پیدا نمودن تعداد اجتماعات می‌بایست روش‌هایی را اجرا نماییم که تقریباً جواب خوبی را به همراه دارند، چرا که اگر روشی نتایج اشتباهی داشته باشد و نتایج آن با روش‌های دیگر تفاوت‌های زیادی داشته باشد تعداد اجتماع بدست آمده با اجتماع حقیقی متفاوت خواهد داشت. بنابراین برای رفع این مشکل از تابع Refine استفاده می‌نماییم، بدین صورت که بعد از بدست آمدن تک‌تک اجتماعات توسط روش‌های گوناگون یک پالایش بر روی تعداد اجتماعات در روش‌های مختلف انجام می‌دهیم. بدین صورت که اگر تعداد اجتماعات بدست آمده تقریباً یکسان باشند همه آن‌ها برای گرفتن میانگین شرکت می‌کنند، ولی اگر در یک یا چند روش مورد استفاده تعداد اجتماعات با روش‌های دیگر متفاوت باشد، آن‌ها را قبل از میانگین گیری حذف می‌نماییم.

۳-۴ توابع توافقی روش پیشنهادی

روش‌های مختلفی برای ترکیب خوشه‌های نهایی وجود دارد، به عبارت دیگر برای تعریف توابع توافقی روش‌های مختلفی وجود دارد که در تشخیص اجتماعات ترکیبی از سه روش گوناگون استفاده کرده‌ایم. این سه روش در اصل در دو طبقه‌بندی توابع توافقی قرار دارند که عبارتند از: روش‌های مبتنی بر ماتریس همبستگی و روش‌های مبتنی بر ابرگراف است.

۴-۳-۱ تابع توافقی مبتنی بر انباشت مدارک

در [۱۰] یک روش برای ترکیب خوشه‌های و بدست آوردن خوشه‌های نهایی ارائه شده است. همان‌طوری که در بخش۳-۶-۲ اشاره شد، ایده اصلی در این روش انباشت مدارک نام دارد. روش کار در اینجا بدین صورت می‌باشد که از فاصله اقلیدسی بین نتایج یک روش تشخیص اجتماع استفاده می‌نماییم که این فاصله توسط تابع pdist در زبان matlab پیاده‌سازی شده است. با اجرای این تابع به تعداد، نتایج ایجاد شده در روش‌های تشخیص اجتماع و جمع نمودن نتایج با هم مقدار فاصله میان هر جفت گره را با هم می‌توان بدست آورد، حال از روی این فاصله ایجاد شده می‌توان ماتریس همبستگی متناظر را بر مبنای فاصله ایجاد نمود. در این پایان‌نامه از این روش برای بدست آوردن شباهت بین نتایج مختلف بدست آمده از روش‌های گوناگون تشخیص اجتماعات استفاده نموده‌ایم. بدین صورت که مقدار یک را از ماتریس همبستگی فاصله بدست آمده کم می‌نماییم تا شباهت بین گره‌ها را از روی ماتریس فاصله بدست آوریم.

۴-۳-۲ تابع توافقی مبتنی بر ابرگراف

همان‌طوری که در بخش۳-۶-۲ اشاره شد، در [۱۴] روش‌هایی برای ترکیب نتایج خوشه‌های ارائه شده است که مبتنی بر ابرگراف می‌باشد. از این الگوریتم‌ها می‌توان به MCLA، HGPA و CSPA اشاره نمود. در هر سه الگوریتم یک ابرگراف ایجاد می‌شود که ارتباطات بین خوشه‌ها در روش‌های مختلف خوشه‌بندی ترکیبی را نشان می‌دهد. هرچه ارتباط میان لبه‌ها در این ابرگراف با توجه به وزنی که به هر لبه داده می‌شود بیشتر باشد احتمال هم خوشه بودن در آن‌ها بیشتر خواهد بود. در این پایان‌نامه از کدهای این روش‌ها با یکسری تغییر در جهت قابل استفاده شدن این روش-های برای تشخیص اجتماعات استفاده نموده‌ایم.

۴-۳-۳ تابع توافقی مبتنی بر پیوند

همان‌طوری که در بخش۳-۶-۲ اشاره شد، این روش در [۳۹] ارائه شده و سه روش مختلف برای بدست آوردن ماتریس همبستگی ارائه شده است، که از این الگوریتم‌ها می‌توان به CTS، SRS و ASRS اشاره نمود. این روش نیز

در اصل همان روش مبتنی بر ماتریس همبستگی می‌باشد، ولی در جهت افزایش کارایی آن روش‌ها می‌باشد. در این پایان‌نامه از کدهای این روش‌ها با یکسری تغییر در جهت قابل استفاده شدن این روش‌های برای تشخیص اجتماعات استفاده نموده‌ایم.

در بخش ۵-۵ شکل ماتریس همبستگی ایجاد شده توسط توابع توافقی را نشان خواهیم داد.

۴-۴ خلاصه‌ی این فصل

در این فصل روشی برای بهبود نتایج تشخیص اجتماعات ارائه گردید. این روش که تشخیص اجتماعات ترکیبی می‌باشد شامل دو مرحله کلی است. در مرحله اول، با استفاده از روش‌های پایه و یا هر روش دیگر در زمینه تشخیص اجتماعات نتایج اولیه روش‌های مختلف را ایجاد نمودیم. برای بدست آوردن این نتایج نیز از الگوریتم‌های مختلف و یا از یک الگوریتم تشخیص اجتماع با تغییر مقادیر اولیه متفاوت و یا با تغییر پارامترهای روش تشخیص اجتماع، نتایج گوناگونی را برای انجام تحلیل‌ها و بررسی‌ها توسط تابع توافقی را ایجاد کردیم. روش تشخیص اجتماعات درصدد پیدا نمودن اجتماعات دقیق‌تر، مطمئن‌تر و مستحکم‌تر با ویژگی‌هایی مانند: عدم نیاز به دانستن تعداد اجتماعات در گراف اولیه، مقیاس‌پذیری بالای این روش، وابسته نبودن به ساختار خاصی از گراف و عدم نیاز به روش‌های خوب و یا نسبتاً خوب برای تشخیص اجتماعات می‌باشد. در بخش بعدی با معرفی مجموعه داده‌های استاندارد مربوط به تشخیص اجتماعات به ارزیابی روش پیشنهادی با روش‌های موجود دیگر می‌پردازیم. لازم به ذکر است که این روش پیشنهادی حتی با داده‌های اشتباه نیز جواب قابل قبولی را ارائه می‌کند که در فصل بعد مورد بررسی قرار می‌گیرد..

فصل پنجم: ارزیابی روش پیشنهادی

در این فصل به ارزیابی روش پیشنهادی می‌پردازیم.

۵-۱ مقدمه

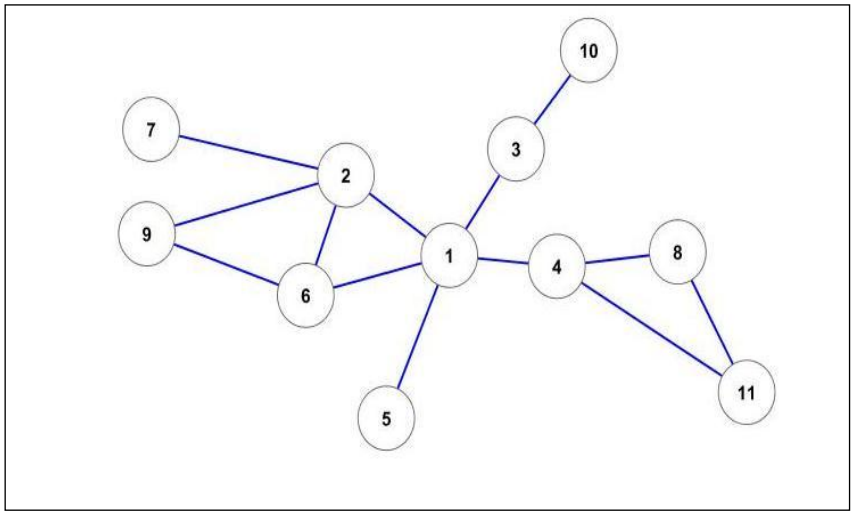
در این قسمت به ارزیابی روش پیشنهادی تشخیص اجتماعات که با استفاده از خوشه‌بندی ترکیبی بدست آمده‌اند، می‌پردازیم. برای این منظور از یکسری داده‌های استاندارد استفاده شده است، که در ادامه این داده‌ها و نتایج حاصله از ارزیابی آن‌ها را مورد بررسی قرار می‌دهیم. به این ترتیب که اجتماعات به دست آمده از طریق روش‌های پیشنهادی با اجتماعات به دست آمده از روش‌های دیگر مقایسه می‌شود.

۵-۲ مجموعه داده‌های مورد استفاده

در اینجا تعدادی داده‌های استاندارد مورد بررسی و آزمایش قرار گرفته است که به تشریح آن‌ها می‌پردازیم.

۵-۲-۱ گراف آزمایشی

برای آزمایش روش پیشنهادی، ابتدا یک گراف آزمایشی ساده طراحی شد. این گراف شامل ۱۱ گره و ۱۳ لبه است که در شکل ۵-۱ نشان داده شده است.

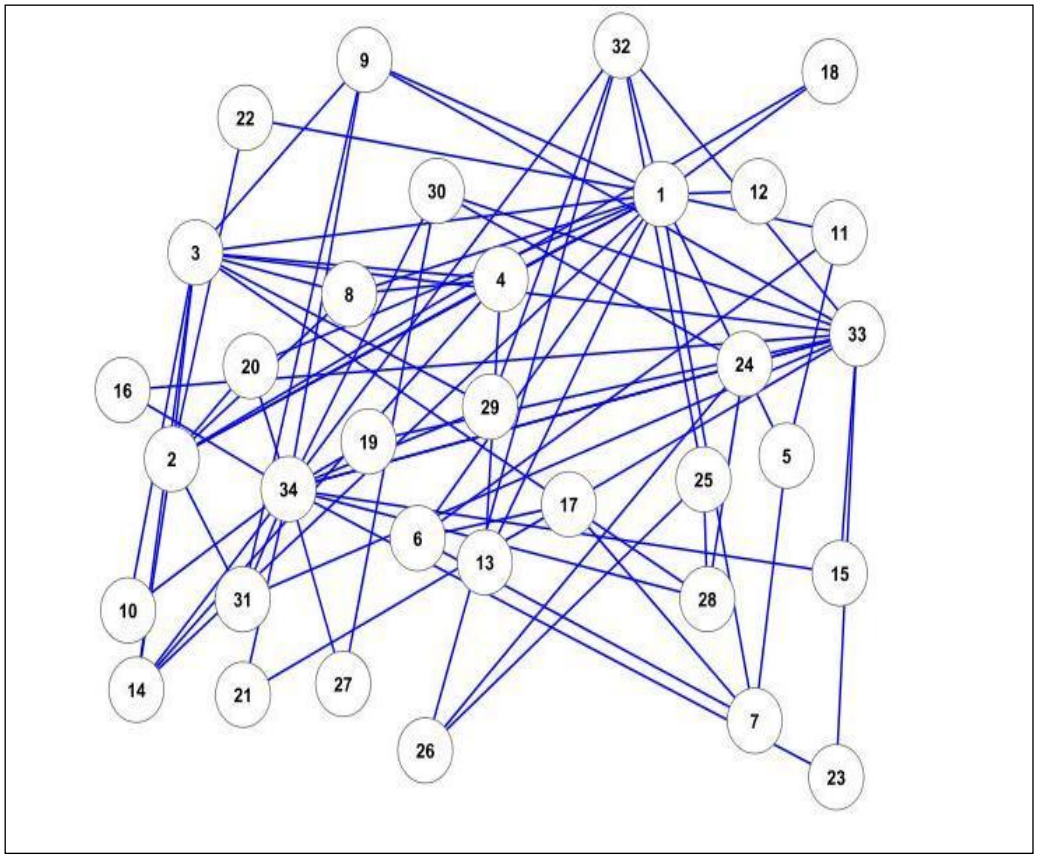


شکل ۵-۱ گراف آزمایشی

۵-۲-۲ باشگاه کاراته زاکاری

در شکل ۵-۲ گراف باشگاه کاراته زاکاری[1] را نشان می‌دهد. این مجموعه یک گراف استاندارد برای تشخیص اجتماع است و در بسیاری از مقالات از آن استفاده شده است. این گراف شامل ۳۴ گره است که اعضای یک باشگاه کاراته در ایالات متحده هستند و در طول ۳ سال مشاهده شده‌اند. لبه‌ها، افرادی را که خارج از فعالیت‌های باشگاه با هم ارتباط دارند به هم متصل می‌کند. در بعضی نقاط درگیری بین مسئول باشگاه با مربی موجب ایجاد شکاف بین اعضای باشگاه شده است و آن‌ها را به دو اجتماع طرفدار این دو فرد تقسیم کرده است[۴۳].

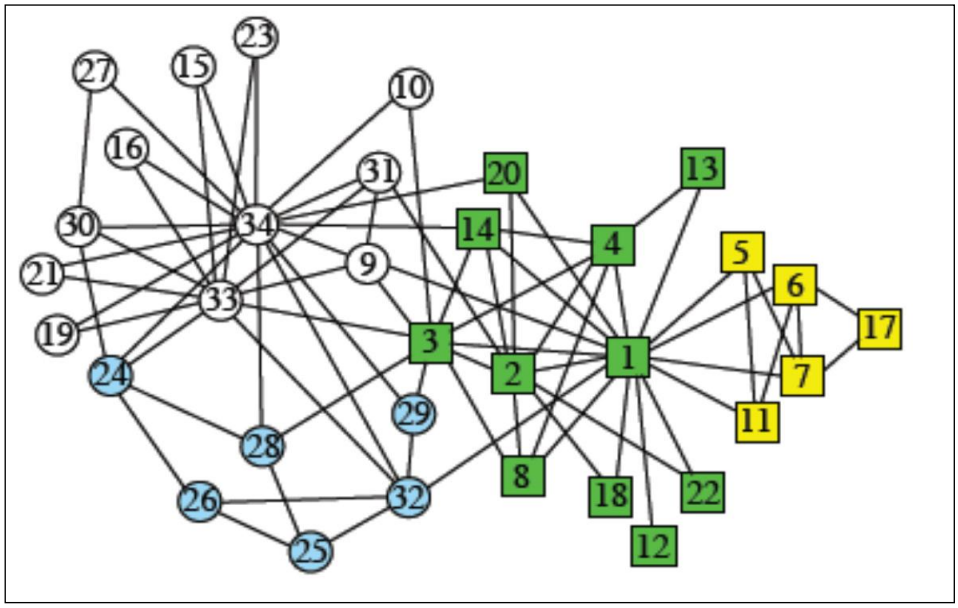
همان‌طور که در شکل ۵-۲ مشاهده می‌شود گراف فوق ارتباطات موجود در باشگاه کاراته را نشان می‌دهد، الگوریتم‌های تشخیص اجتماعات می‌بایست از روی گراف‌هایی بدین شکل اجتماعات گوناگون را پیدا نماید.



شکل ۵-۲ گراف اولیه باشگاه کاراته

[1] Zachary karate club

با نگاه به شکل می‌توان دو تجمع را تشخیص داد. یکی اطراف گره ۳۳ و ۳۴ (۳۴ مسئول باشگاه است) و دیگری اطراف گره۱(مربی) قرار دارد. گره‌های دیگری نیز مانند ۳، ۹ و ۱۰ بین این دو ساختار وجود دارد. این گره‌های معمولاً به درستی توسط روش‌های تشخیص اجتماع، تشخیص داده نمی‌شوند.



شکل ۵-۳ گراف باشگاه کاراته با اعمال روش‌های تشخیص اجتماع[۱۱]

در شکل ۵-۳ دو روش مختلف تشخیص اجتماع را بر روی باشگاه کاراته نشان می‌دهد، یک روش به وسیله دایره و مربع نشان داده شده است، تعداد اجتماعات در این روش دو می‌باشد. روش دیگر توسط رنگ‌های گوناگون اجتماعات مختلف را نشان می‌دهد و تعداد اجتماعات در این روش چهار می‌باشد.

۵-۲-۳ لیگ فوتبال دانشگاه امریکا

این شبکه که لیگ فوتبال دانشگاه امریکا[1] را نشان می‌دهد، از ۱۱۵ گره تشکیل شده است. هر گره نشان دهنده یک تیم می‌باشد و در صورت بازی بین دو تیم در یک فصل یک لبه بین آن دو ایجاد می‌گردد. تعداد لبه‌ها در این شبکه ۶۱۶ لبه می‌باشد که نشان دهنده تعداد بازی‌های انجام شده بین دو تیم بوده است. این مشاهدات برای سال-های ۲۰۰۰ تا ۲۰۰۱ بوده است[۲۵].

[1] American College Football League

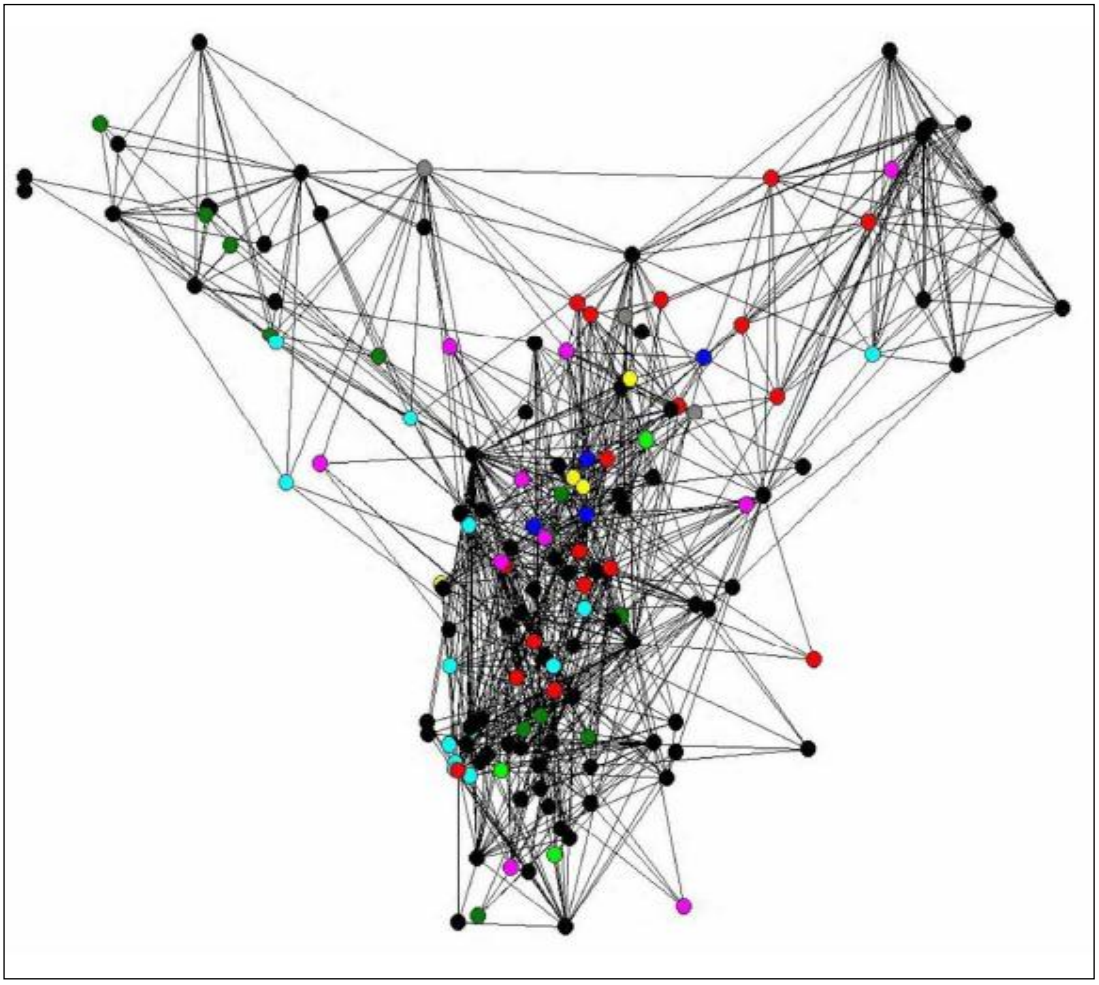
۵-۲-۴ موسیقی دانان جاز

در این پایگاه داده ۱۹۸ باند موسیقی وجود دارد که بین سال‌های ۱۹۱۲ و ۱۹۴۰ فعالیت می‌کردند که بیشتر آن‌ها در سال‌های اطراف ۱۹۲۰ مشغول بوده‌اند. این پایگاه داده موسیقی‌دان‌هایی را لیست می‌کند که در این باندها می-نواختند و شامل ۱۲۷۵ نام موسیقی‌دان مختلف است. اگر دو موسیقی‌دان در یک باند می‌نواختند لبه‌ها بین آن‌ها ترسیم می‌شود[۴۴].

۵-۲-۵ مجموعه داده‌های ایمیل Enron

مجموعه داده‌های ایمیل Enron توسط FERC[1] منتشر شد و البته اشکالاتی هم داشت. پس از آن توسط افراد دیگری بررسی شد و بسیاری از اشکالات آن برطرف شد. این داده‌ها همه نوع ایمیل‌های شخصی و اداری است. این نسخه از داده‌ها شامل ۴۳۱،۵۱۷ ایمیل از ۱۵۱ کاربر است که در ۳۵۰۰ فولدر توزیع شده است. این ایمیل‌ها بخش ضمیمه شده(مثل عکس، فیلم یا هر فایل دیگری) ندارد[۴۵].

[1] Federal Energy Regulatory Commission



شکل ۵-۴ نمایی از قسمتی از شبکه ایمیل[۴۵]

۵-۲-۶ شبکه متابولیکی

در یک سلول یا میکرو ارگانیزم فرآیندهایی که انتقال انرژی و اطلاعات را تولید می‌کنند، به طور یکپارچه از طریق شبکه پیچیده‌ای از مولکول‌های سلولی و واکنش‌هایشان است. با وجود این که نقش کلیدی این شبکه‌ها حفظ کارکردهای سلولی است. اساساً ساختار بزرگ آن‌ها ناشناخته است. این شبکه، یک شبکه متابولیکی کرم‌ها به نام C.Elegans است که ۴۵۳ گره دارد. گره‌ها نمایانگر متابولیک‌ها و لبه‌ها نشان‌دهنده واکنش‌های متابولیکی است.

۵-۲-۷ شبکه علمی

در این شبکه تعداد ۱۵۸۹ نویسنده وجود دارد که با هم در مورد تئوری شبکه پژوهش می‌کنند. این مجموعه داده با بررسی انجام شده بر روی مقالات مختلف در تئوری شبکه توسط نیومن در سال ۲۰۰۶ گردآوری شد. بدین‌صورت که ارتباطات این نویسندگان در مقالات مختلف را به صورت یک شبکه مدل می‌نماید. تعداد لبه‌ها در این شبکه ۲۷۴۲ لبه می‌باشد که نشان دهنده ارتباطات بین نویسندگان است.

حال که پیش‌زمینه لازم برای روش پیشنهادی فراهم شد و داده‌های مورد استفاده در ارزیابی نیز مورد بررسی قرار گرفت، حال به بررسی نتایج گوناگون می‌رسیم.

۵-۳ نتایج اجرای الگوریتم‌های مختلف بر روی مجموعه داده‌ها

نتایج اجرای چند الگوریتم مختلف را بر روی مجموعه داده‌ای تشریح شده در این بخش، در جدول ۵-۱ آورده شده است. در اینجا به بررسی چهار الگوریتم پیاده‌سازی شده در زبان MATLAB R2009a برای تشخیص اجتماعات می-پردازیم که هر یک به چه صورتی اجتماعات را بدست می‌آورند. کلیه این روش‌ها از روی ماتریس مجاورتی ایجاد شده از روی گراف استفاده می‌کنند، کلیه این روش‌ها نیز قابل تعمیم بر روی گراف‌های جهت‌دار و وزن‌دار می‌باشند. در اینجا به بررسی چهار الگوریتم مختلف تشخیص اجتماعات پایه می‌پردازیم. این الگوریتم‌ها در واقع همان روش‌هایی هستند که در تابع پیشنهادی در بخش ۴-۲ برای پیدا نمودن تعداد اجتماعات استفاده شده‌اند. ولی برای اجرای روش پیشنهادی در تشخیص اجتماعات ترکیبی به نتایج بیشتری نیاز داریم بنابراین با دو روش ارائه شده در بخش ۴-۲ با استفاده از روش‌های دیگر تشخیص اجتماع و با ایجاد تغییر پارامتر و مقادیر اولیه بر روی یک الگوریتم این مشکل را رفع نموده‌ایم. حال به توضیح الگوریتم‌های پایه‌ای در تشخیص اجتماعات و بررسی نتایج حاصل از آن‌ها بر روی داده-های استاندارد مختلف می‌پردازیم.

الگوریتم اول که با نام Newman یا N در این پایان‌نامه معرفی شده است، یک روش بر مبنای مرکزیت بینابینی است، که در بخش ۳-۳-۱ به آن اشاره شد. در این روش لبه‌هایی که بیشترین مرکزیت را داشته باشند، به صورت بازگشتی حذف نموده و به این ترتیب یکسری اجتماعات جدا از هم را خواهیم داشت[۲۲].

الگوریتم دوم که با نام Newman Greedy یا NG در این پایان‌نامه معرفی شده است، یک روش بر مبنای روش تجمعی ارائه شده است ایده اصلی در این روش بهینه‌سازی پودمانی و حریصانه است، در این روش با تعریف رابطه پودمانی همانند رابطه در بخش ۲-۶، درصدد پیدا نمودن اجتماعاتی با بیشنه‌ترین حالت پودمانی هستند. این روش چون از انواع روش‌های تجمعی می‌باشد به یک معیار مشابهت نیاز داریم که در این روش پودمانی معیار مشابهت خواهد بود[۲۷].

الگوریتم سوم که با نام Newman & Girvan یا GN در این پایان‌نامه معرفی شده است، یک روش مشابه مرکزیت بینابینی است با این تفاوت که به جای بررسی تک‌تک گره‌ها در هر مرحله و بدست آوردند یک لبه برای حذف شدن در هر مرحله لبه‌ها با ضریب بالا را حذف نموده و در نهایت یک درختواره از کل گراف ایجاد می‌شود که یک نمایش بایگان می‌باشد، با برش این درختواره می‌توان اجتماعات را تفکیک نمود ولی در برش می‌بایست دقت نمود که بهینه‌ترین حالت را داشته باشد، برای پیدا نمودن بهینه‌ترین حالت نیز از پودمانی استفاده شده است[۱۱].

الگوریتم چهارم که با نام Radicchi & et al یا R در این پایان‌نامه معرفی شده است، نیز یک روش بر مبنای بهینه‌سازی پودمانی است که با استفاده از پودمانی و تعریف متغیرهای دیگری که در آن مقادیر پودمانی و نیز حالات گره‌ها در آن وضعیت ذخیره می‌شود. این روش باعث افزایش سرعت در پیدا نمودن اجتماعات و نیز با داشتن یک پیش‌زمینه از مراحل قبلی با دقت بالاتری به ترکیب گره‌ها می‌پردازد[۴۶].

در جدول ۵-۱ پودمانی محاسبه شده توسط چهار الگوریتم بررسی شده بر روی داده‌های استاندارد ارائه شده در بخش ۵-۲ می‌باشد. همان‌طوری که در جدول ۵-۱ مشاهده می‌شود، چهار روش مختلف تشخیص اجتماعات با داده-های مختلف مورد بررسی قرار گرفته است، مقادیر پودمانی هر روش به همراه تعداد اجتماعات بدست آمده توسط روش‌های مختلف در هر ستون جدول ۵-۱ نوشته شده است. با توجه به داده‌های استاندارد ورودی در هر سطر از جدول یکی از الگوریتم‌ها بر روی داده‌های مورد نظر بهتر جواب می‌دهد. در مجموع می‌توان گفت که بعضی روش‌ها نسبت به روش‌های دیگر جواب بهتری دارند ولی این اطمینان به طور قطعی وجود ندارد. بنابراین به روش‌های ترکیبی نیاز داریم که از مجموع روش‌های فوق در نهایت یک جواب قطعی و نهایی را ایجاد نماید. حال به بررسی روش پیشنهادی با داده‌های استاندارد ارائه شده در بالا می‌پردازیم.

جدول ۵-۱ پودمانی محاسبه شده توسط روش‌های مختلف پایه بر روی مجموعه داده‌های استاندارد

Data Set	تعداد گره‌ها	تعداد لبه‌ها	Newman [۲۲]		Newman Greedy [۲۷]		Newman & Girvan [۱۱]		Radicchi & et al [۴۶]	
آزمایشی	۱۱	۱۳	۳ اجتماع	0.4258	۳ اجتماع	0.4518	۴ اجتماع	0.4304	۳ اجتماع	0.3984
کاراته	۳۴	۷۸	۴ اجتماع	0.4086	۴ اجتماع	0.3980	۴ اجتماع	0.4009	۳ اجتماع	0.3853
فوتبال	۱۱۵	۶۱۶	۷ اجتماع	0.5788	۹ اجتماع	0.6057	۶ اجتماع	0.5714	۶ اجتماع	0.5581
جاز	۱۶۸	۲۷۴۲	۳ اجتماع	0.4389	۴ اجتماع	0.4442	۵ اجتماع	0.4397	۴ اجتماع	0.4394
متابولیکی	۴۵۳	۲۰۳۲	۱۰ اجتماع	0.4332	۱۳ اجتماع	0.4269	۱۱ اجتماع	0.4334	۱۰ اجتماع	0.4019
ایمیل	۱۱۳۳	۵۴۵۱	۱۱ اجتماع	0.5485	۱۰ اجتماع	0.5766	۱۰ اجتماع	0.5463	۱۱ اجتماع	0.5036
شبکه علمی	۱۵۸۹	۲۷۴۲	۴۰۷ اجتماع	0.9547	۴۰۹ اجتماع	0.9511	۴۰۵ اجتماع	0.9472	۴۰۳ اجتماع	0.9419

همان‌طوری که در جدول ۵-۱ مشاهده می‌شود، مقدار پودمانی و تعداد اجتماعات بدست آمده توسط روش‌های مختلف پایه را بر روی داده‌های متفاوت مورد مقایسه قرار داده‌ایم.

۵-۴ نتایج اجرای روش پیشنهادی با روش‌های مختلف اولیه

در این بخش معیار پودمانی را برای ارزیابی روش‌های پیشنهادی با روش‌های پایه، مورد بررسی قرار می‌دهیم.

برای اجرای روش پیشنهادی در تشخیص اجتماعات ترکیبی از الگوریتم‌های پایه‌ای در تشخیص اجتماعات استفاده نموده‌ایم. تعداد الگوریتم‌های ارائه شده در قسمت قبلی برای بدست آوردن تعداد اجتماعات کفایت می‌کند، بدین معنی که با اجرای روش‌های فوق می‌توان تعداد اجتماعات را پیدا نمود. ولی برای ایجاد ماتریس ترکیبی که این ماتریس نتایج تشخیص اجتماعات گوناگون است، نیاز به روش‌های بیشتری از روش‌های تشخیص اجتماعات داریم. بنابراین از الگوریتم‌ها مختلف و یا با تغییر در مقادیر اولیه و پارامترهای یک الگوریتم می‌توان نتایج مختلفی را بدست آورد که در روش پیشنهادی تشخیص اجتماعات ترکیبی از آن استفاده می‌کنیم.

توابع توافقی که برای تشخیص اجتماعات استفاده شده‌اند با اجرای الگوریتم‌های خوشه‌بندی متفاوت نتایج مختلفی را ارائه می‌کنند در اینجا نتایج توابع توافقی مختلف را در جدول نشان می‌دهیم.

در ادامه این فصل به بررسی نتایج بدست آمده از روش‌های مختلف تشخیص اجتماعات ترکیبی با روش‌های پایه می‌پردازیم، در ابتدا روش ترکیبی EAC و بعد از آن به ترتیب روش‌های CTS، ASRS، SRS و در انتها روش‌های ابرگراف را مورد مطالعه قرار می‌دهیم.

اعمال تابع توافقی EAC بر روی مجموعه داده‌های مختلف نتایج جدول ۵-۲ را می‌دهد. نتایج بدست آمده از روش ترکیبی EAC در تمام حالات از نتایج بدست آمده از الگوریتم‌های تشخیص اجتماعات پایه‌ای بیشتر بوده است، و این روش بهترین روش نسبت به روش‌های دیگر ترکیبی برای تشخیص اجتماعات برای این مجموعه داده‌ها بوده است.

جدول ۵-۳ پودمانی محاسبه شده توسط روش‌های مختلف پایه و EAC بر روی مجموعه داده‌های استاندارد

Avg E	EAC			Avg B	Basic Community Detection Method				Data Set
	EAC-C	EAC-A	EAC-S		R	GN	NG	N	
0.4528	0.4528	0.4528	0.4528	0.4266	0.3984	0.4304	0.4518	0.4258	آزمایشی
0.4191	0.4188	0.4188	0.4197	0.3982	0.3853	0.4009	0.3980	0.4086	کاراته
0.5953	0.5858	0.5983	0.6019	0.5777	0.5581	0.5714	0.6027	0.5788	فوتبال
0.4444	0.4444	0.4438	0.4450	0.4405	0.4394	0.4397	0.4442	0.4389	جاز
0.4355	0.4334	0.4377	0.4356	0.4238	0.4019	0.4334	0.4269	0.4332	متابولیکی
0.5716	0.5746	0.5634	0.5769	0.5437	0.5036	0.5463	0.5766	0.5485	ایمیل
0.9515	0.9544	0.9548	0.9551	0.9487	0.9419	0.9472	0.9511	0.9547	شبکه علمی

همان‌طوری که در جدول ۵-۴ مشاهده می‌شود، نتایج بدست آمده توسط روش EAC دارای دقت بالاتر و پایدارتری نسبت به روش‌های پایه تشخیص اجتماعات است.

اعمال تابع توافقی CTS بر روی مجموعه داده‌های مختلف نتایج جدول ۵-۵ را می‌دهد. میانگین نتایج بدست آمده از روش‌های پایه در تمام نتایج بدست آمده از روش تشخیص اجتماعات ترکیبی در CTS کمتر بوده است، ولی روش CTS بهترین نتایج را نسبت به روش‌های دیگر تشخیص اجتماعات ترکیبی مثل EAC به همراه ندارد.

جدول ۵-۶ پودمانی محاسبه شده توسط روش‌های مختلف پایه و CTS بر روی مجموعه داده‌های استاندارد

Avg E	CTS			Avg B	Basic Community Detection Method				Data Set
	CTS-C	CTS-A	CTS-S		R	GN	NG	N	
0.4528	0.4528	0.4528	0.4528	0.4266	0.3984	0.4304	0.4518	0.4258	آزمایشی
0.4056	0.3990	0.4188	0.3990	0.3982	0.3853	0.4009	0.3980	0.4086	کاراته
0.5965	0.5894	0.5983	0.6019	0.5777	0.5581	0.5714	0.6027	0.5788	فوتبال
0.4440	0.4435	0.4438	0.4447	0.4405	0.4394	0.4397	0.4442	0.4389	جاز
0.4319	0.4273	0.4317	0.4368	0.4238	0.4019	0.4334	0.4269	0.4332	متابولیکی
0.5647	0.5542	0.5634	0.5765	0.5437	0.5036	0.5463	0.5766	0.5485	ایمیل
0.9534	0.9544	0.9512	0.9546	0.9487	0.9419	0.9472	0.9511	0.9547	شبکه علمی

همان‌طوری که در جدول ۵-۷ مشاهده می‌شود، نتایج بدست آمده توسط روش CTS دارای دقت بهتری نسبت به روش‌های پایه تشخیص اجتماعات است.

اعمال تابع توافقی ASRS بر روی مجموعه داده‌های مختلف نتایج جدول ۵-۸ را می‌دهد. میانگین نتایج بدست آمده از روش‌های پایه در تمام نتایج بدست آمده از روش تشخیص اجتماعات ترکیبی در ASRS نیز کمتر بوده است، با این حال روش ASRS نیز بهترین نتایج را نسبت به روش‌های دیگر تشخیص اجتماعات ترکیبی به همراه ندارد.

جدول ۵-۹ پودمانی محاسبه شده توسط روش‌های مختلف پایه و ASRS بر روی مجموعه داده‌های استاندارد

Avg E	ASRS			Avg B	Basic Community Detection Method				Data Set
	ASRS-C	ASRS-A	ASRS-S		R	GN	NG	N	
0.4044	0.3088	0.4522	0.4522	0.4266	0.3984	0.4304	0.4518	0.4258	آزمایشی
0.4119	0.3990	0.4188	0.4179	0.3982	0.3853	0.4009	0.3980	0.4086	کاراته
0.5939	0.5894	0.5985	0.5939	0.5777	0.5581	0.5714	0.6027	0.5788	فوتبال
0.4412	0.4357	0.4434	0.4445	0.4405	0.4394	0.4397	0.4442	0.4389	جاز
0.4328	0.4320	0.4331	0.4334	0.4238	0.4019	0.4334	0.4269	0.4332	متابولیکی
0.5634	0.5547	0.5670	0.5685	0.5437	0.5036	0.5463	0.5766	0.5485	ایمیل
0.9519	0.9502	0.9521	0.9534	0.9487	0.9419	0.9472	0.9511	0.9547	شبکه علمی

همان‌طوری که در جدول ۵-۱۰ مشاهده می‌شود، نتایج بدست آمده توسط روش ASRS دارای نتایج محکم‌تری نسبت به روش‌های پایه تشخیص اجتماعات است.

اعمال تابع توافقی SRS بر روی مجموعه داده‌های مختلف نتایج جدول ۵-۱۱ را می‌دهد. میانگین نتایج بدست آمده از روش‌های پایه در تمام نتایج بدست آمده از روش تشخیص اجتماعات ترکیبی در SRS نیز کمتر بوده است، با این حال روش SRS نیز بهترین نتایج را نسبت به روش‌های دیگر تشخیص اجتماعات ترکیبی به همراه ندارد.

جدول ۵-۱۲ پودمانی محاسبه شده توسط روش‌های مختلف پایه و SRS بر روی مجموعه داده‌های استاندارد

Avg E	SRS			Avg B	Basic Community Detection Method				Data Set
	SRS-C	SRS-A	SRS-S		R	GN	NG	N	
0.4518	0.4518	0.4518	0.4518	0.4266	0.3984	0.4304	0.4518	0.4258	آزمایشی
0.4119	0.3990	0.4188	0.4197	0.3982	0.3853	0.4009	0.3980	0.4086	کاراته
0.5984	0.5985	0.5983	0.5985	0.5777	0.5581	0.5714	0.6027	0.5788	فوتبال
0.4433	0.4431	0.4438	0.4432	0.4405	0.4394	0.4397	0.4442	0.4389	جاز
0.4348	0.4333	0.4337	0.4374	0.4238	0.4019	0.4334	0.4269	0.4332	متابولیکی
0.5694	0.5713	0.5606	0.5763	0.5437	0.5036	0.5463	0.5766	0.5485	ایمیل
0.9494	0.9443	0.9513	0.9528	0.9487	0.9419	0.9472	0.9511	0.9547	شبکه علمی

همان‌طوری که در جدول ۵-۱۳ مشاهده می‌شود، نتایج بدست آمده توسط روش SRS دارای نتایج مطمئن‌تری نسبت به روش‌های پایه تشخیص اجتماعات است.

نتایج بدست آمده از روش‌های EAC، CTS، ASRS و SRS در مجموع در تمامی مجموعه داده‌های مورد بررسی دارای نتایجی با دقت، اطمینان، پایداری و صحت بالاتری نسبت به روش‌های پایه‌ای در تشخیص اجتماعات می‌باشند.

اعمال تابع توافقی ابرگراف بر روی مجموعه داده‌های مختلف نتایج جدول ۵-۱۴ را می‌دهد.

جدول ۵-۱۵ پودمانی محاسبه شده توسط روش‌های مختلف پایه و ابرگراف بر روی مجموعه داده‌های استاندارد

Avg E	Hayper Graph			Avg B	Basic Community Detection Method				Data Set
	CSPA	HGPA	MCLA		R	GN	NG	N	
0.4512	0.4522	0.4488	0.4528	0.4266	0.3984	0.4304	0.4518	0.4258	آزمایشی
0.3905	0.3836	0.3683	0.4197	0.3982	0.3853	0.4009	0.3980	0.4086	کاراته
0.5816	0.5687	0.5775	0.5988	0.5777	0.5581	0.5714	0.6027	0.5788	فوتبال
0.4397	0.4375	0.4390	0.4428	0.4405	0.4394	0.4397	0.4442	0.4389	جاز
0.4203	0.3989	0.4288	0.4334	0.4238	0.4019	0.4334	0.4269	0.4332	متابولیکی
0.5326	0.5137	0.5246	0.5597	0.5437	0.5036	0.5463	0.5766	0.5485	ایمیل
0.9428	0.9432	0.9394	0.9458	0.9487	0.9419	0.9472	0.9511	0.9547	شبکه علمی

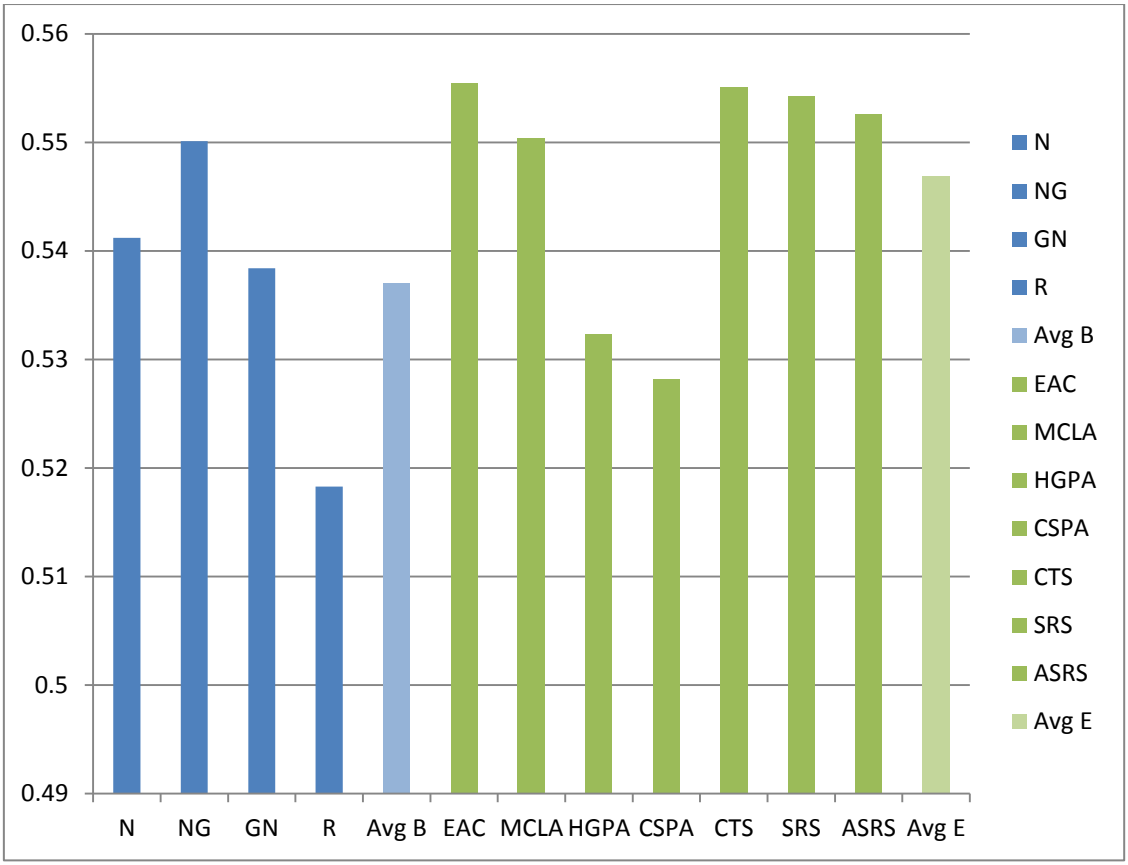
همان‌طوری که در جدول ۵-۱۶ مشاهده می‌شود، نتایج بدست آمده توسط روش‌های مختلف ابرگراف دارای نتایجی نزدیک به میانگین نتایج روش‌های پایه تشخیص اجتماعات می‌باشند. نتایج روش MCLA نسبت به دو روش دیگر HGPA و CSPA بهتر بوده است در مجموع می‌توان گفت که نتایج روش‌های ابرگراف برای تشخیص اجتماعات ترکیبی مناسب نمی‌باشند.

در جدول ۵-۱۷ نتایج روش‌های ECD یا تشخیص اجتماعات ترکیبی مختلف را با روش‌های پایه به صورت کلی نمایش می‌دهد.

جدول ۵-۱۷ پودمانی محاسبه شده توسط روش‌های پایه و مقایسه با کل روش‌های ECD

Avg E	ECD							Avg B	Basic Community Detection Method				Data Set
	ASRS	SRS	CTS	CSPA	HGPA	MCLA	EAC		R	GN	NG	N	
0.4519	0.4522	0.4518	0.4528	0.4522	0.4488	0.4528	0.4528	0.4266	0.3984	0.4304	0.4518	0.4258	آزمایشی
0.4068	0.4179	0.4197	0.4188	0.3836	0.3683	0.4197	0.4197	0.3982	0.3853	0.4009	0.3980	0.4086	کاراته
0.5922	0.5985	0.5985	0.6019	0.5687	0.5775	0.5988	0.6019	0.5777	0.5581	0.5714	0.6027	0.5788	فوتبال
0.4424	0.4445	0.4438	0.4447	0.4375	0.4390	0.4428	0.4450	0.4405	0.4394	0.4397	0.4442	0.4389	جاز
0.4294	0.4334	0.4374	0.4368	0.3989	0.4288	0.4334	0.4377	0.4238	0.4019	0.4334	0.4269	0.4332	متابولیکی
0.5566	0.5685	0.5763	0.5765	0.5137	0.5246	0.5597	0.5769	0.5437	0.5036	0.5463	0.5766	0.5485	ایمیل
0.9491	0.9534	0.9528	0.9546	0.9432	0.9394	0.9458	0.9551	0.9487	0.9419	0.9472	0.9511	0.9547	شبکه علمی
0.5469	0.5526	0.5543	0.5551	0.5282	0.5323	0.5504	0.5555	0.5370	0.5183	0.5384	0.5501	0.5412	AVG

همان‌طوری که در جدول ۵-۱۷ ملاحظه می‌شود میانگین اعمال نتایج مختلف یک الگوریتم یا روش خاص بر روی تمامی مجموعه داده‌ها آورده شده است که نشان دهنده نحوه اجرای و یا چگونگی نتایج آن الگوریتم یا روش مورد استفاده بر روی مجموعه داده‌های متفاوت می‌باشند. همان‌طوری که در نتایج مختلف بدست آمده مشاهده می‌شود استفاده از خوشه‌بندی ترکیبی باعث افزایش مقدار پودمانی می‌گردد. در شکل ۵-۵ یک نمودار کلی از نتایج مختلف را نشان می‌دهد.



شکل ۵-۵ بررسی مقدار پودمانی کلیه روش‌ها

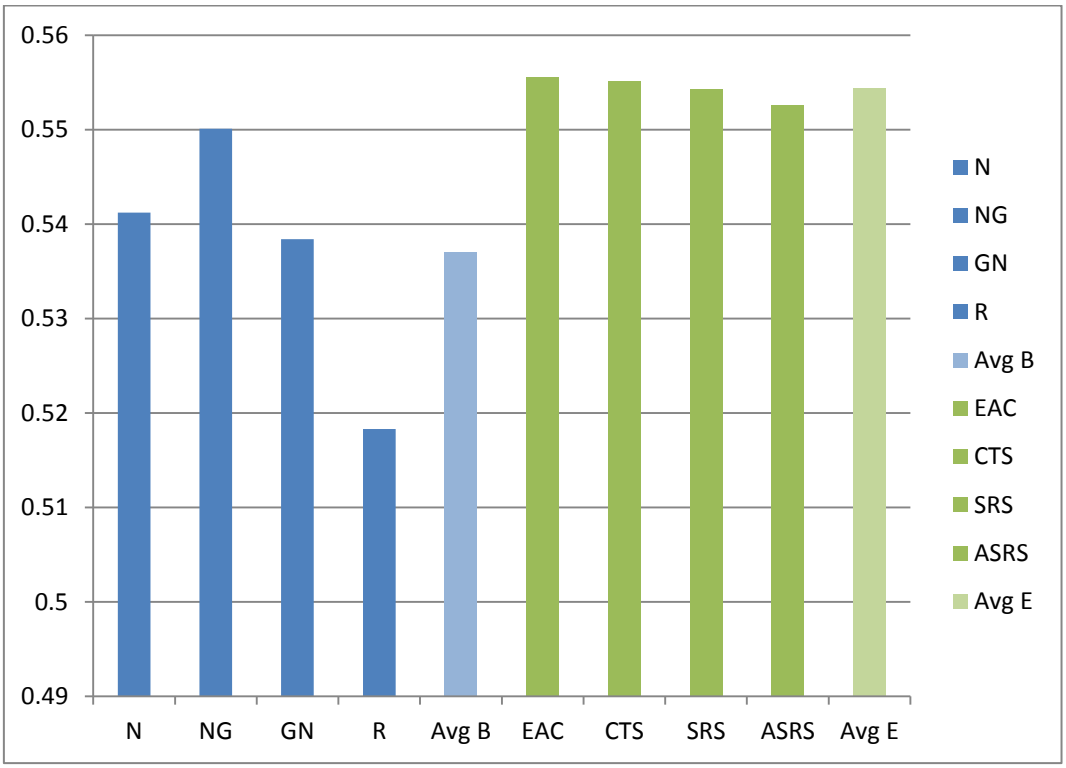
همان‌طوری که در نمودار مشاهده می‌شود، رنگ آبی نتایج روش‌های پایه تشخیص اجتماعات و آبی کم رنگ میانگین نتایج پایه را نشان می‌دهد. رنگ سبز نیز نتایج روش‌های تشخیص اجتماعات ترکیبی را و سبز کم رنگ میانگین نتایج ترکیبی را نشان می‌دهد. با ترکیب نتایج مختلف از تشخیص اجتماعات می‌توان به نتایج پایدارتر و مستحکم‌تر رسید. نتایج بدست آمده از ابرگراف نسبت به روش‌های دیگر مناسب نیستند، بنابراین در تشخیص اجتماعات ترکیبی استفاده نکردن از آن‌ها می‌تواند بهتر باشد بنابراین می‌توان آن‌ها را نادیده گرفت چرا که روش‌های دیگر تشخیص اجتماعات ترکیبی به نسبت ابرگراف بهتر جواب می‌دهند.

در جدول ۵-۱۸ نتایج الگوریتم‌های پایه را با بعضی روش‌های تشخیص اجتماعات ترکیبی مقایسه می‌کنیم. در این مقایسه روش‌های ابرگراف را به علت دقت کم آن‌ها به نسبت روش‌های دیگر حذف نموده‌ایم.

جدول ۵-۱۸ پودمانی محاسبه شده توسط روش‌های اولیه و مقایسه با بعضی روش‌های ECD

Avg E	ECD				Avg B	Basic Community Detection Method				Data Set
	ASRS	SRS	CTS	EAC		R	GN	NG	N	
0.4519	0.4522	0.4518	0.4528	0.4528	0.4266	0.3984	0.4304	0.4518	0.4258	آزمایشی
0.4068	0.4179	0.4197	0.4188	0.4197	0.3982	0.3853	0.4009	0.3980	0.4086	کاراته
0.5922	0.5985	0.5985	0.6019	0.6019	0.5777	0.5581	0.5714	0.6027	0.5788	فوتبال
0.4424	0.4445	0.4438	0.4447	0.4450	0.4405	0.4394	0.4397	0.4442	0.4389	جاز
0.4294	0.4334	0.4374	0.4368	0.4377	0.4238	0.4019	0.4334	0.4269	0.4332	متابولیکی
0.5566	0.5685	0.5763	0.5765	0.5769	0.5437	0.5036	0.5463	0.5766	0.5485	ایمیل
0.9491	0.9534	0.9528	0.9546	0.9551	0.9487	0.9419	0.9472	0.9511	0.9547	شبکه علمی
0.5543	0.5526	0.5543	0.5551	0.5555	0.5370	0.5183	0.5384	0.5501	0.5412	AVG

در شکل ۵-۶ نتایج مختلف روش‌های پایه و بعضی روش‌های پیشنهادی را بر روی نمودار نشان می‌دهد، همان‌طوری که مشاهده می‌شود با استفاده از این روش پیشنهادی می‌توان با اطمینان بیشتری به نتایج حاصل از روش‌های تشخیص اجتماعات اتکا نمود.



شکل ۵-۶ نمودار حاصل از نتایج مختلف پایه و پیشنهادی

همان‌طوری که در نمودار مشاهده می‌شود، رنگ آبی نتایج روش‌های پایه تشخیص اجتماعات و آبی کم رنگ میانگین نتایج پایه را نشان می‌دهد. رنگ سبز نیز نتایج روش‌های تشخیص اجتماعات ترکیبی را و سبز کم رنگ میانگین نتایج ترکیبی را نشان می‌دهد. نتایج روش‌های تشخیص اجتماعات ترکیبی به علت دقت، اطمینان و پایداری بالای آن می‌تواند در تصمیمات بلند مدت و یا کوتاه مدت و سیاست‌های سازمانی و یا مواردی از این قبیل تأثیر بسزای داشته باشند.

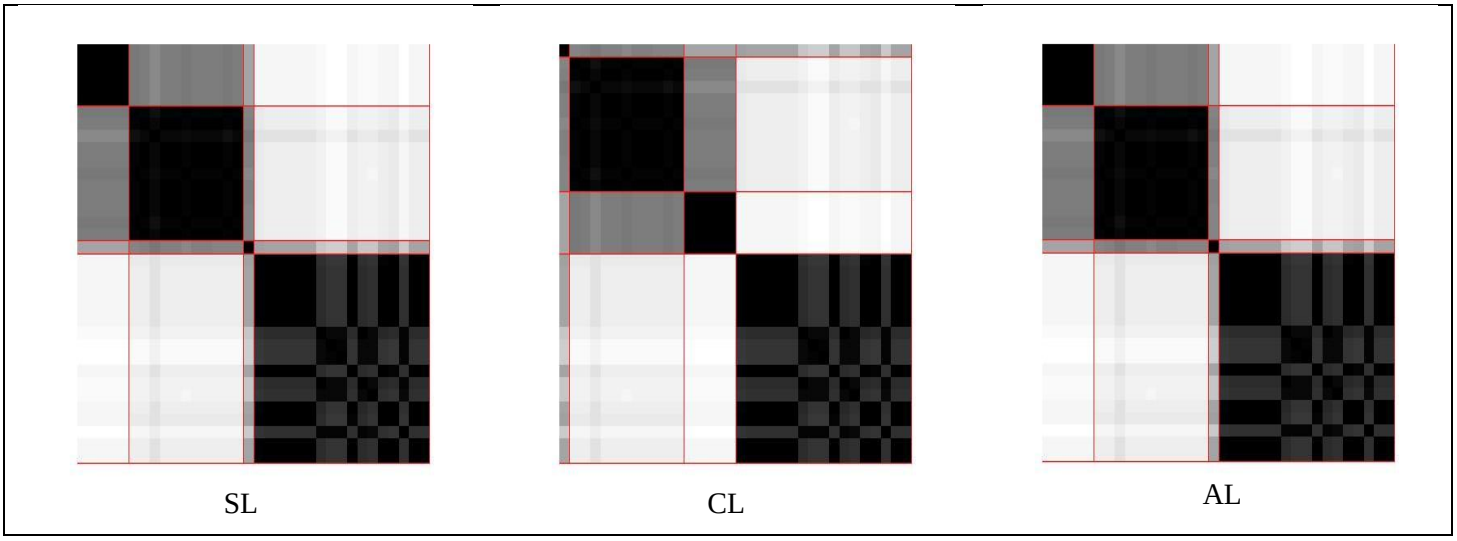
امکان استفاده از الگوریتم‌ها و روش‌های دیگر برای تشخیص اجتماعات در روش پیشنهادی وجود دارد و این نشان از مقیاس‌پذیری بالای این روش است، اضافه نمودن آن روش‌ها نتایج روش ترکیبی را نیز بهبود می‌دهد.

امکان استفاده از داده‌های وب نیز همانند داده استاندارد ایمیل وجود دارد. در این پایان‌نامه صرفاً یکسری داده‌های استاندارد با تعداد کمی گره را مورد بررسی قرار دادیم که هزینه اجرایی آن برای یک دستگاه رایانه خانگی قابل انجام است. برای اجرای داده‌های حجیم تر نیاز به سیستم‌های پردازشی بهتری داریم که توانایی انجام پردازش برای شبکه اجتماعی همانند شبکه جهانی را داشته باشد.

در بخش بعدی نمایی از ماتریس همبستگی ایجاد شده از روش پیشنهادی را بر روی چند داده را نمایش داده‌ایم. با اجرای یکی از روش‌های خوشه‌بندی بایگان می‌توان اجتماعات دقیق‌تر را بدست آورد.

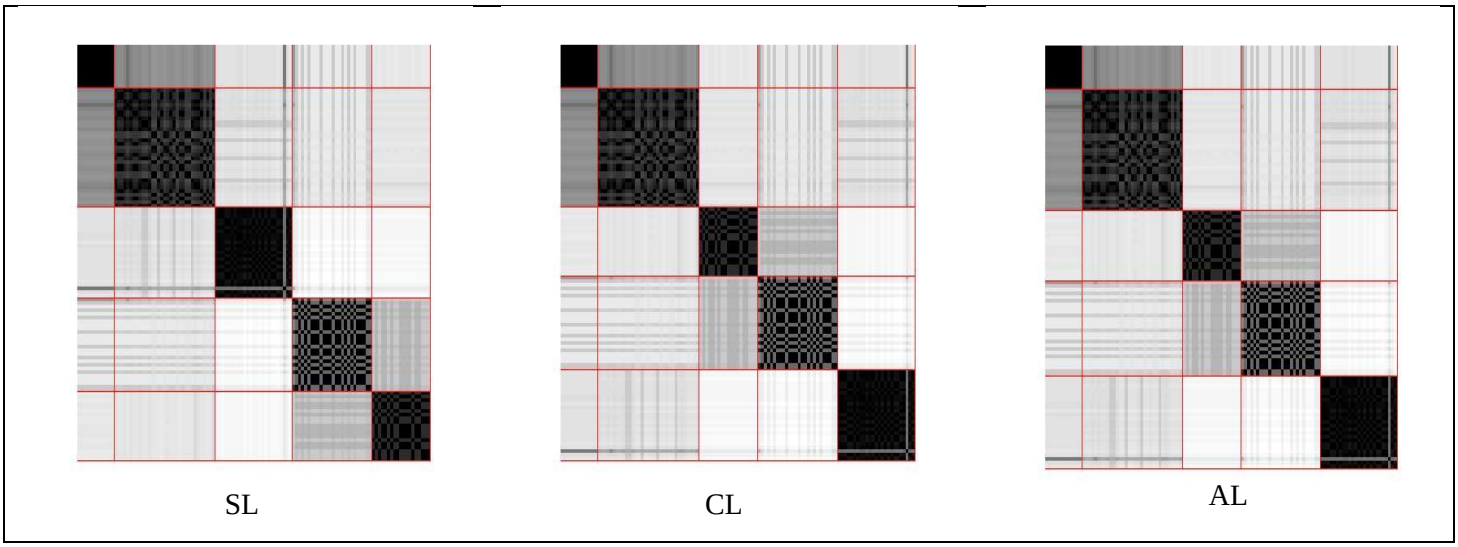
۵-۵ نمایش ماتریس همبستگی ایجاد شده از توابع توافقی

این قسمت ماتریس همبستگی ایجاد شده از تابع توافقی پیشنهادی را نشان می‌دهد. این ماتریس را با درصدی از رنگ‌ها سفید و سیاه نمایش می‌دهیم. در این قسمت صرفاً مجموعه داده‌های باشگاه کاراته و لیگ فوتبال دانشگاه امریکا را به عنوان ورودی به تابع توافقی پیشنهادی خواهیم داد. اجتماعات بدست آمده از این روش‌ها در تصاویر بدین صورت می‌باشد که هرچه ارتباط میان گره‌ها(وابستگی گره‌ها در ماتریس همبستگی ایجاد شده) با هم بیشتر است رنگ نمایش آن‌ها تیره‌تر خواهد بود. با اعمال یکی از روش‌های خوشه‌بندی بایگان می‌توان اجتماعات مختلف را از روی ماتریس همبستگی بدست آورد.



شکل ۵-۷ اجتماعات بدست آمده از ماتریس همبستگی باشگاه کاراته

در شکل ۵-۷ اجتماعات بدست آمده از ماتریس همبستگی توسط تابع توافقی پیشنهادی را با اعمال الگوریتم‌های بایگان از قبیل SingleLinkage(SL)، CompleteLinkage(CL) و AverageLinkage(AL) برای باشگاه کاراته را نشان می‌دهد.



شکل ۵-۸ اجتماعات بدست آمده از ماتریس همبستگی لیگ فوتبال دانشگاه امریکا

در شکل ۵-۸ اجتماعات بدست آمده از ماتریس همبستگی توسط تابع توافقی پیشنهادی را با اعمال الگوریتم‌های بایگان از قبیل SL، CL و AL برای لیگ فوتبال دانشگاه امریکا را نشان می‌دهد.

۵-۶ خلاصه‌ی این فصل

در این فصل ارزیابی روش پیشنهادی را مورد بررسی قرار دادیم، در ابتدای این فصل تعدادی مجموعه داده‌های استاندارد که در اکثر روش‌های تشخیص اجتماعات مورد استفاده قرار می‌گیرد را توضیح دادیم. سپس با استفاده از الگوریتم‌های پایه در تشخیص اجتماعات به بررسی این داده‌ها بر روی روش‌های مختلف تشخیص اجتماعات و روش پیشنهادی تشخیص اجتماعات ترکیبی پرداختیم. در نهایت نتایج مختلف بدست آمده از روش‌های مختلف اولیه را با روش پیشنهادی مقایسه نمودیم که میانگین نتایج بدست آمده در تشخیص اجتماعات ترکیبی بهتر از روش‌های تشخیص اجتماعات پایه‌ای بود. باید توجه نمود که حتی اگر یک یا چند روش نتایج نسبتاً بد یا خیلی بدی در روش-های اولیه برای تولید نتایج مختلف وجود داشته باشند در نهایت تفاوت زیادی در نتایج نهایی ایجاد نمی‌شود و باز هم خوشه‌بندی ترکیبی توانایی پیدا نمودن اجتماعات مناسب را خواهد داشت. در فصل بعدی نتیجه‌گیری کلی از این پایان‌نامه ارائه خواهد شد.

فصل ششم: نتیجه‌گیری

۱-۶ مروری بر گزارش پایان نامه

وب، محیطی وسیع، متنوع و پویا است که کاربران متعدد اسناد خود را در آن منتشر می‌کنند. وب طی یک فرآیند آشفته و غیرمتمرکز رشد می‌کند و این روند منجر به تولید حجم وسیعی از مستندات متصل به یکدیگر گشته است که از هیچ‌گونه سازماندهی منطقی برخوردار نیستند. بنابراین به ابزارهایی برای مدیریت این اطلاعات در وب نیاز است. خوشه‌بندی یا تشخیص اجتماعات در شبکه‌های اجتماعی مشکل مهمی می‌باشد که ساختار گروه‌ها در شبکه‌های اجتماعی، ارتباطات پنهان بین مؤلفه‌های آن را آشکار خواهد نمود. با در نظر گرفتن افزایش پایگاه داده‌های مربوط به شبکه‌های اجتماعی، به الگوریتم‌های مقیاس‌پذیری برای تجزیه تحلیل آن‌ها نیاز داریم. مسئله اصلی در تشخیص اجتماعات این است که بدانیم چگونه به بهترین حالت شبکه را به اجتماعات اصلی آن تقسیم کنیم. در شبکه‌های واقعی هیچ اطلاعاتی درباره تعداد اجتماعات وجود ندارد. این چالش، تشخیص اجتماعات را دچار مشکل می‌کند. در بعضی از روش‌های تشخیص اجتماعات فرض بر آن است که تعداد اجتماعات شبکه را از قبل می‌دانیم. در حالیکه در بسیاری از شبکه‌ها، هیچ دانش اولیه‌ای در مورد اجتماعات شبکه وجود ندارد. روش‌های جدید به دنبال بر طرف کردن این نقیصه هستند، در این پایان‌نامه نیز یک روش ترکیبی با نام تشخیص اجتماعات ترکیبی برای پیدا نمودن تعداد اجتماعات و ایجاد نتایج دقیق‌تر و پایدارتر و مستحکم‌تر از روی گراف حاصله از شبکه مورد تحلیل، ارائه می‌شود.

اکثر روش‌های رایج تشخیص اجتماعات موجود قطعی نیستند، و نتایج آن‌ها به مقادیر اولیه‌ای که در اکثر مواقع به صورت تصادفی انتخاب می‌شود بستگی دارد. خوشه‌بندی ترکیبی در تحلیل داده‌ها برای رسیدن به نتایج پایدار بدون توجه به مقادیر اولیه تصادفی استفاده می‌گردد. روش پیشنهادی ارائه شده مستقل از نوع روش استفاده شده برای تشخیص اجتماعات است، این روش تنها به نتایج نهایی هر الگوریتم نگاه می‌کند و با کنار هم قرار دادن نتایج متفاوت سعی در ارائه یک نتیجه دقیق‌تر و مستحکم‌تر برای تشخیص اجتماعات می‌باشد. در این پایان‌نامه نشان دادیم که خوشه‌بندی ترکیبی توانایی ترکیب با هر روش دیگری را خواهد داشت به گونه‌ای که دقت نتایج اجتماعات را افزایش می‌دهد.

از مزایای مهم خوشه‌بندی ترکیبی در تشخیص اجتماعات می‌توان به استفاده از روش‌های قبلی به تعداد زیاد که نشان از مقیاس‌پذیری بالای این روش و نیز دقت بالای آن به نسبت روش‌های دیگر و در مجموع نتایج دقیق‌تر و پایدارتری را خواهد داشت. مزایای مهم دیگر در این روش اینست که، نیازی به دانستن تعداد اجتماعات نیز وجود ندارد. در بسیاری از روش‌های تشخیص اجتماع نیازمند دانستن تعداد اجتماعات هستیم، با توجه به اینکه تعداد

اجتماعات در وب برخلاف روش‌های خوشه‌بندی معین نیست این روش می‌تواند کمک شایانی را در این خصوص داشته باشد. از معایب این روش به هزینه اجرایی آن که با بدترین حالت هزینه اجرایی در الگوریتم‌های استفاده شده برابر است و حافظه زیادی را نیز مصرف می‌نماید، همچنین ممکن است در همه موارد به جواب صددرصد بهینه نرسیم و الگوریتم‌های پایه برای داده‌های خاصی جواب دقیق‌تری را داشته باشند. به طور میانگین در تمامی روش‌ها نتیجه نهایی از میانگین کلی نتایج اولیه بهتر بوده است و در بعضی موارد از تمامی نتایج نیز بهتر مشاهده شده است.

یکی از ویژگی‌های روش پیشنهادی آن است که وابسته به ساختار خاصی از گراف برای تشخیص اجتماع نیست و تقریباً بر روی تمام ساختارها به صورت گرافی بدون جهت در این پایان‌نامه و جهت‌دار جواب می‌دهد، امکان استفاده از گراف‌های جهت‌دار نیز برای این روش پیشنهادی وجود دارد. در حالی که برخی از روش‌های تشخیص اجتماع ساختار خاصی از گراف مانند گراف کامل دوبخشی را جستجو می‌کنند و بخش‌هایی از گراف وب را که شامل این ساختار می‌باشند به عنوان اجتماع معرفی می‌کنند. ویژگی دیگر مقیاس‌پذیری بالای این روش است که امکان استفاده از هر روش تشخیص اجتماعات وجود دارد، هر چه تعداد روش‌ها زیاد باشد تشخیص اجتماعات ترکیبی نتیجه مناسب‌تری را ارائه می‌کند. ویژگی دیگر پراکندگی در نتایج اولیه تشخیص اجتماع و نتایج نسبتاً بد در روش‌های اولیه است که هر چه پراکندگی نتایج اولیه بدست آمده بیشتر باشد، در نهایت دقت نتایج در خوشه‌بندی ترکیبی نهایی بهتر خواهد بود.

۶-۲ نتیجه‌گیری

همان‌طوری که مشاهده شد روش‌های مختلفی برای تشخیص اجتماعات ارائه وجود دارد. با توجه به شباهت‌های خوشه‌بندی و تشخیص اجتماعات در این پایان‌نامه از روش‌های خوشه‌بندی ترکیبی برای تشخیص اجتماعات استفاده شد. روش پیشنهادی مستقل از نوع روش استفاده شده برای تشخیص اجتماعات است، این روش تنها به نتایج نهایی هر الگوریتم نگاه می‌کند و با کنار هم قرار دادن نتایج متفاوت سعی در ارائه یک نتیجه دقیق‌تر و مستحکم‌تر می‌باشد.

از نتایج این بررسی‌ها می‌تواند در مسائل بسیاری از جمله: بهبود موتورهای جستجو، درک ساختار شبکه، تشخیص دقیق‌تر اجتماعات، بازاریابی، تبلیغات، و مورد استفاده قرار گیرد.

۶-۳ کارهای آتی

از کارهای آتی می‌توان به تعریف معیارهای جدید برای ارزیابی روش‌های تشخیص اجتماعات اشاره نمود. همچنین امکان استفاده از این روش‌ها برای رفع مشکل ترافیک در شبکه، تشخیص خطا در نرم‌افزار و نیز تشخیص گروه‌های جاسوسی و.... وجود دارد. همچنین می‌تواند با ایجاد یک ساختار از ارتباطات موجود در وب و در نهایت اعمال الگوریتم-های تشخیص اجتماعات پایه و تشخیص اجتماعات ترکیبی در راستای پیدا نمودن اجتماعات در وب برای بهبود نتایج موتورهای جستجو، بازاریابی و تبلیغات بود. همچنین می‌توان از این روش‌ها در سازمان‌های مختلف برای مثال مخابرات، تلفن همراه و یا هر سازمان دیگری برای بررسی ارتباطات بین افراد پرداخت.

۶-۴ خلاصه‌ی این فصل

این فصل به نتیجه‌گیری و جمع‌بندی کل گزارش اختصاص داشت. در این راستا ابتدا گزارش را به اجمال مرور کردیم. سپس به نتیجه‌گیری پرداختیم. در این راستا مزایا و معایب روش تشخیص اجتماعات ترکیبی را مورد بررسی قرار دادیم. همان طور که دیدیم روش پیشنهادی دارای نتایج بهتری نسبت به روش‌های پایه‌ای در تشخیص اجتماعات می‌باشد. همچنین تعدادی پیشنهاد برای انجام کارهای آتی در این زمینه ارائه گردید.

مراجع

[1] Kosala, R., and Blockeel, H., "Web mining research: A survey," *In SIGKDD Explorations Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining,* Vol. 2, No. 1, pp. 1-15, 2001.

[2] Fortunato, S., "Community detection in graphs," *Physics Reports,* Vol. 486, No. 3-5, pp. 75–174, 2010.

[3] Easley, D., and Kleinberg, J., "Networks, Crowds, and Markets: Reasoning about a Highly Connected World," *Cambridge University Press,* June 10, 2010.

[4] Liu, B., "Web DataMining Exploring Hyperlinks, Contents, and Usage Data," *Springer,* 2007.

[5] Porter, M.A., Onnela, J.P., and Mucha, P.J., "Communities in networks," *Notices of the American Mathematical Society,*V. 56, No. 9, pp. 1082–1097, 2009.

[6] Desikan, P., Srivastava, J., Kumar, V., and Tan, P., "Hyperlink Analysis -- Techniques & Applications," *Army High Performance Computing Center Technical Report,* 2002.

[7] Newman, M.E.J., "Communities, modules and large-scale structure in networks," *Nature Physics,* Vol. 8, January 2012.

[8] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D., "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 101, No. 9, p. 2658, 2004.

[9] Lancichinetti, A., "Community detection algorithms: a comparative analysis," *Physical Review E,* Vol. 80, No. 5, p. 056117, 2009.

[10] Fred, A., and Jain, A.K., "Data Clustering Using Evidence Accumulation," *In Conference on Pattern Recognition, ICPR02,* Quebec City, pp. 276 – 280, 2002.

[11] Newman, M.E.J., and Girvan, M., "Finding and evaluating community structure in networks," *Physical review E,* Vol. 69, No. 2, p. 26113, 2004.

[12] Clauset, A., Newman, M.E.J., and Moore, C., "Finding community structure in very large networks," *Physical Review E,* Vol. 70, No. 6, p. 66111, 2004.

[13] Faceli, K., Marcilio, C.P., and Souto, D., "Multi-objective Clustering Ensemble," *Proceedings of the Sixth International Conference on Hybrid Intelligent Systems,* 2006.

[14] Strehl, A., and Ghosh, J., "Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions," *Journal of Machine Learning Research,* pp. 583-617, 2003.

[15] Kleinberg, J., "Authoritative sources in a hyperlinked environment," *Journal of the ACM,* Vol. 46, 1999.

[16] Borodin, A., Roberts, G., Rosenthal, J., and Tsaparas, P., "Link analysis ranking: algorithms, theory, and experiments," *ACM Trans. Inter. Tech.,* Vol. 5, No. 1, pp. 231-297, 2005.

[17] Lin C., and Chen, M.S., "VIPAS: Virtual Link Powered Authority Search in the web," *29th Conference on Very Large Databases (VLDB),* 2003.

[18] Page, L., Brin, S., Motwani, R., and Winograd, T., "The PageRank citation ranking: bringing order to the web," *Stanford Publications,* 1998.

[19] Haveliwala, T.H., "Topic-sensitive PageRank," *Proceedings of the Eleventh International World Wide Web Conference*, 2002.

[20] Kamvar, S.D., Haveliwala, T.H., Manning, C.D., and Golub, G.H., "Exploiting the block structure of the web for computing PageRank," *Stanford University Technical Report*, 2003.

[21] Baeza-Yates, R., and Davis, E., "Web page ranking using link attributes," *International World Wide Web Conference, Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, New York, NY, USA, pp. 328-329, 2004.

[22] Newman, M.E.J., "A measure of betweenness centrality based on random walks," *Social networks*, Vol. 27, No. 1, pp. 39–54, 2005.

[23] Arenas, A., Cabrales, A., Diaz-Guilera, A., Guimera, R., and Vega-Redondo, F., "Search and congestion in complex networks," *Statistical Mechanics of Complex Networks*, pp. 175–194, 2003.

[24] Duch, J., and Arenas, A., "Community detection in complex networks using extremal optimization," *Physical Review E*, Vol. 72, No. 2, p. 027104, 2005.

[25] Girvan, M., and Newman, M.E.J., "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 99, No. 12, p. 7821, 2002.

[26] Tyler, J.R., Wilkinson, D.M., and Huberman, B.A., "E-mail as spectroscopy: Automated discovery of community structure within organizations," *The Information Society*, Vol. 21, No. 2, pp. 143–153, 2005.

[27] Newman, M.E.J., "Fast algorithm for detecting community structure in very large networks," *Physical review E*, Vol. 69, 2004.

[28] Shen, H., Cheng, X., Cai, K., and Hu, M. B., "Detect overlapping and hierarchical community structure in networks," *Physica A: Statistical Mechanics and its Applications*, Vol. 388, No. 8, pp. 1706–1712, 2009.

[29] Zhang, X. S., et al., "Modularity optimization in community detection of complex networks," *Euro Physics Letters(EPL)*, Vol. 87, p. 38002, 2009.

[30] Seary, A.J., and Richards, W.D., "Partitioning networks by eigenvectors," *In Proceedings of the International conference on Social Networks*, Vol. 1, pp. 47–58, 1995.

[31] Newman, M.E.J., "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, Vol. 74, No. 3, p. 36104, 2006.

[32] Newman, M.E.J., "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, Vol. 103, No. 23, p. 8577, 2006.

[33] Newman, M.E.J., "Detecting community structure in networks," *The European Physical Journal B-Condensed Matter and Complex Systems*, Vol. 38, No. 2, pp. 321–330, 2004.

[34] Duch J., and Arenas, A., "Community detection in complex networks using extremal optimization," *Physical Review E*, Vol. 72, No. 2, p. 027104, 2005.

[35] Tasgin, M., and Bingol, H., "Community detection in complex networks using genetic algorithm," *Arxiv preprint cond-mat/0604419*, 2006.

[36] Pizzuti, C., "Ga-net: A genetic algorithm for community detection in social networks," *Parallel Problem Solving from Nature–PPSN X*, pp. 1081–1090, 2008.

[37] Lipczak, M., and Milios, E., "Agglomerative genetic algorithm for clustering in social networks," *In Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, pp. 1243–1250, 2009.

[38] Leskovec, J., Lang, K.J., and Mahoney, M., "Empirical comparison of algorithms for network community detection," *In Proceedings of the 19th international conference on World Wide Web*, pp. 631–640, 2010.

[39] Iam-on, N., and Garrett, S., "LinkCluE: A MATLAB Package for Link-Based Cluster Ensembles," *In Journal of Statistical Software*, Issue 9, Volume 36, August 2010.

[40] Ayad, H., and Kamel, M., "Cluster-based cumulative ensembles," *In the 6th International Workshop on Multiple Classifier Systems*, pp. 236–245. LNCS 3541, 2005.

[41] Vega-Pons, S., and Ruiz-Shulcloper, J., "A survey of clustering ensemble algorithms," *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 25, No. 3, pp. 337-372, 2011.

[42] Lancichinetti, A., and Fortunato, S., "Consensus clustering in complex networks," *Nature. Scientific Reports*, March 2012.

[43] Zachary, W.W., "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, No. 4, pp. 452–473, 1977.

[44] Gleiser, P., and Danon, L., "Community structure in jazz," *Arxiv preprint condmat/ 0307434*, 2003.

[45] Shetty, J., and Adibi, J., "The Enron email dataset database schema and brief statistical report," *Information Sciences Institute Technical Report*, University of Southern California, 2004.

[46] Reichardt, J., and Bornholdt, S., "Statistical mechanics of community detection," *Physical Review E*, 2006.

[۴۷] مطیعی س.، میبدی م.ر، "داده‌کاوی ساختار وب با استفاده از اتوماتای یادگیر توزیع‌شده و سلولی و کاربردهـای آن"، پایـان نامه کارشناسی ارشد، دانشگاه صنعتی امیرکبیر، بهار ۱۳۸۷.

[۴۸] علیزاده ح.، مینایی بیدگلی ب.، "خوشه‌بندی ترکیبی مبتنی بر زیرمجموعه‌ای از نتایج اولیه"، پایان نامه کارشناسی ارشد، دانشگاه علم و صنعت، اسفند ۱۳۸۷.

واژه‌نامه

A	
Adjacency	مجاورتی، همسایگی
Adjacency matrix	ماتریس مجاورتی
Agglomerative	تجمعی
Approach	رویکرد، روش
Association Rule	قواعد انجمنی
Average Linkage	اتصال میانگین
B	
Betweenness centrality	مرکزیت بینابینی
Base Set	مجموعه پایه
Biased	جهت‌دار، دارای سو گیری
Bridge	پل
C	
Cache	نهانگاه
Caching	نهان کردن
Capacity	ظرفیت
Clique	پیوسته
Clustering	خوشه‌بندی
Coarse-grained	درشت دانه
Community	اجتماع
Community detection	تشخیص اجتماع
Complete Linkage	اتصال کامل
Congruence	متجانس
Consensus Function	تابع توافقی
Crawling	پیمایش
Customer Relationship Management	مدیریت ارتباط با مشتری
D	
Data Mining	داده‌کاوی

Data set	مجموعه داده
Data Warehouse	انبار داده‌ها
Dendrogram	درختواره
Directed	مستقیم
Distributed	توزیع‌شده
Diversity	پراکندگی
Document Object Model	مدل شیئ سند
Document Structure Analysis	تحلیل ساختار سند
Divisive	تقسیم‌کننده
Dynamic	پویا
E	
Edge	لبه
Ensemble Clustering	خوشه‌بندی ترکیبی
Ensemble Community Detection	تشخیص اجتماعات ترکیبی
Evidence Accumulation Clustering	خوشه‌بندی انباشت مدارک
Evolutionary approach	رویکرد تکاملی
F	
Fine-grained	ریزدانه
Focused Crawling	پیمایش موضوعی
G	
General	عمومی
Generalization	تعمیم
Graph theory	تئوری گراف
Graph visualization	بصری سازی گراف
Greedy techniques	تکنیک‌های حریصانه
H	
Heuristic	بر مبنای تجربه
Hierarchical	بایگان
I	
Incongruence	نامتجانس

Indexing	شاخص گذاری
Inter Document Hyperlink	پیوند درون سند
Intra Document Hyperlink	پیوند بین سند
Isolate	مجزا
L	
Link Analysis	تحلیل پیوند
Log File	فایل ثبت وقایع
M	
Minimal Cut	برش کمینه
Modularity	پودمانی
Modularity maximization	بیشینه‌سازی پودمانی
Modularity optimization	بهینه‌سازی پودمانی
N	
Neighborhood Graph	گراف همسایگی
Node	گره
Normalized cut	برش نرمال‌شده
O	
Overlapping clique	پیوستگی دارای همپوشانی
P	
Partitional	افرازبندی، تقسیم‌بندی
Personalization	شخصی سازی
Power Law	قانون توان
Precision	دقت
Q	
Quasi-clique	شبه پیوستگی
Query	پرس‌وجو
Query Dependent Schemes	وابسته به پرس‌وجو
Query Independent Schemes	مستقل از پرس‌وجو
R	
Random Walk	قدم زدن تصادفی
Ranking	رتبه بندی

Recall	فراخوان
Recommendation Systems	سیستم‌های پیشنهاد دهنده
Related Page Algorithm	الگوریتم صفحات مرتبط
Root Set	مجموعه ریشه
S	
Search Engine	موتور جستجو
Seed URLs	پیوندهای آغازین
Session	نشست
Similarity measurement	معیار مشابهت
Single Linkage	اتصال منفرد
Social Network Analysis(SNA)	تحلیل شبکه‌های اجتماعی
Sparse	پراکنده، تنک
Spectral methods	روش‌های طیفی
Stemming	کاهش کلمات به ریشه آن‌ها
Stop Words	کلمات زائد
T	
Thread	ریسمان، نخ
Topic Discovery	تشخیص موضوع
U	
Undirected	بدون جهت
Usability	قابلیت استفاده
V	
Validation	اعتبار سنجی
W	
Web Community	اجتماع وب
Web Content Mining	داده‌کاوی محتوای وب
Web Mining	وب کاوی
Web Structure Mining	داده‌کاوی ساختار وب
Web Usage Mining	داده‌کاوی استفاده از وب
WWW	شبکه جهانی وب

Abstract

Nowadays social networks have different applications especially among internet users, so social network analysis (SNA) is an important and effective area of research. Web mining techniques have been introduced recently for the purpose of utilizing the enormous volume of web data. Web mining is the use of data mining techniques to automatically discover and extract information from web documents and services. One of the categories of web mining is web structure mining which obtains information about web pages through the existing hyperlinks between these pages. One of the great challenges in SNA is community detection. Community is a group of vertices with high intra-connection and sparse inter-connection.

The community detection in complex networks or social networks is an important problem. Community detection or Clustering show community structure of social networks and hidden relationships among their constituents. By considering the increase of datasets related to social networks, it needs scalable algorithms to analyze these networks.

Most community detection methods currently available are not deterministic and their results typically depend on the specific random seeds, initial conditions. In this project, it will be proposed an approach with name of Ensemble Community Detection (ECD) which goals in finding robust communities, with using Ensemble Clustering. Ensemble clustering is used in data analysis to generate stable results out of a set of partitions delivered by stochastic methods. In this paper, it will be shown that Ensemble clustering can be combined with any existing method in a self-consistent way, enhancing considerably the accuracy of the resulting partitions. With this result it can use in: improving search engines, realizing network structure and finding robust communities, advertisement and etc.

KeyWords: *Clustering, Community Detection, Ensemble Community Detection, Ensemble Clustering, Modularity, Social Network, Web Mining.*

Shahid Beheshti University

Department of Electrical & Computer Engineering

A Novell method for improving clustering algorithm in community detection by using web mining

By

Rasoul Hosseinzadeh

A THESIS SUBMITTED

TO FOR THE DEGREE OF

MASTER OF SCIENCE

Supervisor:
Dr. Eslame Nazemi

Advisor:
Hossein Alizadeh

Winter (2013)