اجتماع‌یابی صفحات وب در شبکه‌های اینترنتی دارای ویژگی با استفاده از رویکرد برنامه‌ریزی ریاضی

دو فصلنامه علمی- پژوهشی

مدیریت اطلاعات

دوره ۴، شماره ۲

پاییز و زمستان ۱۳۹۷

اسماعیل علی نژاد

دانشجوی دکترای مهندسی صنایع، دانشکده مهندسی صنایع و سیستم‌های دانشگاه تربیت مدرس

بابک تیمور پور

استادیار گروه مهندسی صنایع، دانشکده مهندسی صنایع و سیستم‌های دانشگاه تربیت مدرس[1]

چکیده: اجتماع‌یابی (کشف اجتماعات) یکی از شاخه‌های نوظهور و پرطرف‌دار در علم داده‌کاوی و تحلیل شبکه‌های اجتماعی است که کاربردهای فراوانی در کشف و تحلیل اجتماع‌ها در سایت‌های اینترنتی، شبکه‌های زیستی، علمی و پژوهشی و غیره دارد. اجتماع‌یابی صفحات اینترنتی می‌تواند به‌طور ویژه به مدیران سایت‌های اینترنتی در تخصیص پهنای بهینه به شبکه صفحات وب تحت نظارتشان کمک کند. در اکثر روش‌های اجتماع‌یابی موجود فقط از توپولوژی شبکه (ارتباطات، یال‌ها) برای گروه‌بندی گره‌ها (صفحات وب) استفاده می‌شود؛ درحالی‌که نتایج پژوهش‌های اخیر نشان داده است که این‌گونه روش‌ها باید به‌گونه‌ای تغییر کند که در آن‌ها علاوه بر توپولوژی، ویژگی‌های ذاتی گره‌ها نیز در فرآیند اجتماع‌یابی لحاظ شود. ازاین‌رو در این مقاله برای اولین بار با لحاظ کردنِ هم‌زمانِ ویژگی‌های ذاتی صفحات وب و ارتباطات میان آن‌ها، یک مدل ریاضی برای کشف اجتماعات در شبکه‌های اینترنتی توسعه داده‌شده است. روش پیشنهادی این پژوهش بدین‌صورت است که برای لحاظ کردن ویژگی‌ها در فرآیند اجتماع‌یابی، ابتدا با استفاده از یک رویکرد ریاضی، میزان شباهتِ صفحات وب به کمک یک سنجه شباهت (مانند جاکارد یا ضریب انطباق) و بردار ویژگی‌ها محاسبه و به‌عنوان وزن به یال‌های موجود بین آن‌ها در شبکهٔ اینترنتی موردنظر افزوده می‌شود. با این کار عملاً یک شبکه اینترنتی ویژگی‌دار با یال‌های غیر موزون به یک شبکه بدون ویژگی با یالهای موزون تبدیل می‌شود. سپس با استفاده از یک مدل ریاضی (که مختص شبکه‌هایی با یال‌های موزون است)، اجتماعات موجود در این شبکهٔ موزون کشف می‌شود. برای اعتبارسنجی و اثبات کارایی، در قالب آزمون‌های فرض آماری ادعاشده است که کیفیت اجتماعات کشف‌شده توسط رویکرد ریاضی پیشنهادی (که ویژگی‌های صفحات وب را لحاظ می‌کند) به‌طور آماری بهتر از مدل‌های ریاضی پیشین (که از ویژگی‌ها چشم‌پوشی می‌کند) است. نتایج آزمون‌های آماری روی شبکه اینترنتی واقعی نشان می‌دهد که مدل پیشنهادی این پژوهش در حالتی که از معیار جاکارد برای محاسبه میزان شباهت صفحات وب استفاده می‌کند به‌طور معنی‌داری (با $P\text{-value}=0.01$) باعث کشف اجتماعاتی بهتر در قیاس با مدل‌های ریاضی پیشین شده است. همچنین نتایج دیگر آزمون‌های آماری نیز نشان می‌دهد که انتخاب سنجهٔ شباهتِ متناسب با ماهیت شبکه، تأثیر بسزایی در میزان کیفیت رویکرد پیشنهادی دارد.

کلیدواژه‌ها: اجتماع‌یابی، بهینه‌سازی پودمانگی، توپولوژی شبکه، شبکه اینترنتی، صفحات وب، مدل ریاضی، ویژگی‌های گره.

۱. نویسنده مسئول: b.teimourpour@modares.ac.ir

۱- مقدمه

اجتماع‌یابی یکی از مشهورترین موضوعات در حوزه علم مدرن شبکه است که توجهات زیادی را در سال‌های اخیر به خود جلب کرده است (Fortunato and Hric 2016). هدف اصلی مسائل اجتماع‌یابی در گراف‌ها، تشخیص گروه‌های نهان در شبکه با استفاده از ویژگی‌های ساختاری است (Beiró et al. 2013). اخیرا اجتماع‌یابی به موضوعی داغ در حوزه شبکه‌های پیچیده تبدیل‌شده است و ازآنجاکه به‌راحتی می‌توان اطلاعات را با یک شبکه یا گراف نمایش داد (Bello-Orgaz, Salcedo-Sanz and Camacho 2018)، توسط بسیاری از محققان در رشته‌های مربوطه مانند کامپیوتر، ریاضی، زیست‌شناسی، فیزیک و علوم اجتماعی موردمطالعه قرارگرفته است (Xiong, Li and Yang 2018). اجتماع‌یابی کاربردهای فراوانی در بسیاری از مسائل دنیای واقعی مانند شبکه‌های اینترنتی (Fortunato 2010a)، شبکه‌های زیستی پروتئینی (J. Chen and Yuan 2006)، تحلیل شبکه‌های بزرگ (Agarwal and Kempe 2008) و غیره دارد.

با بررسی خلاصه مهم‌ترین مقالات و نوآوری‌های ارائه‌شده در زمینه اجتماع‌یابی می‌توان دریافت که در اکثر روش‌های اجتماع‌یابی تنها از ویژگی‌های ساختاری (توپولوژی) شبکه برای اجتماع‌یابی استفاده می‌شود. به‌عنوان‌مثال، برخی از تازه‌ترین مقالات در این حوزه عبارت‌اند از (Liu et al. 2018; Xiong, Li and Yang 2018; Said et al. 2018; Bello-Orgaz, Salcedo-Sanz and Camacho 2018). در پیشینه پژوهش به چنین شبکه‌هایی همگن یا بدون ویژگی گفته می‌شود. این در حالی است که در بیشتر شبکه‌های اجتماعی گره‌ها خود نیز دارای ویژگی‌های شخصی هستند. در نظر گرفتن این ویژگی‌ها ازآن‌جهت مهم است که در برخی مواقع، دلیل اصلی ایجاد ارتباطات میان اعضا، تناسب شخصیتی و شباهت در خصوصیات فکری- جسمی و موقعیت اجتماعی-شغلی آن‌هاست. در پیشینه پژوهش به چنین شبکه‌هایی ناهمگن یا ویژگی‌دار گفته می‌شود.

دریکی از پژوهش‌های اخیر در زمینه شبکه‌های ناهمگن، (Hric, Darst and Fortunato 2014) با استفاده از شبکه‌های واقعی و ساختگی مشهور نشان داده‌اند مدل‌های پیشین اجتماع‌یابی باید به‌گونه‌ای تغییر کنند که علاوه بر توجه به ساختار ارتباطات، به ویژگی‌های ذاتی گره‌ها نیز توجه کنند. ازاین‌رو در این پژوهش یک مدل ریاضی جدید برای اجتماع‌یابی صفحات وب در شبکه‌های اینترنتی دارای ویژگی با تلفیق دو رویکرد مهم در ادبیات یعنی توجه هم‌زمان به ویژگی‌های ساختاری شبکه (روش‌های اجتماع‌یابی) و ویژگی‌های اختصاصی گره‌ها (روش‌های خوشه‌بندی) پیشنهادشده است. با استفاده از این مدل می‌توان به‌طور هم‌زمان از مزایای هردو دسته از روش‌های خوشه‌بندی و اجتماع‌یابی در کشف اجتماعات نهان در میان صفحات وب بهره‌مند شد که این امر باعث نزدیک‌تر شدن نتایج به گروه‌بندی واقعی صفحات وب می‌شود.

ساختار ادامه این پژوهش بدین‌صورت است که در بخش دوم به توضیح مبانی اجتماع‌یابی و پیشینه رویکردهای اجتماع‌یابی در شبکه‌های دارای ویژگی و بدون ویژگی پرداخته می‌شود. سپس در بخش سوم یک مدل ریاضی با ایده تلفیق ماتریس مجاورت و ماتریس شباهت‌ها پیشنهاد و ساختار آن تبیین می‌شود. در بخش چهارم آزمون فرض‌های آماری جهت بررسی صحت و کارایی مدل پیشنهادی ارائه و سپس در بخش پنجم با استفاده از داده‌های شبکه‌های واقعی، صحت مدل پیشنهادی در مقایسه با مدل‌های پیشین

در هر دو نوع شبکه دارای ویژگی و بدون ویژگی اعتبارسنجی می‌شود. در پایان نیز مهم‌ترین دستاوردهای پژوهش موردبحث و نتیجه‌گیری قرار می‌گیرد و همچنین چند پیشنهاد برای کار آیندگان ارائه می‌شود.

۲- پیشینه پژوهش

۲-۱- تعریف اجتماع و معیار پودمانگی (پیمانه‌ای)

علی‌رغم تعاریف متعدد ارائه‌شده در پیشینه، نمی‌توان تعریف واحدی برای اجتماع یافت که موردقبول تمامی محققان باشد. یکی از مشهورترین تعاریف موجود از اجتماع در ادبیات موضوع، تعریفی است که توسط دو دانشمند به نام‌های نیومن[1] و گیروان[2](۲۰۰۴) ارائه‌شده است. از دیدگاه این دو دانشمند، هرچه خصوصیات یک افراز از شبکه از ویژگی‌های شبکه تصادفی دورتر باشد آن افراز به مفهوم اجتماع نزدیک‌تر است. این دیدگاه باعث ایجاد معیار بسیار مشهوری در زمینه اجتماع‌یابی به نام معیار پودمانگی (ماجولاریتی) شد که قادر است اختلاف ساختار یک شبکه را از شبکه تصادفی متناظر آن به‌صورت کمّی اندازه‌گیری کند (که معمولاً با نماد اختصاری Q نشان داده می‌شود). با استفاده از این تعریف می‌توان مسئله اجتماع‌یابی را به یک مسئله بهینه‌سازی با هدف بیشینه‌سازی معیار پودمانگی تبدیل کرد. این معیار در سال‌های بعد به‌طور گسترده توسط محققان و دانشمندان بسیاری در روش‌های ابتکاری مختلف اجتماع‌یابی مورداستفاده قرار گرفت نحوه محاسبه این معیار در رابطه (۱) آورده شده است که یکی از مشهورترین فرم‌های مورداستفاده از پودمانگی در مدل‌های ریاضی است.

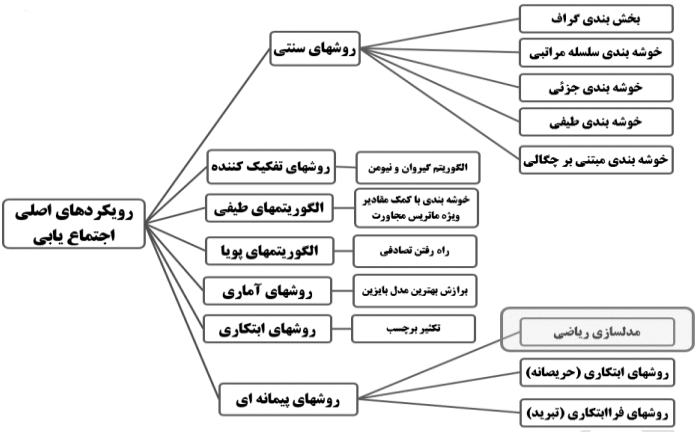
$$Q=\sum_{c=1}^{n_c}\left(\frac{l_c}{m}-\left(\frac{d_c}{2m}\right)^2\right) \qquad (۱)$$

در این رابطه n_c تعداد اجتماعات، m تعداد کل یال‌های شبکه، lc تعداد یال‌های موجود در اجتماع c و d_c مجموع درجه گره‌های موجود در اجتماع c است. طبق تعریف نیومن هرچه این مقدار بزرگ‌تر باشد نزدیکی به ساختار اجتماع واقعی بیشتر است. پس مقادیر مختلف Q این امکان را ایجاد می‌کند که بتوان افرازهای مختلف آن شبکه را باهم مقایسه کرد (Sheldon 2010).

۲-۲- رویکردهای اجتماع‌یابی در شبکه‌های بدون ویژگی (همگن)

با توجه به تعاریف و برداشت‌های متفاوت از تعریف اجتماع، محققان برای تشخیص ساختارهای اجتماعی از روش‌های متفاوتی استفاده می‌کنند. یکی از کامل‌ترین و جامع‌ترین دسته‌بندی‌ها مربوط به پژوهش فورتوناتو[3] (۲۰۱۰b) است که تقریباً بیشتر دسته‌بندی‌های موجود را پوشش می‌دهد. در *شکل ۱* هستان‌نگاری رویکردهای اجتماع‌یابی در شبکه‌های بدون ویژگی و حوزه تمرکز اصلی این مقاله به تصویر کشیده شده است. ازآنجاکه تمرکز اصلی این پژوهش بر مدل‌های بهینه‌سازی اجتماع‌یابی با روش پودمانگی است جهت رعایت اختصار در ادامه تنها به پیشینه این قسمت پرداخته می‌شود.

1. Newman
2. Girvan
3. Fortunato



شکل ۱. هستان‌نگاری رویکردهای اجتماع‌یابی و حوزه تمرکز این پژوهش

می‌توان گفت اولین پژوهش صورت گرفته در این زمینه بهینه‌سازی پودمانگی مقاله ژو[1]، تسوکا[2] و پاپاجورجیو[3] (۲۰۰۷) است. اگرچه این مدل بنیان‌گذار رویکردهای مدل‌سازی ریاضی در اجتماع‌یابی و اولین چارچوب با تعریف مشخص ریاضی است اما فقط در حل شبکه‌های غیر وزنی با ابعاد پایین تا متوسط کارایی داشته و تنها مختص شبکه‌های همگن است. در ادامه، ژو و همکاران (۲۰۱۰) مدل را برای شبکه‌هایی با تعداد گره زیاد و پیچیده گسترش دادند. بنتا[4] و همکاران (۲۰۱۲) نیز یک مدل ریاضی دومرحله‌ای برای رفع مشکل عدم همپوشانی اجتماعات و وزن‌دار نبودن شبکه ارائه کردند. از محدودیت‌های این پژوهش نیز می‌توان عدم توانایی به‌کارگیری در شبکه‌های دارای ویژگی عنوان کرد. آگاروال[5] و کمپه[6] (۲۰۰۸) الگوریتم‌های خود را بر پایه دو مسئله بهینه‌سازی با تابع هدف خطی و درجه دوم (کوآدراتیک) ارائه کردند. چن[7]، درث[8] و یو[9] (۲۰۰۸) یک مدل بر مبنای برنامه‌ریزی خطی ارائه کردند که با حذف یال‌های موجود و یافتن یک مجموعه یال ثانویه با کمترین هزینه، شبکه را به چند واحد مجزا از گراف‌های کلیک تبدیل می‌کند. در یکی از تازه‌ترین پژوهش‌ها، لی[10] (۲۰۱۳) با استفاده از معیار پودمانگی یک مدل بهینه‌سازی غیرخطی بر مبنای روش لاگرانژین فزاینده ارائه کرده است.

1. Xu
2. Tsoka
3. Papageorgiou
4. Bennetta
5. Agarwal
6. Kempe
7. Chen
8. Dress
9. Yu
10. Li

۲-۳- رویکردهای اجتماع‌یابی در شبکه‌های ویژگی‌دار (ناهمگن)

همان‌گونه که اشاره شد، تاکنون در هیچ مطالعه‌ای به بررسی هم‌زمان ویژگی‌های شبکه (مانند اطلاعات موجود در مورد صفحات وب) و توپولوژی شبکه در قالب مدل‌های ریاضی پرداخته نشده است. ازاین‌رو در این پژوهش سعی شده این فضای خالی با ارائه یک رویکرد جدید پوشانده شود. با توجه به پژوهش‌های صورت گرفته در زمینه شبکه‌های ناهمگن، به‌طورکلی رویکردهای مواجهه با گراف‌هایی با گره‌های دارای ویژگی را می‌توان به‌صورت زیر دسته‌بندی کرد:

۱. افزایش وزن به یال‌ها بر اساس مشخصه‌های گره‌ها: ایده اصلی این دسته از روش‌ها تبدیل گراف مشخصه‌دار به یک گراف موزون یکتا و استفاده از هریک از روش‌های رایج خوشه‌بندی موزون است. علاقه‌مندان به مطالعه بیشتر می‌توانند به‌عنوان‌مثال به مقالات (Neville, Adler and Jensen 2003) و (Steinhaeuser and Chawla 2008) مراجعه کنند.

۲. ترکیب خطی ابعاد ساختاری و مشخصه‌ای: در این رویکرد دقیقاً روندی برعکس نسبت به حالت قبل طی می‌شود. بدین‌صورت که اطلاعات ساختاری (یال‌ها) به‌عنوان تابع شباهت بر مبنای فاصله ذخیره می‌شود. سپس با استفاده از روش‌های سنتی مبتنی بر فاصله، عمل خوشه‌بندی صورت می‌گیرد. به‌عنوان برخی از پژوهش‌های این حوزه می‌توان به مقالات ژیا[1] و همکاران (۲۰۱۷) و فلیح[2] و همکاران (۲۰۱۷) اشاره کرد.

۳. روش‌های مبتنی بر پیمایش: ژو[3]، چنگ[4] و یو[5] (۲۰۰۹) برای شبکه‌های مشخصه‌دار از فرآیند پیمایش تصادفی استفاده و چنین عنوان کردند که هرچه مقادیر یکسان مشخصه‌های دو گره بیشتر باشد، تعداد مسیرها (یال‌ها)ی بیشتری میان آن دو گره وجود خواهد داشت. پس در این مورد می‌توان از روش پیمایش تصادفی برای محاسبه تقریب گره‌ها با توجه به هردو نوع یال‌های ساختاری و ترکیبی (حاصل از مشخصه‌ها) استفاده کرد.

۴. روش‌های مبتنی بر استنتاج آماری: در پیشینه پژوهش از این رویکرد بیشتر برای خوشه‌بندی شبکه مستندات استفاده‌شده است که در آن، متن‌ها به‌عنوان گره‌ها و لغات موجود در متن‌ها به‌عنوان معرف مشخصه‌های گره‌ها مطرح‌اند. از مقالات این حوزه می‌توان به پژوهش لیو[6]، نیکولسکیو –میزیل[7] و گریک[8] (۲۰۰۹) و ژو و همکاران (۲۰۱۲) اشاره کرد.

۵. روش‌های مبتنی بر زیر فضا: در برخی پژوهش‌ها چنین عنوان‌شده که استفاده بدون دقت از تمامی مشخصه‌ها در فرآیند خوشه‌بندی ممکن است باعث افت کیفیت خوشه‌بندی شود (پدیده نفرین بعد) (Villa-Vialaneix, Olteanu, and Cierco-Ayrolles 2013) . این پدیده باعث گسترش

1. Jia
2. Falih
3. Zhou
4. Cheng
5. Yu
6. Liu
7. Niculescu-Mizil
8. Gryc

روش‌هایی برای خوشه‌بندی به نام روش‌ها خوشه‌بندی زیر فضایی شده است. در یکی از تازه‌ترین پژوهش‌ها در این حوزه، به مسئله‌ای کاربرد-محور از واکاوی ساختارهای اجتماعات در شبکه‌های اجتماعی دارای ویژگی پرداخته شده است (Wu and Pan 2018).

۶. سایر روش‌ها (ابتکاری و فرا ابتکاری): بیشتر پژوهش‌های ارائه‌شده در این دسته به توسعه روش‌های مشهور و کارای مبتنی بر گراف پرداخته‌اند مانند (Cruz, Bothorel, and Poulet 2011). در برخی دیگر از پژوهش‌های این حوزه از مدل‌های مبتنی بر تقسیم غیر انتزاعی ماتریس برای لحاظ کردن اطلاعات معنایی گره‌ها استفاده‌شده است (Li et al. 2017; Qin et al. 2017). تمرکز دسته‌ای دیگر در این حوزه نیز بر کشف الگوهای قابل‌توجه مانند قواعد انجمنی یا ساختارهای منظم در گراف معطوف است (Pool, Bonchi, and Leeuwen 2014).

۲-۴- جمع‌بندی پیشینه مدل‌های ریاضی اجتماع‌یابی و جایگاه پژوهش حاضر

با بررسی خلاصه مهم‌ترین مقالات و نوآوری‌های ارائه‌شده در زمینه اجتماع‌یابی می‌توان دریافت که در سال‌های اخیر پژوهش‌های زیادی در زمینه بهینه‌سازی معیار پودمانگی انجام‌شده است اما تمرکز اصلی آن‌ها عمدتاً بر ارائه روش‌های ابتکاری و فرا ابتکاری در شبکه‌های بدون ویژگی بوده است و تعداد انگشت‌شماری از آن‌ها به مبحث مدل‌سازی ریاضی پرداخته‌اند. این در حالی است که مدل‌های ریاضی می‌توانند با به دست آوردن مقادیر بهینه پودمانگی به‌عنوان بهترین محک برای محاسبه کیفیت دیگر روش‌های ابتکاری استفاده شوند. از سوی دیگر، همان تعداد انگشت‌شمار مدل‌های ریاضی که در پیشینه وجود دارند نیز تنها برای اجتماع‌یابی در شبکه‌های بدون ویژگی مناسب هستند. درحالی‌که در اغلب شبکه‌های دنیای واقعی و به‌خصوص شبکه‌های اینترنتی، گره‌ها دارای برخی ویژگی‌های ذاتی هستند که لحاظ کردن آن‌ها می‌تواند به بهبود فرآیند اجتماع‌یابی کمک کند. ازاین‌رو در این پژوهش سعی شده است که این فضای خالی با ارائه یک رویکرد ریاضی برای اجتماع‌یابی در شبکه‌های اینترنتی ویژگی‌دار پوشانده شود. به‌عنوان جمع‌بندی و نتیجه‌گیری در مورد پیشینه و شکاف تحقیقاتی، مدل‌های ریاضی موجود در اجتماع‌یابی به‌طور خلاصه در *جدول ۱* ارائه و مدل پیشنهادی با آن‌ها در سطر آخر جدول مقایسه شده است.

جدول ۱. خلاصه مهم‌ترین مقالات موجود در زمینه مدل‌سازی ریاضی برای اجتماع‌یابی

مقاله	نوع مدل	شبکه	تابع هدف		تعریف	مبنای اجتماع‌یابی	
			معیار	نوع	کران	توپولوژی	ویژگی‌ها
(G. Xu, Tsoka, and Papageorgiou 2007)	MIQP*۱	غموز*۲	پود*۳	تک*۴	√	√	-
(Gang Xu et al. 2010)	MIQP	غموز	پود	تک	-	√	-
(Bennetta et al. 2012)	MIQP	موز*۵،غموز	پود	تک	-	√	-

مقاله	نوع مدل	شبکه	تابع هدف		تعریف	مبنای اجتماع‌یابی	
			معیار	نوع	کران	توپولوژی	ویژگی‌ها
(Agarwal and Kempe 2008)	IQP[*۶]	غموز	پود	تک	-	√	-
(W. Li 2013)	MIQP	غموز	پود	تک	-	√	-
مدل پیشنهادی	**IQP**	**موز،غموز**	**پود**	**تک**	√	√	√

[*۱] مختلط درجه دوم عدد صحیح، [*۲] غیر موزون، [*۳] پودمانگی، [*۴] تک‌هدفه، [*۵] موزون، [*۶] عدد صحیح درجه دوم

۳- روش پژوهش

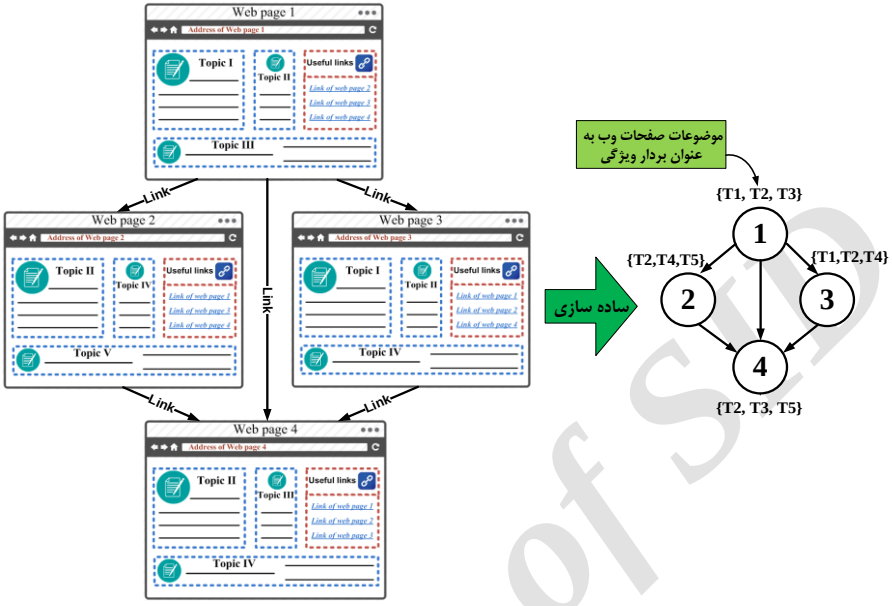
۳-۱- تعریف مسئله

همانند *شکل ۲* یک شبکه اینترنتی از صفحات وب را به‌صورت $G=(V,E,F)$ در نظر بگیرید که در آن V مجموعه وب‌سایت‌های اینترنتی (گره‌ها)، E مجموعه ارتباطات بین صفحات اینترنتی (یال‌ها)، F مجموعه موضوعات منحصربه‌فرد موجود در صفحات اینترنتی (مجموعه ویژگی‌های گره‌ها) و A ماتریس مجاورت شبکه G (ماتریس حاوی اطلاعات یال‌ها) است. هدف این پژوهش، ارائه یک مدل ریاضی برای اجتماع‌یابی است که با استفاده از G و در نظر گرفتن هم‌زمان توپولوژی شبکه (A) و ویژگی‌های گره‌ها (F)، صفحات وب را به تعدادی اجتماعِ حتی‌المقدور همگن (حاوی گره‌هایی با ویژگی‌های مشابه) و چگال (حاوی گره‌هایی با یال‌های متراکم) تخصیص دهد.

۳-۲- رویکرد ریاضی پیشنهادی برای مواجهه با ویژگی‌ها (در شبکه‌های ویژگی‌دار)

مدل‌سازی ریاضی برای اجتماع‌یابی در شبکه‌های اینترنتی ویژگی‌دار به‌طور مستقیم امری پیچیده و غیر سرراست است. ازاین‌رو در این پژوهش در رویکردی ابتکاری پیشنهادشده است که ابتدا با استفاده از یک تکنیک، ویژگی‌های صفحات وب را در ساختار شبکه اینترنتی (G) ذخیره کرد. در این رویکرد، ابتدا میزان مشابهت گره‌های موجود در دو سر یال‌ها در G طبق یک سنجه تشابه (مانند جاکارد، ضریب انطباق و غیره) محاسبه و به‌عنوان وزن به آن یال اضافه می‌شود. به بیان دقیق‌تر، فرض کنید دو گره دلخواه مانند n و e در شبکه G با یک یال به یکدیگر متصل باشند. رویه چنین است که ابتدا میزان شباهت بین این دو گره (w_{ne}) با استفاده از بردار ویژگی‌های F و یک سنجه تشابه (مانند جاکارد، ضریب انطباق و غیره) محاسبه می‌شود که عددی بین صفر و یک خواهد بود. حال این عدد به‌عنوان وزنِ یالِ متناظر بین n و e در شبکه ثانویه‌ای مانند G' در نظر گرفته می‌شود. با این کار، شبکه ویژگی‌دار G با مجموعه یال‌های غیر موزون $A=[a_{ne}]$ به شبکه ثانویه بدون ویژگی G' با مجموعه یال‌های موزونِ $A^w=[a^w_{ne}]$ تبدیل می‌شود. به بیان ریاضی، اگر در شبکه G یالی بین دو صفحه وب دلخواه مانند n و e وجود داشته باشد (یعنی $a_{ne}=1$) آنگاه وزن یال متناظر بین این دو گره در شبکه G' برابر است با $a^w_{ne}=w_{ne}$ اگر یالی بین این دو گره وجود نداشته باشد (یعنی $a_{ne}=0$) آنگاه در G' نیز یالی بین این دو وجود نخواهد داشت یعنی $a^w_{ne}=0$. حال به‌جای اینکه عملیات اجتماع‌یابی با استفاده از مدل ریاضی روی شبکه G انجام شود، این کار با

استفاده از شبکه G' صورت می‌گیرد. در بخش بعدی جزئیات مدل ریاضی پیشنهادی که برای عملیات اجتماع‌یابی استفاده می‌شود تشریح خواهد شد.



شکل ۲. نمونه‌ای از یک شبکه اینترنتی ناهمگن از صفحات وب و فرم ساده‌شده آن

۳-۳- مدل ریاضی پیشنهادی

مدل پیشنهادی در این پژوهش مشابه مدل‌های ریاضی موجود در پیشینه پژوهش برای شبکه‌های موزون است. البته با این تفاوت که برخی از متغیرهای واسطه که تعریف آن‌ها غیرضروری بوده است حذف و همچنین برای تعداد اعضای موجود در یک اجتماع نیز کران‌های کارا تعریف‌شده است (محدودیت (۴)) که در انتهای مدل ریاضی به مزیت این کار اشاره می‌شود. مجدداً یادآوری می‌شود که همان‌گونه که در بخش ۳-۲ اشاره شد، این مدل ریاضی به‌جای پیاده‌سازی بر روی شبکه اینترنتی G بر روی شبکه ثانویه G' پیاده ولی نتایج آن برای شبکه G لحاظ می‌شود.

اندیس‌ها و مجموعه‌ها:

n و e: اندیس صفحات وب (گره‌ها) $n,e=1,2,...,N$

m: اندیس اجتماع‌ها $m=1,2,...,M$

پارامترها:

a_{ne}^{w}: وزن یال متناظر بین دو صفحه وب n و e در شبکه G' که نحوه محاسبه آن در بخش ۳-۲ توضیح داده شد.

متغیرها:

Y_{nm}: متغیر تصمیم اصلی مسئله و از نوع دودویی[1] است که اگر صفحه وب n در اجتماع m قرار بگیرد مقدار یک و در غیر این صورت مقدار صفر می‌گیرد.

با توجه به توضیحات و تعاریف فوق، مدل بهینه‌سازی پودمانگی پیشنهادی برای اجتماع‌یابی صفحات وب در شبکه‌های اینترنتی ویژگی‌دار به‌صورت زیر است:

$$Max\ Q = \sum_{m=1}^{M}\left[\frac{\sum_{n=1}^{N}\sum_{\substack{e=1\\e<n}}^{N} a_{ne}^{w} Y_{nm} Y_{em}}{\sum_{n=1}^{N}\sum_{\substack{e=1\\e<n}}^{N} a_{ne}^{w}} - \left[\frac{\sum_{n=1}^{N}(\sum_{e=1}^{N} a_{ne}^{w}) Y_{nm}}{2\times\sum_{n=1}^{N}\sum_{\substack{e=1\\e<n}}^{N} a_{ne}^{w}}\right]^2\right] \quad (۲)$$

$$S.t.\ \sum_{m=1}^{M} Y_{nm} = 1\ \forall n \quad (۳)$$

$$2 \le \sum_{n=1}^{N} Y_{nm} \le N - 2(M-1)\ \forall m \quad (۴)$$

$$Y_{nm} \in \{0,1\} \quad (۵)$$

تابع هدف این مسئله از نوع غیرخطی درجه دوم و باهدف بیشینه‌سازی معیار پودمانگی است (رابطه (۲)). این رابطه حالت بسط یافتهٔ رابطه (۱) و مختص شبکه‌های موزون است که پیش‌تر توضیحات تکمیلی در مورد آن در بخش ۲-۱ و ۳-۲ ارائه شد. محدودیت (۳) تضمین می‌کند که هر صفحه وب حتماً و فقط به یک گروه (اجتماع) تخصیص یابد. ازآنجاکه طبق مفهوم و ماهیت اجتماع، وجود حداقل دو صفحه وب در هریک از آن‌ها ضروری است پس می‌توان گفت که حد پایین تعداد اعضای هر گروه برابر با ۲ است که در نتیجه حد بالای آن نیز با حداقل قرار دادن اعضای دیگر اجتماعات قابل‌محاسبه است. این حد بالا و پایین به‌طور ریاضی در محدودیت (۴) نشان داده‌شده است. البته در پیشینه پژوهش نیز ثابت‌شده است که در جواب بهینه، هیچ اجتماع تک عضوی نخواهیم داشت (Brandes et al. 2008) که دلیل دیگری بر صحت این محدودیت است. پس با ذکر این محدودیت که طبق ماهیت مسئله است می‌توان بخشی غیرضروری از فضای جواب شدنی که در آن‌ها اجتماعات تک عضوی وجود دارند (ولی جواب بهینه هیچ‌وقت در آن نواحی نخواهد بود) را حذف کرد. شایان ذکر است که کران پایین را تنها در مواردی باید در مدل لحاظ کرد که تعداد اجتماعات به عنوان ورودی از ابتدا مشخص باشد؛ در غیر اینصورت و در حالتی که هیچ اطلاعاتی در مورد تعداد بهینه اجتماعات وجود ندارد نباید حد پایین را به عنوان محدودیت در مدل ریاضی وارد کرد. درنهایت در رابطه (۵) ماهیت متغیر مسئله مشخص‌شده است. شایان‌ذکر است که در پیشینه نشان داده‌شده است که پیچیدگی محاسباتی همه مسائل اجتماع‌یابی باهدف بهینه‌سازی پودمانگی (اعم از

1. Binary

شبکه‌های موزون/غیر موزون) از نوع NP-Hard است (Brandes et al. 2008). ازآنجاکه رویکرد پیشنهادی بخش ۳-۲ موجب تبدیل (ساده‌سازی) مسئله به مسئله اجتماع‌یابی در شبکه‌های موزون می‌شود پس می‌توان نتیجه گرفت که مدل پیشنهادی این پژوهش نیز ازنظر پیچیدگی محاسباتی در این دسته قرار می‌گیرد.

۴- شیوه اعتبارسنجی مدل پیشنهادی

ابتدا قبل از هر چیز نیاز است که تعریف دقیقی از شبکه‌های محک برای رویکردهای اجتماع‌یابی ارائه شود:

منظور از شبکه‌های محک یا بنچمارک، شبکه‌هایی هستند که برای آن‌ها، مقدار بهینه پودمانگی و بالتبع گروه‌بندی بهینه گره‌ها مشخص است. ازاین‌رو به‌راحتی می‌توان با استفاده از آن‌ها، میزان کیفیت یک روش اجتماع‌یابی را اعتبارسنجی کرد. طبق این توضیحات مقدماتی، نحوه اعتبارسنجی مدل پیشنهادی در دو مرحله انجام می‌شود که در ادامه این مراحل تشریح خواهد شد.

مرحله اول) استفاده از شبکه‌های محکِ بدون ویژگی: از آنجا که شبکه‌های محک موجود در پیشینه پژوهش که برای اعتبارسنجیِ رویکردهای اجتماع‌یابی استفاده می‌شوند همگی بدون ویژگی هستند، برای نمایش اعتبار کلی و صحت عملکرد مدل پیشنهادی در وهله اول نشان داده خواهد شد که مدل پیشنهادی این پژوهش قادر است مقدار بهینه پودمانگی و گروه‌بندی صحیح گره‌ها را در برخی از شبکه‌های محک بدون ویژگی به دست آورد. بدین منظور نیاز است مدل پیشنهادی به نحوی ساده‌سازی شود که قابلیت به‌کارگیری روی چنین شبکه‌هایی را داشته باشد. این ساده‌سازی با در نظر گرفتن یک بردار ویژگی یکسان برای تمامی گره‌ها (به عبارتی لحاظ کردن وزن ۱ برای همه یال‌های موجود در شبکه) قابل انجام است. لازم به ذکر است که اگر شبکه‌های محکی وجود داشت که گره‌های آن دارای ویژگی بودند آنگاه اعتبارسنجی مدل پیشنهادی نیز با استفاده از همین شبکه‌ها قابل انجام بود اما چون چنین شبکه‌هایی هنوز در پیشینه پژوهش وجود ندارد مرحله دوم برای اعتبارسنجی مدل پیشنهادی طراحی‌شده است:

مرحله دوم) استفاده از شبکه‌های اینترنتی ویژگی‌دار: طبق توضیحاتی که در انتهای مرحله اول داده شد، شبکه محک ویژگی‌داری در پیشینه پژوهش وجود ندارد که حاوی اطلاعات بهینگی در مورد مقدار پودمانگی باشد اما شبکه‌هایی وجود دارند که گروه‌بندی واقعی گره‌ها برای آن‌ها مشخص است (توضیحات در مورد این شبکه در بخش ۵-۲ آورده شده است). ازاین‌رو در مرحله دوم (که اعتبارسنجی اصلی این پژوهش نیز محسوب می‌شود) می‌بایست اعتبار مدل در قیاس با مدل‌های پیشین روی این شبکه بررسی شود. به‌بیان‌دیگر، در قالب طراحی آزمون فرض آماری باید اثبات شود که کارایی مدل پیشنهادی این پژوهش (که ویژگی‌های ذاتی صفحات وب را در فرآیند اجتماع‌یابی لحاظ می‌کند) بیشتر از کارایی مدل‌های پیشین است (که در آن‌ها از ویژگی‌های صفحات وب صرف‌نظر می‌شود).

نکته حائز اهمیتی که در اینجا وجود دارد آن است که در اینجا نمی‌توان گفت که برای یک شبکه ویژگی‌دار، اگر مقدار تابع هدفِ (پودمانگی) مدلِ پیشنهادی بیشتر از مدل‌های قبل باشد آنگاه کارایی آن نیز بیشتر است؛ زیرا طبق رابطه (۱) و (۲)، مقدار پودمانگی تابع مستقیمی از وزن یال‌ها (ماتریس مجاورت) در شبکه است؛ درحالی‌که در مدل پیشنهادی وزن یال‌ها به‌طور هدفمند دست‌کاری و اصلاح می‌شود.

ازاین‌رو باید از گروه‌بندی واقعی گره‌ها برای اعتبارسنجی استفاده کرد که معیار اعتبارسنجی مربوطه در بخش ۴-۱ توضیح داده خواهد شد.

همچنین نکته دیگری که در اینجا مطرح می‌شود این است که مطابق بخش ۳-۲، در رویکرد پیشنهادی نیاز است یک سنجه شباهت (طبق ماهیت شبکهٔ موردمطالعه و ماهیت سنجه‌ها) برای محاسبه تشابه بین گره‌ها مورداستفاده قرار گیرد. به‌عنوان‌مثال در شبکه‌های اینترنتی که ویژگی‌ها شامل لغات و کلیدواژه‌های منحصربه‌فرد هستند (توضیحات تکمیلی در بخش ۵-۳)، باید از ضریب جاکارد که مختص چنین شبکه‌هایی است استفاده کرد و اگر از سنجه دیگری مانند ضریب انطباق که اصلاً مناسب مسائل متن کاوی نیست استفاده شود ممکن است باعث افت کارایی مدل پیشنهادی شود (برای مطالعه بیشتر در مورد خواص و حوزه کاربرد سنجه‌ها به کتاب‌های موجود در علم داده‌کاوی مراجعه نمایید). پس علاوه بر آزمون آماری درباره نوآوری اصلی مسئله (آزمون شماره (۱) در بخش ۴-۲)، این نکته نیز به‌طور آماری آزمون خواهد شد که آیا استفاده از سنجه‌های نامناسب مانند ضریب انطباق منجر به افت کیفیت اجتماعات کشف‌شده می‌شود یا خیر؟ (آزمون‌های شماره (۲) و (۳) در بخش ۴-۲).

جهت رعایت اختصار در ادامه مقاله، ابتدا نمادهای اختصاری زیر تعریف می‌شود:

۱) مدل پیشنهادی با سنجه جاکارد: رویکرد اصلی پیشنهادی این پژوهش است که به‌اختصار مدل ناهمگن موزون بر مبنای ضریب جاکارد (*JWHM*) نامیده می‌شود. در این مدل، با استفاده از سنجه جاکارد و رویکرد پیشنهادی در بخش ۳-۲، اجتماعات موجود در شبکه اینترنتی ویژگی‌دار کشف می‌شود.

۲) مدل پیشنهادی با سنجه ضریب انطباق: در این مدل که به‌اختصار مدل ناهمگن موزون بر مبنای ضریب انطباق (*MWHM*) نامیده می‌شود، طبق رویکرد پیشنهادی در بخش ۳-۲ و با استفاده از سنجه ضریب انطباق، فرآیند اجتماع‌یابی روی شبکه اینترنتی ویژگی‌دار انجام می‌شود.

۳) مدل‌های پیشین: در این‌گونه مدل‌ها که در ادامه به‌اختصار مدل همگن غیر موزون (*UWHM*) نامیده می‌شود، برای اجتماع‌یابی تنها از توپولوژی (اطلاعات یال‌های) شبکه اینترنتی استفاده و از ماتریس ویژگی‌ها صرف‌نظر می‌شود.

۴-۱- معیار ارزیابی (اعتبارسنجی)

به دلیل اهمیت موضوع مجدداً تأکید می‌شود که همان‌گونه که در رابطه (۲) مشخص است معیار پودمانگی تابعی از ماتریس مجاورت موزون است و به همین دلیل، مقایسه مقادیر پودمانگی نمی‌تواند نشان‌دهنده کارایی مدل پیشنهادی باشد؛ به همین دلیل در این پژوهش از گروه‌بندی (برچسب‌های) واقعی گره‌ها و یکی از مشهورترین معیارهای اعتبارسنجی خارجی به نام شاخص رند (RI) برای ارزیابی کارایی مدل استفاده‌شده است. این شاخص به‌طورکلی میزان مشابهت دو افراز مختلف را روی یک شبکه مورد ارزیابی قرار می‌دهد که از رابطه (۶) محاسبه می‌شود:

$$RI = \frac{TP + TN}{TP + FP + FN + TN} = \frac{TP + TN}{\binom{N}{2}} \qquad (۶)$$

در این رابطه، TP تخصیص دو گره با برچسب مشابه به اجتماع‌های مشابه، TN تخصیص دو گره با برچسب متفاوت به اجتماع‌های متفاوت، FP تخصیص دو گره متفاوت به اجتماع‌های مشابه، FN تخصیص دو گره مشابه به اجتماع‌های متفاوت و N تعداد گره‌های شبکه است. مقادیر این معیار در بازه صفر و یک قرار دارد که هرچه نزدیک‌تر به یک باشد به معنای تشابه بیشتر گروه‌های کشف‌شده با گروه‌های واقعی است.

۴-۲- فرضیه‌ها (آزمون‌های آماری) پژوهش

طبق توضیحات ارائه‌شده از ابتدای بخش چهار تاکنون، با در دست داشتن گروه‌بندی واقعی گره‌ها در یک شبکه ویژگی‌دار می‌توان نتایج اجتماع‌یابی را در دو حالت «بدون در نظر گرفتن ویژگی گره‌ها (مدل‌های پیشین)» و «با در نظر گرفتن ویژگی‌ها (مدل پیشنهادی)» مقایسه کرد. به‌عبارت‌دیگر باید اثبات شود که میانگین مقدارِ RI برای اجتماعات کشف‌شده در مدل پیشنهادی ازلحاظ آماری بیشتر از مدل‌های پیشین است. این ادعا در قالب آزمون فرض اصلی (آزمون شماره (۱) در *جدول ۲*) مطرح می‌شود. علاوه بر فرضیه فوق، طبق توضیحات ارائه‌شده در بخش چهار، دو فرضیه (۲) و (۳) نیز جهت نتیجه‌گیری‌های تکمیلی آزمون می‌شود. این سه آزمون به‌طور خلاصه در *جدول ۲* توضیح داده‌شده است.

جدول ۲. فرضیه‌های پژوهش

شماره آزمون فرض	توضیحات آزمون فرض	نماد ریاضی آزمون فرض
آزمون فرض (۱)	فرض صفر: تفاوت معناداری بین میانگین مقدار RI برای اجتماع‌های کشف‌شده در "مدل پیشنهادی با سنجه جاکارد" و "مدل‌های پیشین" وجود ندارد. فرض یک (ادعا): میانگین RI در "مدل پیشنهادی با سنجه جاکارد" بیشتر از "مدلهای پیشین" است.	$\begin{cases} H_0 : \mu_{RI}^{JWHM} = \mu_{RI}^{UWHM} \\ H_1 : \mu_{RI}^{JWHM} \geq \mu_{RI}^{UWHM} \end{cases}$
آزمون فرض (۲)	فرض صفر: تفاوت معناداری بین میانگین مقدار RI برای اجتماع‌های کشف‌شده در "مدل پیشنهادی با سنجه ضریب انطباق" و "مدل‌های پیشین" وجود ندارد. فرض یک (ادعا): میانگین RI در "مدل پیشنهادی با ضریب انطباق" بیشتر از "مدل‌های پیشین" است.	$\begin{cases} H_0 : \mu_{RI}^{MWHM} = \mu_{RI}^{UWHM} \\ H_1 : \mu_{RI}^{MWHM} \geq \mu_{RI}^{UWHM} \end{cases}$
آزمون فرض (۳)	فرض صفر: تفاوت معناداری بین میانگین مقدار RI برای اجتماع‌های کشف‌شده در "مدل پیشنهادی با سنجه جاکارد" و "مدل پیشنهادی با سنجه انطباق" وجود ندارد.	$\begin{cases} H_0 : \mu_{RI}^{JWHM} = \mu_{RI}^{MWHM} \\ H_1 : \mu_{RI}^{JWHM} \geq \mu_{RI}^{MWHM} \end{cases}$

شماره آزمون فرض	توضیحات آزمون فرض	نماد ریاضی آزمون فرض
	فرض یک (ادعا): میانگین RI در "مدل پیشنهادی با سنجه جاکارد" بیشتر از "مدل پیشنهادی با سنجه انطباق" است.	

در بخش بعدی به واکاوی نتایج و یافته‌های اجرای مدل در هر دو نوع شبکه‌های محک بدون ویژگی و شبکه اینترنتی ویژگی‌دار پرداخته می‌شود.

۵- تجزیه‌وتحلیل یافته‌ها

۵-۱- روش کلی حل

طبق توضیحات بخش چهار به‌طور خلاصه باید گفت که روش حل و نحوه اعتبارسنجی مدل پیشنهادی در دو مرحله انجام می‌شود: برای شبکه‌های محک بدون ویژگی، ابتدا اجتماعات آن شبکه با استفاده از مدل ریاضی پیشنهادی در بخش ۳-۳ و وزن یکسان یک برای تمامی یال‌ها با کمک نرم‌افزار بهینه‌سازی GAMS نسخه ۲۴ به دست می‌آید و سپس با مقدار بهینه پودمانگی (که برای آن شبکه گزارش‌شده است) مقایسه می‌شود. در شبکه‌های ویژگی‌دار نیز ابتدا اجتماعات صفحات وب در شبکه ثانویه G' (توضیحات در بخش ۳-۲) با استفاده از مدل ریاضی پیشنهادی در بخش ۳-۳ و به کمک نرم‌افزار GAMS کشف می‌شود. سپس با استفاده از گروه‌بندی واقعی گره‌ها و نرم‌افزار R نسخه ۲۳، مقدار کیفیت اجتماعات کشف‌شده ازنظر معیار RI مشخص می‌شود. برای مدل‌های پیشین نیز پروسه‌ای مشابه اما روی شبکه G (بدون توجه به ویژگی‌ها) طی و کیفیت اجتماعات ازنظر معیار RI محاسبه می‌شود. حال از آنجا که M باید به‌عنوان ورودی به مسئله مشخص شود، به ازای تنظیمات مختلف از این پارامتر در بازه مجاز (رابطه (۷))، فرضیه‌های پژوهش (بخش ۴-۲) با استفاده از نرم‌افزار MINITAB نسخه ۱۶ برای میانگین RIهای حاصله از مدل پیشنهادی و مدل‌های پیشین آزمون می‌شود. تمامی این محاسبات بر روی یک رایانه خانگی تحت سیستم‌عامل Microsoft WIN8.1 با حافظه 4GB و پردازنده Corei5 2.3~2.9GHZ صورت گرفته است. لازم به ذکر است که طبق ماهیت غیرخطی مدل پیشنهادی، هریک از مسائل چهار بار و هربار با یکی از اجراکننده‌های BARON، DICOPT، SBB و حالت پیش‌فرض نرم‌افزار GAMS اجرا و بهترین جواب به‌عنوان جواب نهایی آن مسئله گزارش شده است.

۵-۲- واکاوی یافته‌ها برای شبکه‌های محک بدون ویژگی

شبکه‌های محک بدون ویژگی که در این پژوهش مورداستفاده قرارگرفته‌اند عبارت‌اند از: کلوپ کاراته زاخاری، دلفین‌های پوزه چکشی، تیم‌های فوتبال، کتاب‌های سیاسی آمریکا و شخصیت‌های رمان بی‌نوایان. اطلاعات مربوط به این شبکه‌ها به همراه نتایج حاصل از اجرای مدل پیشنهادی بر روی آنها در *جدول* ۳ گزارش شده است.

جدول ۳. مقایسه پودمانگی به‌دست‌آمدهٔ مدل پیشنهادی با پودمانگی بهینه در شبکه‌های محک بدون ویژگی

نام شبکه	تعداد گره	تعداد یال	پودمانگی بهینه[*1]	پودمانگی به‌دست‌آمده[*2]
کلوپ کاراته زاخاری	۳۴	۷۸	۰/۴۱۲	۰/۴۱۲
دلفین‌های پوزه چکشی	۶۲	۱۵۹	۰/۵۲۹	۰/۵۲۹
تیم‌های فوتبال	۱۱۵	۶۱۳	۰/۶۰۵	۰/۶۰۵
کتاب‌های سیاسی آمریکا	۱۰۵	۴۴۱	۰/۵۲۷	۰/۵۲۷
شخصیت‌های رمان بینوایان	۷۷	۲۵۴	۰/۵۶۰	۰/۵۶۰

[*1] گزارش‌شده در پیشینه پژوهش [*2] توسط مدل پیشنهادی

همان‌گونه که در *جدول* ۳ مشاهده می‌شود مدل پیشنهادی توانسته است به مقدار بهینه پودمانگی در این شبکه‌های محک دست یابد که این امر دلیلی بر صحت کارکرد مدل و اعتبار آن در اجتماع‌یابی شبکه‌های بدون ویژگی است.

۵-۳- واکاوی یافته‌ها برای شبکه‌های اینترنتی دارای ویژگی

در این پژوهش از داده‌های شبکه اینترنتی دانشگاه کورنل (که بخشی از یک پروژه جهانی به نام گروه یادگیری متنی CMU است) استفاده‌شده است. به‌طورکلی صفحات وب موجود در این شبکه اینترنتی در پنج حوزه کلی دروس، دانشکده، دانشجویان، پروژه و کارکنان جمع‌آوری‌شده‌اند که هریک از آن‌ها با یک رنگ منحصربه‌فرد در *شکل ۳* نشان داده‌شده‌اند. همچنین در هریک از این صفحات وب، تعدادی از موضوعات درسی و دانشگاهی موردبحث قرارگرفته است که مجموع تمامی این موضوعات (۱۷۰۴ لغت منحصربه‌فرد) به‌عنوان ویژگی‌های این صفحات وب در نظر گرفته‌شده است. ساختار این شبکه شامل ۱۹۵ گره و ۳۰۴ یال است که با استفاده از نرم‌افزار تحلیل و ترسیم شبکه (Gephi) در *شکل ۳* به تصویر کشیده شده است. در این شکل، هر گره نشان‌دهنده یک صفحه وب و هریک از یال‌ها نیز نشان‌دهنده ارتباط میان دو صفحه است؛ بدین‌صورت که اگر کاربری در حین بازدید از یک صفحه از طریق یک لینک به صفحه دیگری انتقال‌یافته است یک یال بین آن دو صفحه در نظر گرفته‌شده است.

ازآنجاکه در مدل پیشنهادی می‌بایست حد بالای تعداد اجتماعات (M) را به‌عنوان ورودی مشخص کرد، مدل پیشنهادی به ازای مقادیر مختلف از این پارامتر اجرا و تحلیل شد. با استفاده از نامعادله $2 \le N - 2(M-1)$ در رابطه (۴) می‌توان چنین نتیجه گرفت که بازه مجاز برای تغییرات M برابر است با:

$$2 \le M \le \left\lfloor \frac{N}{2} \right\rfloor \qquad (۷)$$

حال به ازای مقادیر مختلف M در بازه معرفی‌شده در رابطه هفت، جواب‌های سه مدل *MWHM*، *UWHM* و *JWHM* با استفاده از حل‌کننده‌های نرم‌افزار GAMS (که در بخش محیط محاسباتی بدان‌ها اشاره شد) اجرا و اجتماع‌های به‌دست‌آمده برای محاسبه مقدار RI وارد نرم‌افزار R شدند. مقادیر مربوط به شاخص RI به ازای Mهای مختلف در *جدول ۴* و *شکل ۴* نشان داده‌شده است.

شکل ۳. شبکه اینترنتی دانشگاه کورنل

نتایج فرضیه‌های (۱) تا (۳) که با استفاده از نرم‌افزار MINITAB محاسبه‌شده است به‌طور خلاصه در جدول ۵ آورده شده است. لازم به ذکر است که در این پژوهش قبل از انجام آزمون t-زوجی، تمامی پیش‌فرض‌های استفاده از آزمون‌های آماری پارامتری مانند نرمال بودن داده‌ها آزمون و تأیید شد.

جدول ۴. مقادیر RI برای *UWHM*، *MWHM* و *JWHM* به ازای Mهای متفاوت

تعداد اجتماع‌ها	*UWHM*	*MWHM*	*JWHM*
M=۲	۰/۴۸۱	۰/۴۸	۰/۴۹۰
M=۳	۰/۵۳۴	۰/۴۹۶	۰/۵۶۸
M=۴	۰/۵۵۳	۰/۵۶۱	۰/۵۸۳
M=۵	۰/۵۷۷	۰/۵۰۹	۰/۶۱۵
M=۶	۰/۵۸۹	۰/۵۷۶	۰/۶۰۲
M=۷	۰/۵۷۴	۰/۵۴۷	۰/۶۱۹
M=۸	۰/۵۴۶	۰/۶۱۱	۰/۶۳۷
M=۹	۰/۶۰۶	۰/۵۶۸	۰/۶۱۹
M=۱۰	۰/۵۹۴	۰/۵۴۰	۰/۶۰۳
M=۲۰	۰/۶۵۷	۰/۶۲۲	۰/۶۰۲
M=۳۰	۰/۶۳۰	۰/۶۰۴	۰/۶۰۲
M=۴۰	۰/۶۱۴	۰/۶۳۲	۰/۶۷۲
M=۵۰	۰/۶۶۱	۰/۶۲۱	۰/۷۲۳
M=۶۰	۰/۶۸۲	۰/۷۰۲	۰/۶۹۸
M=۷۰	۰/۷۱۳	۰/۷۱۸	۰/۷۲۱

تعداد اجتماع‌ها	*UWHM*	*MWHM*	*JWHM*
M=۸۰	۰/۷۲۱	۰/۷۲۳	۰/۷۲۷
M=۹۰	۰/۷۲۷	۰/۷۲۷	۰/۷۲۷
M=۹۷	۰/۷۲۹	۰/۷۲۸	۰/۷۲۸

در ادامه نتایج هریک از فرضیه‌های پژوهش به‌طور مجزا موردبحث و تحلیل قرار می‌گیرد:

آزمون فرض شماره (۱): همان‌گونه که انتظار می‌رفت نتایج این آزمون نشان می‌دهد که فرض صفر با سطح اطمینان بسیار بالا (P-Value=0.01) رد و ادعای مطرح‌شده در فرض یک اثبات می‌شود. پس طبق توضیحات مندرج در سطر اول از *جدول* ۲، در نظر گرفتن ویژگی‌های صفحات اینترنتی و اجتماع‌یابی با استفاده از مدل پیشنهادی و سنجه جاکارد به‌طور معناداری توانسته است باعث بهبود میانگین RI نسبت به مدل‌های پیشین شود. این امر مؤید اعتبار و کارایی بالاتر مدل پیشنهادی این مقاله نسبت به مدل‌های پیشین دریافتن اجتماعات بهتر در شبکه اینترنتی ویژگی‌دار کورنل است.

شکل ۴. مقادیر RI برای *UWHM*، *MWHM* و *JWHM* به ازای تعداد اجتماع‌های مختلف

جدول ۵. خلاصه نتایج آزمون آماری فرضیه‌های پژوهش

شماره آزمون فرض	P-Value	نتیجه آزمون
۱	۰/۰۱	رد فرض H_0
۲	۰/۹۴	دلیلی برای رد فرض H_0 وجود ندارد
۳	۰/۰۱	رد فرض H_0

آزمون فرض شماره (۲): با توجه به مقدار بالای P-Value حاصله (۰/۹۴) در *جدول ۵* می‌توان چنین نتیجه گرفت که دلیلی بر رد فرض H_0 در آزمون دوم (سطر دوم *جدول ۲*) وجود ندارد. به‌عبارت‌دیگر همان‌گونه که از قبل پیش‌بینی می‌شد، این فرضیه که لحاظ کردن ویژگی‌های صفحات اینترنتی با استفاده از سنجه

ضریب انطباق باعث بهبود کیفیت اجتماع‌ها می‌شود پذیرفته نشد. البته با توجه به ساختار و ماهیت شبکه انتظار می‌رفت که این اتفاق رخ بدهد زیرا ضریب انطباق معیاری متقارن (سیمتریک) است و در حالتی که ویژگی‌ها ماهیت نامتقارن دارد قابلیت کاربرد ندارد.

آزمون فرض شماره (۳): با توجه به مقدار P-Value=0.01، در اینجا این ادعا که «میانگین RI به‌دست‌آمده در مدل پیشنهادی با سنجه جاکارد به‌طور معنی‌داری بهتر از سنجه ضریب انطباق است» اثبات شد. همان‌گونه که پیش‌تر گفته شد این آزمون نیز از اهداف فرعی پژوهش محسوب می‌شود و همانند آزمون شماره (۲)، تنها تأکید مجددی بر این نتیجه‌گیری است که با توجه به ماهیت ویژگی‌های موجود در شبکه اینترنتی کورنل (که ماهیتی نامتقارن دارد)، استفاده از ضریب انطباق نه توجیه علمی و نه پشتوانه آماری مناسبی خواهد داشت.

به‌عنوان جمع‌بندی باید گفت که استفاده از ویژگی‌های ذاتی صفحات وب می‌تواند منجر به کشف اجتماعات معنادارتر شود. این امر می‌تواند به‌عنوان یک گام میانی، به مدیران سایت‌های اینترنتی کمک کند که صفحات وب موجود در شبکه اینترنتی تحت مدیریت‌شان را به‌طور بهتری دسته‌بندی و مدیریت کنند. این اجتماع‌بندی می‌تواند به تخصیص بهینه پهنای باند به صفحات وب کمک کند.

۶- نتیجه‌گیری

در این پژوهش برای اولین بار یک رویکرد مدل‌سازی ریاضی برای اجتماع‌یابی صفحات وب در شبکه‌های اینترنتی دارای ویژگی ارائه شد که ایده آن، استفاده هم‌زمان از هر دو عامل توپولوژی شبکه اینترنتی و ویژگی صفحات وب برای تشخیص بهتر اجتماعات پنهان موجود در شبکه است. برای رسیدن به این هدف، ابتدا با استفاده از یک رویکرد پیشنهادی، میزان شباهت صفحات وب با استفاده از ویژگی‌های ذاتی آن‌ها و یک سنجه تشابه (مانند جاکارد) محاسبه و به‌عنوان وزن یال‌ها در فرآیند اجتماع‌یابی بکار گرفته شد. برای سنجش صحت و کارایی مدل از هر دودسته شبکه‌های بدون ویژگی و دارای ویژگی استفاده شد. در مورد شبکه‌های محک بدون ویژگی، اجتماع‌های کشف‌شده توسط مدل پیشنهادی دقیقاً منطبق بر اجتماعات کشف‌شده در ادبیات موضوع است که نشان از صحت عملکرد و دقت مدل در اجتماع‌یابی در شبکه‌های محکِ موردمطالعه است. همچنین جهت سنجش میزان کارایی مدل پیشنهادی در مورد شبکه‌های اینترنتی دارای ویژگی، فرآیند اجتماع‌یابی روی شبکه کورنل انجام و کیفیت اجتماعات با توجه به یکی از سنجه‌های مشهور اعتبارسنجی به نام شاخص RI مورد آزمون آماری قرار گرفت.

نتایج فرضیه‌های پژوهش نشان داد که در نظر گرفتن شبکه اینترنتی به‌صورت ویژگی‌دار و استفاده از معیار جاکارد به‌طور معنی‌داری باعث بهبود کیفیت اجتماع‌های کشف‌شده نسبت به روش‌های پیشین شده است. هرچند که مشاهده شد که استفاده از ضریب انطباق نتوانست به بهبود کیفیت اجتماعات کمک کند. این امر نشان می‌دهد که انتخاب سنجه شباهتِ متناسب با ماهیت شبکه، نقش حیاتی در کیفیت جواب‌های حاصل دارد. به‌عنوان پیشنهاد کار آیندگان توصیه می‌شود که بر ارائه یک روش فراابتکاری برای حل مدل پیشنهادی در شبکه‌هایی با ابعاد بالا تمرکز شود.

فهرست منابع

Agarwal, Gaurav, and David Kempe. 2008. “Modularity-Maximizing Graph Communities via Mathematical Programming.” The European Physical Journal B-Condensed Matter and Complex Systems 66 (3): 409–18.

Beiró, Mariano G., Jorge R. Busch, Sebastian P. Grynberg, and J. Ignacio Alvarez-Hamelin. 2013. “Obtaining Communities with a Fitness Growth Process.” Physica A: Statistical Mechanics and Its Applications 392 (9): 2278–93.

Bello-Orgaz, Gema, Sancho Salcedo-Sanz, and David Camacho. 2018. “A Multi-Objective Genetic Algorithm for Overlapping Community Detection Based on Edge Encoding.” Information Sciences 462 (September): 290–314.

Bennetta, Laura, Songsong Liub, Lazaros G. Papageorgioub, and Sophia Tsokaa. 2012. “A Mathematical Programming Approach to Community Structure Detection in Complex Networks.” In Symposium on Computer Aided Process Engineering, 17:20.

Brandes, U., D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner. 2008. “On Modularity Clustering.” IEEE Transactions on Knowledge and Data Engineering 20 (2): 172–88.

Chen, Jingchun, and Bo Yuan. 2006. “Detecting Functional Modules in the Yeast Protein-Protein Interaction Network.” Bioinformatics (Oxford, England) 22 (18): 2283–90.

Chen, William Y. C., Andreas W. M. Dress, and Winking Q. Yu. 2008. “Community Structures of Networks.” Mathematics in Computer Science 1 (3): 441–57.

Cruz, Juan David, Cécile Bothorel, and François Poulet. 2011. “Entropy Based Community Detection in Augmented Social Networks.” In Computational Aspects of Social Networks (cason), 2011 International Conference on, 163–68. IEEE.

Fortunato, Santo. 2010a. “Community Detection in Graphs.” Physics Reports 486 (3-5): 75–174.

Fortunato, Santo, and Darko Hric. 2016. “Community Detection in Networks: A User Guide.” Physics Reports 659 (November): 1–44.

Hric, Darko, Richard K. Darst, and Santo Fortunato. 2014. “Community Detection in Networks: Structural Communities versus Ground Truth.” Physical Review E 90 (6).

Liu, Chuang, Linan Fan, Zhou Liu, Xiang Dai, Jiamei Xu, and Baoren Chang. 2018. “Community Detection in Complex Networks by Using Membrane Algorithm.” International Journal of Modern Physics C 29 (01): 1850003.

Liu, Yan, Alexandru Niculescu-Mizil, and Wojciech Gryc. 2009. “Topic-Link LDA: Joint Models of Topic and Author Community.” In Proceedings of the 26th Annual International Conference on Machine Learning, 665–72. ACM.

Li, Wenye. 2013. “Revealing Network Communities with a Nonlinear Programming Method.” Information Sciences 229 (April): 18–28.

Li, Zhen, Zhisong Pan, Guyu Hu, Guopeng Li, and Xingyu Zhou. 2017. “Detecting Semantic Communities in Social Networks.” IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences E100.A (11): 2507–12.

Neville, Jennifer, Micah Adler, and David Jensen. 2003. “Clustering Relational Data Using Attribute and Link Information.” In Proceedings of the Text Mining and Link Analysis Workshop, 18th International Joint Conference on Artificial Intelligence, 9–15.

Newman, MEJ, and M Girvan. 2004. “Finding and Evaluating Community Structure in Networks.” Physical Review E, 1–16.

Pool, Simon, Francesco Bonchi, and Matthijs van Leeuwen. 2014. “Description-Driven Community Detection.” ACM Transactions on Intelligent Systems and Technology (TIST) 5 (2): 28.

Qin, Meng, Di Jin, Dongxiao He, Bogdan Gabrys, and Katarzyna Musial. 2017. “Adaptive Community Detection Incorporating Topology and Content in Social Networks.” In , 675–82. ACM Press.

Said, Anwar, Rabeeh Ayaz Abbasi, Onaiza Maqbool, Ali Daud, and Naif Radi Aljohani. 2018. "CC-GA: A Clustering Coefficient Based Genetic Algorithm for Detecting Communities in Social Networks." Applied Soft Computing 63 (February): 59–70.

Sheldon, Prof Ben. 2010. "Community Detection Algorithms : A Comparative Evaluation on Artificial and Real-World Networks."

Steinhaeuser, Karsten, and Nitesh V. Chawla. 2008. "Community Detection in a Large Real-World Social Network." In Social Computing, Behavioral Modeling, and Prediction, 168–75. Springer.

Villa-Vialaneix, Nathalie, Madalina Olteanu, and Christine Cierco-Ayrolles. 2013. "Carte Auto-Organisatrice Pour Graphes Étiquetés." In Atelier Fouilles de Grands Graphes (FGG)-EGC'2013, Article – numéro.

Wu, Peng, and Li Pan. 2018. "Mining Application-Aware Community Organization with Expanded Feature Subspaces from Concerned Attributes in Social Networks." Knowledge-Based Systems 139 (January): 1–12.

Xiong, Lu, Kangshun Li, and Lei Yang. 2018. "A Parallel Immune Genetic Algorithm for Community Detection in Complex Networks." International Journal of High Performance Computing and Networking 11 (3): 242–50.

Xu, Gang, Laura Bennett, Lazaros G. Papageorgiou, and Sophia Tsoka. 2010. "Module Detection in Complex Networks Using Integer Optimisation." Algorithms for Molecular Biology 5: 36.

Xu, G., S. Tsoka, and L. G. Papageorgiou. 2007. "Finding Community Structures in Complex Networks Using Mixed Integer Optimisation." The European Physical Journal B 60 (2): 231–39.

Xu, Zhiqiang, Yiping Ke, Yi Wang, Hong Cheng, and James Cheng. 2012. "A Model-Based Approach to Attributed Graph Clustering." In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 505–16. ACM.

Zhou, Yang, Hong Cheng, and Jeffrey Xu Yu. 2009. "Graph Clustering Based on Structural/attribute Similarities." *Proceedings of the VLDB Endowment* 2 (1): 718–29.

Detecting web communities in attributed internet networks using a mathematical programming approach

Esmaeil Alinezhad
Ph.D. Student in Industrial Engineering, Faculty of Industrial and Systems Engineering, Tarbiat Modares University
Babak Teimourpour
Assistant Prof., Faculty of Industrial and Systems Engineering, Tarbiat Modares University[1]

Abstract: Community detection is one of the emerging and well-known topics in the area of data mining and social network analysis, which has wide variety applications in discovering communities in real-world networks such as biological networks, internet weblogs, scientific and research websites, etc. Web community detection can especially help admins assign the optimal bandwidth to the websites of their own networks. Most of web community detection approaches only use the network topology to discover the web communities. However, the results of the most recent researches show that traditional community detection methods have to be substantially modified to consider web attributes as well as network topology. Therefore, in this paper, a mathematical programming approach is developed for community detection in attributed internet networks by simultaneously considering both network topology and node attributes. In this approach, first, similarities of web pages are calculated using node attributes and a desired similarity measure and are considered as the weight of the corresponding edges. Then, communities of the resulted weighted network will be detected by the proposed mathematical model. To validate and prove the efficiency, it is hypothesized that the detected communities of the proposed approach have a better quality than that of previous models. Experimental results demonstrate that the proposed approach has the ability to significantly improve the quality of detected web communities, when the model uses the Jaccard index. However, the results of other hypotheses indicate that the correct selection of similarity measure has a significant impact on the quality of the detected communities.

Keywords: Community detection, Internet network, Mathematical model, Modularity optimization, Network topology, Node attributes, Web pages.

1. Corresponding Author: b.teimourpour@modares.ac.ir