مقاله پژوهشی

تعبیه‌سازی شبکه‌های اجتماعی مبتنی بر کاربست روش‌های تشخیص جوامع و استخراج ویژگی‌های معنایی نهفته

محدثه طاهرپرور، فاطمه احمدی آبکناری و پیمان بیات

چکیده: هدف از تعبیه‌سازی شبکه‌های اجتماعی که اخیراً توجه زیادی را به خود جلب کرده، یادگیری نمایش در ابعاد پایین برای هر گره در شبکه با حفظ ساختار و خصوصیات شبکه است. در این مقاله، تأثیر نحوه تشخیص جوامع در حالت‌های مختلف مانند تشخیص جامعه حین یا قبل از روند پیاده‌روی تصادفی و هچنین تأثیر معنایی اطلاعات متنی هر گره بر روی تعبیه‌سازی شبکه مورد بررسی قرار گرفته و دو چارچوب اصلی با نام‌های تعبیه‌سازی شبکه آگاه به جامعه و متن و تعبیه‌سازی شبکه مبتنی بر جامعه و ویژگی‌های معنایی پیشنهاد شده است. در این مقاله، در تعبیه‌سازی شبکه آگاه به جامعه و متن، تشخیص جوامع قبل از روند پیاده‌روی تصادفی با به‌کارگیری روش غیرهمپوشان ادموت و همپوشان اگونت‌اسپلیتر انجام گرفته است. با این حال در تعبیه‌سازی شبکه مبتنی بر جامعه و ویژگی‌های معنایی، تشخیص جوامع حین رخداد پیاده‌روی تصادفی و با استفاده از مدل موضوعی جفت‌کلمه اعمال شده است. در تمامی روش‌های ارائه‌شده، تحلیل متنی مورد بررسی قرار گرفته و نهایتاً نمایش نهایی با به‌کارگیری مدل Skip-Gram در شبکه انجام می‌گردد. آزمایش‌های انجام‌شده نشان داده‌اند که روش‌های پیشنهادی این مقاله از روش‌های با نام‌های پیاده‌روی عمیق، CARE، CONE و COANE بهتر عمل کرده‌اند.

***کلیدواژه:* تعبیه‌سازی شبکه، شبکه‌های اجتماعی همپوشان، مدل‌های موضوعی جفت‌کلمه، یادگیری عمیق.**

۱- مقدمه

شبکه‌های اجتماعی در زندگی روزمره، کاربرد فزاینده‌ای یافته‌اند. شبکه‌هایی همانند شبکه‌های استنادی دانشگاهی[1]، شبکه‌های اطلاعاتی[2]، شبکه‌های زیست‌شناسی[3] و ... با هدف کشف و مدل‌سازی سیستم‌های پیچیده و متنوع به وفور مورد استفاده قرار می‌گیرند. نظر به گستردگی اندازه و ابعاد شبکه‌ها، استفاده بهینه از ویژگی‌ها و خواص شبکه‌ها، وظیفه‌ای پیچیده و چالش‌برانگیز است. در شبکه‌های با مقیاس بزرگ به دلیل پراکنده‌بودن ماتریس همسایگی، ویژگی‌های ساختاری سراسری نمی‌توانند به خوبی منعکس شوند. اخیراً تعبیه‌سازی شبکه[4] به‌صورت فزاینده‌ای توجه پژوهشگران را به خود جلب نموده است. تعبیه‌سازی شبکه با هدف یادگیری نمایش گره‌ها با بعد کم و با رعایت حفظ خصوصیات ذاتی شبکه، توانایی بالایی را در تجزیه و تحلیل و استخراج داده‌ها از شبکه نشان داده است؛ به نحوی که مسأله محاسبات فشرده مرتبط با شبکه‌های با مقیاس بزرگ را تسهیل می‌نماید. در این حوزه می‌توان از الگوریتم‌های یادگیری ماشینی برای کاربردهای مختلف بهره‌مند شد [۱] تا [۴]. روابط اطلاعاتی موجود در شبکه می‌توانند از یادگیری ماشین [۵] و [۶]، همانند طبقه‌بندی رئوس[5] [۷] و [۸]، پیش‌بینی یال[6] [۹] و ... بهره‌مند شوند.

در چند سال گذشته، تلاش‌های زیادی برای توسعه الگوریتم‌های تعبیه‌سازی شبکه اختصاص یافته که پژوهش‌های اولیه عمدتاً بر کاهش ابعاد شبکه مبتنی بر استخراج ویژگی تأکید داشتند. با این حال هزینه بالای محاسبه ماتریس همسایگی شبکه‌ها با مقیاس بزرگ، یک چالش اساسی است. اخیراً با الهام از موفقیت Word۲Vec [۱۰]، تحقیقات کارآمد زیادی در چارچوب تعبیه‌سازی شبکه مانند پیاده‌روی عمیق [۱۱]، Node۲Vec [۱۲] و LINE [۱۳] ارائه گردیده است. این روش‌های کلاسیک، عملکرد امیدوارکننده‌ای را در بسیاری از کاربردهای یادگیری ماشینی نشان داده‌اند.

روش پیاده‌روی عمیق [۱۱]، پیاده‌روی‌های تصادفی برای هر رأس را ایجاد می‌کند و از آنها به عنوان اطلاعات زمینه‌ای برای یادگیری نمایش[7]های رئوس استفاده می‌نماید. Node۲Vec [۱۲] با استفاده از دو پارامتر ازپیش‌تعیین‌شده برای کنترل روش پیاده‌روی تصادفی که یک معامله بین جستجوهای اول سطح[8] و اول عمق[9] است، پیاده‌روی عمیق را گسترش می‌دهد. پیاده‌روی عمیق و Node۲Vec با مسأله نمونه‌برداری ناکافی در شبکه‌های متراکم روبه‌رو هستند؛ بنابراین در این روش‌ها برخی از الگوهای محلی منعکس نمی‌شوند.

علاوه بر این، برخی از پژوهش‌ها در خصوص بررسی و استخراج ویژگی‌های صفات رئوس، مانند متن [۱۴] و [۱۵] یا برچسب‌ها [۱۶] و [۱۷]، در تعبیه‌سازی شبکه ارائه شده است. این تحقیقات عمدتاً بر حفظ ساختار میکروسکوپی شبکه‌ها متمرکز هستند و گاه مجبور به نادیده‌گرفتن الگوهای سراسری شبکه می‌گردند. در نتیجه، نمایش‌های یادگرفته‌شده

این مقاله در تاریخ ۱۹ اردیبهشت ماه ۱۴۰۱ دریافت و در تاریخ ۶ آبان ماه ۱۴۰۱ بازنگری شد.

محدثه طاهرپرور، گروه مهندسی کامپیوتر، واحد رشت، دانشگاه آزاد اسلامی، رشت، ایران، (email: mtaherparvar@phd.iaurasht.ac.ir).

فاطمه احمدی آبکناری (نویسنده مسئول)، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام نور، تهران، ایران، (email: fateme.abkenari@pnu.ac.ir).

پیمان بیات، گروه مهندسی کامپیوتر، واحد رشت، دانشگاه آزاد اسلامی، رشت، ایران، (email: bayat@iaurasht.ac.ir).

1. Academic Citation Network
2. Information Network
3. Biology Network
4. Network Embedding
5. Vertex Classification
6. Link Prediction
7. Representation Learning
8. Breadth-First Search
9. Depth-First Search

نمی‌توانند به خوبی در کاربردهای مختلف تجزیه و تحلیل شبکه به کار گرفته شوند.

شبکه‌های دنیای واقعی معمولاً حاوی اطلاعاتی غنی از جوامع گوناگون هستند که از اهمیت ویژه‌ای در برنامه‌های سطح جامعه برخوردار است. با توجه به جوامع موجود در شبکه، برخی از محققان، الگوریتم‌های ماژولار را برای حفظ اطلاعات جامعه معرفی کردند [۱۸] تا [۲۰]؛ اما بهینه‌سازی ماژولار به تنهایی به مسأله محدودیت وضوح[1] می‌انجامد [۲۱]. در نتیجه بررسی مفهومی محتوای متن رئوس برای تجزیه و تحلیل شبکه ضروری است و باید در تعبیه‌سازی شبکه در نظر گرفته شود. اکثر روش‌های موجود، مستقیماً کاهش ابعاد را در ماتریس‌های TF-IDF[2] انجام می‌دهند می‌دهند [۲۲] در حالی که این روش فقط قادر به اندازه‌گیری تشابه متن است و به تشابه معنایی کلمات توجه نمی‌کند [۲۳]. رئوس در یک جامعه معمولاً اطلاعات مشترکی را به اشتراک می‌گذارند. به عنوان مثال، مقالاتی در زمینه تحقیقاتی مشابه حاوی عناوین و چکیده‌های نزدیک به هم هستند. انتظار می‌رود که این شباهت‌ها برای تسهیل کاوش در ساختار شبکه و به‌دست‌آوردن نمایش‌های مؤثر به کار گرفته شود. از این رو تلفیق معناشناسی متنی و الگوهای جامعه در تعبیه‌سازی شبکه ضروری است.

در پیاده‌روی عمیق [۱۱] اثبات شده که کلمات موجود در متن و رئوس موجود در دنباله‌های پیاده‌روی تصادفی دارای یک توزیع قانونمند هستند. در این مقاله فرض شده که نقش رئوس در شبکه، مشابه کلمات در موضوعات است.

در تعبیه‌سازی شبکه آگاه به جامعه و متن (CCNE_Type۱&۲)[3] در ابتدا از روش‌های تخصیص جامعه متنوع مانند لوویان، ادموت و اگونت‌اسپلیتر برای تخصیص جوامع استفاده شده و در ادامه با توجه به دو مسأله همسایگی و جامعه، دنباله پیاده‌روی تصادفی به ازای هر گره تولید می‌شود و در نهایت دنباله‌های پیاده‌روی تولیدشده در قالب رشته‌های متنی به مدل Skip-Gram برای تولید بردارهای نمایش داده می‌شوند. متفاوت از پیاده‌روی ماژولار تولیدشده در روش CARE [۲۴]، روش CCNE_Type۱&۲ برای تشخیص جامعه علاوه بر روش غیرهم‌پوشان لوویان از روش‌های تخصیص جوامع مبتنی بر زیرگراف موتیف[4] با نام ادموت و روش هم‌پوشان اگونت‌اسپلیتر استفاده کرده است تا نشان دهد که تشخیص بهتر جوامع می‌تواند تأثیر زیادی در تعبیه‌سازی شبکه داشته باشد.

در تعبیه‌سازی شبکه مبتنی بر جامعه و ویژگی‌های معنایی (CSNE)[5] مدل موضوعی جفت‌کلمه برای توالی‌های رئوس به کار می‌رود تا توزیع جامعه برای رئوس به دست آید؛ به مرور با توزیع جوامع، مرزبندی‌هایی در شبکه ایجاد خواهد شد. این مرزبندی‌ها در CSNE برای ارزیابی اطمینان از تقسیم درست جوامع استفاده می‌شود. در ادامه، دنباله پیاده‌روی‌های تصادفی بر اساس فاصله گره‌ها از لحاظ معیار عضویت در جوامع به دست می‌آید و در نهایت دنباله‌های پیاده‌روی تولیدشده در قالب رشته‌های متنی به مدل Skip-Gram برای تولید بردارهای نمایش داده می‌شود. متفاوت از پیاده‌روی تصادفی سفت و سخت [۱۱] یا پیاده‌روی نیمه‌نظارت‌شده [۱۲]، روش پیاده‌روی تصادفی سفارشی ارائه‌شده در CSNE انعطاف‌پذیر[6] انعطاف‌پذیر[6] و خودسازگار[7] است. دامنه پیاده‌روی‌های تصادفی با مرزبندی مرزبندی محدود می‌شوند و بنابراین نهایتاً می‌توان جوامع واضح‌تری را تشخیص داد.

در روش‌های CCNE_Type۱&۲ و CSNE به ازای هر گره در دنباله پیاده‌روی‌های تصادفی، محتوای متنی رئوس مجاور در یک سند واحد جمع‌آوری شده و الگوی موجود در سند با به‌کارگیری مدل موضوعی جفت‌کلمه استخراج می‌شود. در نهایت خروجی تولیدشده از بخش تحلیل متن با بردار نمایش حاصل از مدل Skip-Gram پیوند می‌خورد. هر دو روش CCNE_Type۱&۲ و CSNE با توجه به جستجوی محلی و سراسری، مسأله نمونه‌برداری ناکافی رو نسبت به روش‌های کلاسیک تعدیل می‌کنند.

به طور خاص، نوآوری مقاله حاضر به این شرح است که روش‌های جدید تعبیه‌سازی شبکه با نام‌های تعبیه‌سازی شبکه آگاه به جامعه (CNE)[8]، تعبیه‌سازی شبکه آگاه به جامعه و متن (CCNE_Type۱&۲) (CCNE_Type۱&۲) و تعبیه‌سازی شبکه مبتنی بر جامعه و ویژگی‌های معنایی (CSNE) پیشنهاد شده است. روش‌های تعبیه‌سازی ارائه‌شده در این مقاله از دو بخش تشخیص جوامع قبل از فرایند تولید پیاده‌روی تصادفی و حین فرایند تولید پیاده‌روی تصادفی تشکیل شده است.

- تشخیص جامعه قبل از فرایند تولید پیاده‌روی تصادفی
  - در CNE و CCNE_Type۱&۲ از روش‌های تشخیص جامعه مانند روش لوویان[9]، ادموت[10] و اگونت‌اسپلیتر[11] استفاده شده است [۲۵] تا [۲۷].
  - روش‌های تخصیص جوامع لوویان و ادموت جزو روش‌های غیرهم‌پوشان و اگونت‌اسپلیتر جزو روش‌های هم‌پوشان هستند.
  - تشخیص جامعه حین فرایند تولید پیاده‌روی تصادفی
  - در CSNE از مدل موضوعی جفت‌کلمه (BTM)[12] برای استخراج جوامع موجود و تحلیل معنایی ویژگی‌های متنی در شبکه استفاده شده است.

تمام روش‌ها برای بررسی نتایج مورد مقایسه قرار گرفته‌اند. روش‌های ارائه‌شده در این مقاله در چند شبکه دنیای واقعی مانند شبکه‌های استنادی ارزیابی شده‌اند. نتایج تجربی نشان داده‌اند که الگوریتم‌های پیشنهادی در حالات مختلف در زمینه‌های تجسم شبکه، طبقه‌بندی رئوس و پیش‌بینی یال از دیگر روش‌های مبتنی بر جامعه و متن پیشی گرفته‌اند.

ادامه مقاله حاضر به شرح زیر سازماندهی شده است: بخش ۲ به طور خلاصه پژوهش‌های پیشین و الگوریتم‌های مربوط به استراتژی پیاده‌روی تصادفی، تشخیص جوامع موجود در شبکه و مدل موضوعی را ارائه داده است. در بخش ۳، الگوریتم پیشنهادی این مقاله به همراه نگاشت‌ها و توضیحات با بیان رسمی تعریف شده‌اند. بخش ۴ حاوی نتایج آزمایش‌های انجام‌شده برای تأیید اثربخشی روش‌های پیشنهادی در پژوهش حاضر است و حساسیت پارامتری روش پیشنهادی، تجزیه و تحلیل شده است. نهایتاً در بخش ۵ نیز بحث در مورد چارچوب روش‌های ارائه‌گردیده، نتیجه‌گیری و پژوهش‌های آینده خلاصه شده است.

1. Resolution Limitation
2. Term Frequency-Inverse Document Frequency
3. Community and Context Aware Network Embedding
4. Motifs
5. Community and Semantic Feature-Oriented Network Embedding
6. Flexible
7. Self-Adaptive
8. Community Aware Network Embedding
9. Louvian
10. EdMot
11. EgoNetSplitter
12. Biterm Topic Modeling

۲- پژوهش‌های پیشین

در این بخش، پژوهش‌های پیشین مربوط به یادگیری نمایش بدون نظارت[1] در خصوص گره‌های شبکه ارائه شده است. برخی از رویکردهای یادگیری با استفاده از ماتریس همسایگی سعی در حفظ همسایگی مرتبه اول گره‌ها دارند. این تحقیقات به‌عنوان روش‌های کاهش ابعاد عمل می‌کنند و بهترین ماتریس بردارهای ویژه[2] شبکه را به عنوان بردار ویژگی ویژگی شبکه‌ها می‌یابند [۲۸] تا [۳۲]. تجزیه بردار ویژه معمولاً از نظر محاسباتی گران است. علاوه بر این، روش‌های کلاسیک فقط همسایگی مرتبه اول گره‌ها را در نظر می‌گیرند و از همسایگی‌های مرتبه بالاتر و اطلاعات مربوط به جوامع استفاده نمی‌کنند. از این رو، روش‌های کلاسیک قادر به حفظ ساختار سراسری شبکه‌ها نیستند. در نتیجه، نمایش‌های یادگرفته‌شده نمی‌توانند عملکرد مناسبی را در وظایف تجزیه و تحلیل شبکه‌های متنوع ارائه دهند.

در سال‌های اخیر، یادگیری عمیق[3] به عنوان جایگزینی برای یادگیری بردار ویژگی گره‌های شبکه استفاده شده است. این روش‌ها از یادگیری عمیق برای یادگیری بردارهای نمایش استفاده کرده‌اند. آنها پیاده‌روی‌های تصادفی را به کمک استراتژی‌های مختلف جستجوی شبکه، ایجاد و آن را به عنوان اطلاعات زمینه‌ای به صورت رشته متنی در مدل Skip-Gram وارد می‌کنند.

پروزی[4] و همکاران در [۹]، تشابه بین رئوس در توالی پیاده‌روی تصادفی و کلمات در متن را تأیید کرده‌اند. آنها مدل پیاده‌روی عمیقی را پیشنهاد نمودند که با استفاده از معماری Skip-Gram بردارهای ویژگی را از توالی پیاده‌روی‌های تصادفی استخراج می‌کند. یک دنباله رأس $s=(v_1,v_2,\ldots,v_{|s|})$ به‌دست‌آمده با یک پیاده‌روی تصادفی از طریق شبکه به‌عنوان یک توالی کلمه در نظر گرفته می‌شود و هر رأس در دنباله به‌عنوان یک کلمه در نظر گرفته می‌شود. در مرحله بعدی، پیاده‌روی عمیق می‌تواند نمایش‌های شبکه را با استفاده از مدل Skip-Gram به دست آورد که هدف آن، به‌حداکثر رساندن میانگین ورود به سیستم مشاهده یک رأس است [۱۱]

$$\max_{\varnothing} \frac{1}{|s|}\sum_{i=1}^{|s|}\log \Pr(\{v_{i-w},\ldots,v_{i-1},v_{i+1},\ldots,v_{i+w}\}\,|\,v_i) \quad (1)$$

که در (۱) متن از کلمات هر دو طرف کلمه مورد نظر به اندازه پنجره w تشکیل شده است. پیاده‌روی عمیق، اولین روشی بود که از مدل Skip-Gram برای تولید بردارهای ویژگی استفاده کرد. پیاده‌روی عمیق برای ایجاد دنباله‌های پیاده‌روی تصادفی از استراتژی جستجوی اول عمق استفاده می‌کند. با وجود اینکه روش پیاده‌روی عمیق، عملکرد خوبی را در طبقه‌بندی رئوس نشان داده است اما به دلیلی درنظرنگرفتن اطلاعات جامعه نتوانست ساختار سراسری شبکه را حفظ کند. تانگ[5] و همکاران [۱۳] در LINE از همسایگی‌های مرتبه اول و دوم همراه با حفظ اطلاعات محلی برای یادگیری نمایش گره‌ها استفاده می‌کنند. در روش LINE دو تابع مستقل برای همسایگی‌های مرتبه اول و دوم تعریف شده؛ اما اطلاعات جوامع در نظر گرفته نشده است. روش LINE و پیاده‌روی عمیق قادر به یادگیری بردار نمایش برای گره‌های مرزی در شبکه نیستند.

گروور[6] و همکاران [۱۲] در Node2Vec پیاده‌روی‌های تصادفی را براساس استراتژی‌هایی مانند جستجوی اول عمق و اول سطح ایجاد می‌کنند. Node2Vec از دو پارامتر کنترلی برای بررسی دو معیار هموفیلی[7] [۳۳] و معادلات ساختاری[8] [۳۴] در شبکه‌ها استفاده می‌کند؛ اما اما در این روش هیچ تضمینی برای دستیابی به گره‌های مختلف از یک جامعه وجود ندارد. دلیل اصلی این مسأله آن است که این الگوریتم‌ها فقط تا همسایگی‌های مرتبه دوم را در نظر گرفته و توان رسیدن به گره‌هایی را که فاصله آنها از گره شروع‌کننده پیاده‌روی تصادفی بیش از دو است ندارند. لازم به ذکر است که در شبکه‌های دنیای واقعی، تعداد زیادی گره در یک جامعه وجود دارد که به صورت آشکار، فاصله آنها از هم بیشتر از دو است. بنابراین روش‌های کلاسیک در هنگام ایجاد پیاده‌روی‌های تصادفی را برای یک گره تمام اعضای جامعه‌ای که آن گره به آن تعلق دارد در نظر نمی‌گیرند.

یانگ[9] و همکاران [۳۵] نشان دادند که پیاده‌روی عمیق، معادل فاکتورسازی ماتریسی است که عنصر M_{ij} را می‌توان به عنوان اطلاعات متقابل مثبت (PMI)[10] از یک جفت متن (v_i,v_j) تفسیر کرد. آنها همچنین محتوای متنی رئوس را در چارچوب فاکتوراسیون ماتریس گنجاندند. چن[11] و همکاران [۳۶] تعبیه‌سازی شبکه اطلاعات جانبی را ارائه دادند که همسایگی معنایی[12] را برای مدل‌سازی شکل هر گره تعریف کردند؛ سپس پیاده‌روی تصادفی را برای کاوش در این همسایگی اعمال نمودند. وانگ[13] و همکاران [۳۷] یک مدل عمیق با یک معماری نیمه‌نظارتی به نام SDNE ارائه دادند که داده‌ها را به یک فضای پنهان غیرخطی نگاشت می‌کند و قادر است که هم‌زمان همسایگی مرتبه اول و مرتبه دوم را بهینه کند. این روش‌ها تنها سعی در حفظ ساختار میکروسکوپی شبکه‌ها دارند. از آنجا که این روش‌ها، عملکرد احتمال شرطی رئوس موجود در متن را به حداکثر می‌رسانند، نمایش رأس‌ها فقط به این مسأله مربوط می‌شوند که در یک پنجره از یک دنباله متنی از رئوس ظاهر شوند؛ اما اطلاعات جامعه که الگوهای سراسری شبکه را حفظ می‌کند در این روش نادیده گرفته می‌شود [۳۸]. بر اساس عملکرد این روش، بردارهای ارائه رئوس، زمانی که در جوامع مختلف با همسایگی مرتبه پایین قرار داشته باشند نزدیک‌تر به هم هستند؛ در حالی که رئوسی که در یک جامعه قرار دارند و در ساختار میکروسکوپی روابط ضعیفی دارند از همدیگر دور در نظر گرفته می‌شوند.

علاوه بر ساختار میکروسکوپی، تشخیص جامعه یکی از ویژگی‌های قابل توجه شبکه‌های پیچیده است. اگرچه برخی رویکردها ویژگی‌های سراسری یک شبکه را حفظ می‌کنند اما اطلاعات جامعه به طور کامل در تعبیه‌سازی شبکه در آنها مورد استفاده قرار نمی‌گیرد. جوامع می‌توانند محدودیت‌هایی را بر نمایش رئوس در سطح ساختاری بالاتر تحمیل کنند که باعث می‌شود نمایش رئوس زمانی که در یک جامعه هستند مشابه شوند. حتی ممکن است برخی از گره‌ها رابطه ضعیفی در ساختار میکروسکوپی داشته باشند؛ اما به این دلیل که در یک جامعه قرار دارند

1. Unsupervised
2. Eigenvector
3. Deep Learning
4. Perozzi
5. Tang
6. Grover
7. Homophily
8. Structural Equivalences
9. Yang
10. Pointwise Mutual Information
11. Chen
12. Semantical Neighborhood
13. Wang

بردارهای نمایش آنها مشابه باشد. بنابراین اطلاعات جامعه باید در تعبیه‌سازی شبکه به کار گرفته شود تا بردارهای نمایش متمایزکننده‌ای[1] برای رئوس یاد گرفته شود. لی[2] و همکاران [۳۹] یک روش تعبیه‌سازی شبکه را بر اساس الگوریتم‌های تکاملی ارائه داده‌اند که می‌تواند با بهینه‌سازی یک عملکرد چندهدفی، همسایگی و جوامع رئوس در شبکه را حفظ کند. چن[3] و همکاران [۴۰] روشی را با اطلاعات گروهی باارزش برای شبکه‌های در مقیاس بزرگ با درنظرگرفتن ساختارهای داخلی گروه‌ها و اطلاعات موجود در بین گروه‌ها پیشنهاد کردند. وانگ[4] و همکاران [۱۹] برای حفظ هر دو ساختار میکروسکوپی و مزوسکوپی[5]، یک مدل فاکتورسازی ماتریس غیرمنفی ماژولار[6] (M-NMF) معرفی کرده‌اند. عملکرد روش‌های موجود به درستی عملکرد تشخیص جامعه وابسته است که این امر یک مسأله دشوار اما بسیار بااهمیت می‌باشد.

در سال‌های اخیر، مدل‌های موضوعی آماری که هدف آنها استفاده از متن مشاهده‌شده برای استنباط توزیع موضوع پنهان است با موفقیت برای استخراج مباحث در متون پیچیده استفاده شده است [۲۳]، [۴۱] و [۴۲]. برخی از پژوهشگران از مدل‌های موضوعی در خصوص شبکه‌های همکاری نویسندگان[7] برای استنباط جامعه پژوهشی استفاده کرده‌اند [۴۳] و [۴۴]؛ در حالی که از ساختار توپولوژی شبکه در بین نویسندگان که برای پالایش موضوعات بسیار مفید است بهره‌ای نبرده‌اند. مای[8] و همکاران [۴۵] یک راه حل کلی از متن‌کاوی با ساختار شبکه به نام NetPLSA ارائه داده‌اند که بهینه‌سازی موضوعی را در شبکه نشان می‌دهد. آنها یال‌های بین اسناد را به عنوان یک قانون ایجادکننده شبکه در نظر گرفته‌اند؛ به گونه‌ای که اسناد مرتبط می‌توانند توزیع‌های مشابه موضوعی را به اشتراک بگذارند. هدف NetPLSA استفاده از اطلاعات متنی شبکه از طریق مدل موضوعی است. با این حال، این مدل بر استخراج ویژگی‌های جامعه از ویژگی‌های متن متمرکز است؛ در حالی که ساختار توپولوژیکی شبکه فقط به عنوان اطلاعات کمکی استفاده می‌شود. از این رو نمی‌تواند هنگام مدیریت اطلاعات متنی نویزی یا ناقص، بردارهای نمایش خوبی را به دست آورد.

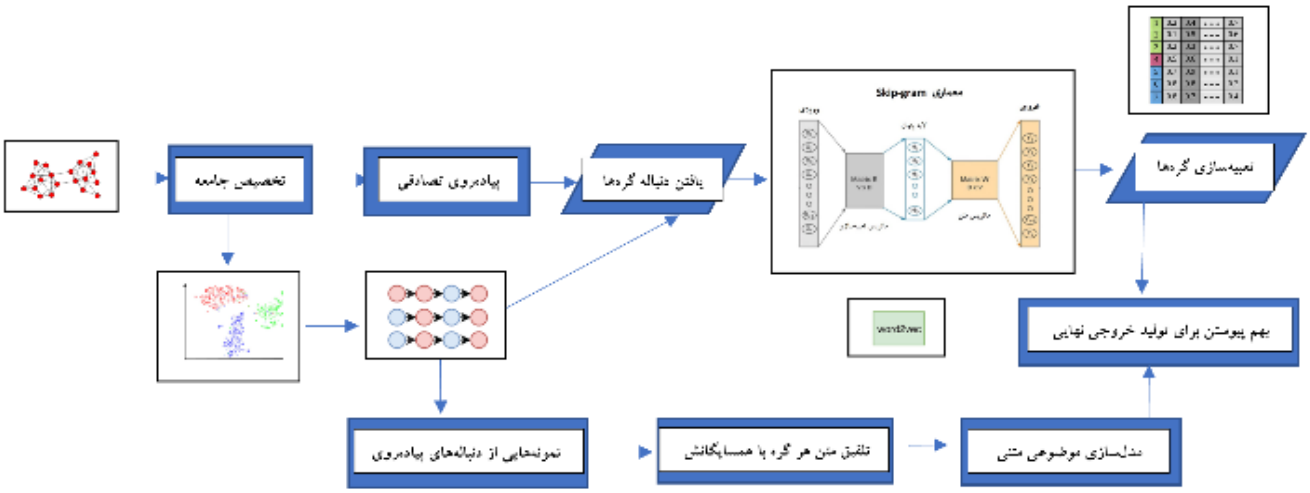
شی[9] و همکاران [۴۶] یک روش تعبیه‌سازی انتشار[10] شبکه را ارائه داده‌اند که هدفشان حل محدودیت‌هایی همچون تمایل به انتخاب گره‌ها با درجه بالا بوده است؛ اما بی‌توجهی به ساختار سراسری شبکه‌های بسیار پیچیده مورد بررسی آنها در پیاده‌روی تصادفی در این پژوهش مشهود است. چن[11] و همکاران [۴۷] یک مدل قابل تعمیم ارائه داده‌اند که هم از اطلاعات یال و هم از اطلاعات مرکزیت[12] گره‌ها برای یادگیری نمایش‌های برداری با ابعاد کم استفاده می‌کند که می‌تواند به حفظ اطلاعات مختلف مرکزیت رئوس اهتمام ورزد. ژائو[13] و همکاران [۴۸] یک چارچوبی یکپارچه را برای توصیه‌های اجتماعی و رفتاری با تعبیه‌سازی شبکه ارائه داده‌اند و یک رویکرد تعبیه‌سازی شبکه مشترک را به عنوان یک مرحله قبل از آموزش برای نمایش‌های پنهان کاربران معرفی کرده‌اند. لی[14] و همکاران [۴۹] یک مدل تعبیه‌سازی شبکه بدون نظارت را برای رمزگذاری اطلاعات رابطه یال‌ها ارائه داده‌اند. بنابراین نمایش ویژگی‌های رئوس می‌تواند بیشتر ضبط شود. وو[15] و همکاران [۵۰] برای یادگیری نمایش‌های شبکه برای کاربردهای خاص، یک مدل LSTM توجه دوگانه چندوظیفه‌ای[16] را ارائه کرده‌اند که می‌تواند ساختار، محتوا و اطلاعات برچسب را ضبط و سپس نمایش‌های رئوس را با توجه به وظیفه پایین‌دست تنظیم کند. یوان[17] و همکاران [۵۱] الگوریتمی را با نام COANE ارائه کرده‌اند که از مدل‌سازی موضوعی LDA برای تشخیص جامعه رئوس موجود در شبکه و ساخت دنباله پیاده‌روی‌های تصادفی استفاده می‌کند. کیخا[18] و همکاران [۲۴] الگوریتمی با نام CARE ارائه کرده‌اند که از روش تشخیص جامعه لوویان برای تشخیص جوامع گره‌های موجود در شبکه و ساخت دنباله پیاده‌روی‌های تصادفی استفاده کرده است. الگوریتمی CARE از روش لوویان برای کشف جوامع استفاده می‌کند. این روش در ساخت پیاده‌روی تصادفی علاوه بر رابطه هموفیلی صریح یعنی انتخاب گره‌ها بین همسایه‌های مرتبه اول و دوم به گره‌هایی که دارای رابطه هموفیلی ضمنی هستند نیز شانس انتخاب‌شدن را می‌دهد. گره‌هایی که در یک جامعه قرار دارند ولی از لحاظ همسایگی در مرتبه اول یا دوم نیستند دارای رابطه هموفیلی ضمنی می‌باشند. روش لوویان جزو روش‌های غیرهم‌پوشان و ازهم‌گسسته است [۲۵].

الگوریتم CARE روش پیاده‌روی تصادفی آگاه به جامعه است و برای تعبیه‌سازی شبکه به ساختار محلی و سراسری شبکه توجه می‌کند؛ اما نسبت به خواص و ویژگی‌های متنی گره‌ها بی‌توجه است. تحلیل معنای متن موجود در گره‌ها می‌تواند خط مشی در خصوص تجزیه و تحلیل کاربردهای شبکه باشد. گره‌هایی که در یک جامعه قرار دارند بر روی هم تأثیرگذار مثبت یا منفی هستند و از لحاظ تحلیل معنایی اطلاعات متنی به هم نزدیک بوده و به موضوعات مشابهی اشاره دارند. پس برای بهبود عملکرد تعبیه‌سازی شبکه بهتر است که علاوه بر ساختارهای محلی و سراسری به اطلاعات متنی گره‌ها در شبکه نیز توجه داشت.

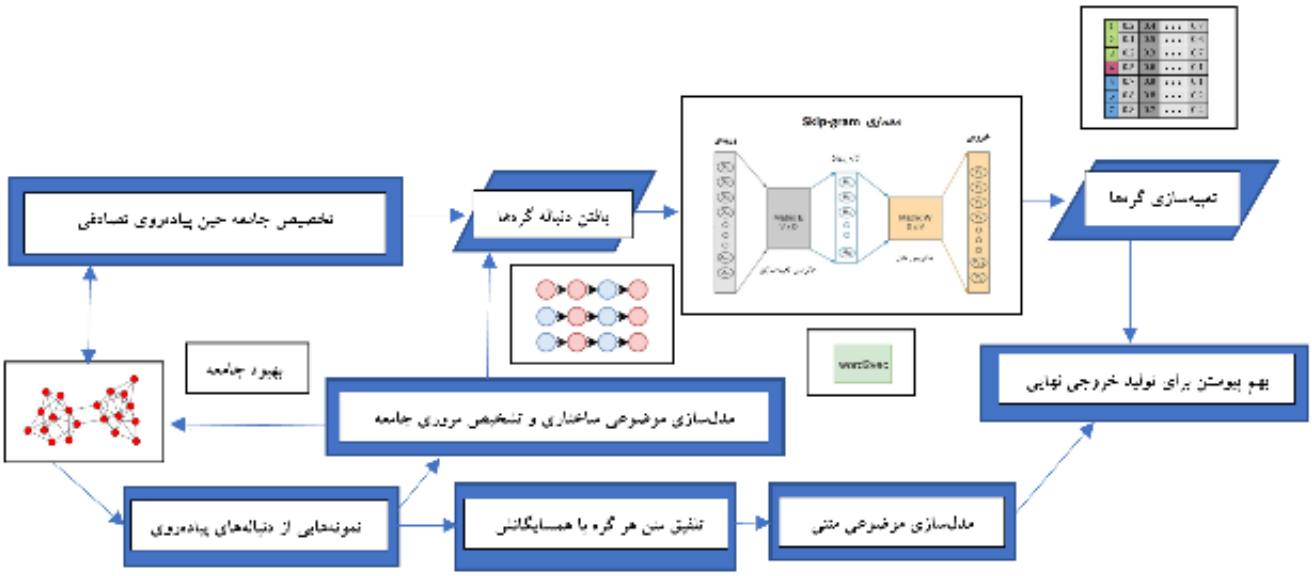
در این مقاله، تخصیص جوامع در تعبیه‌سازی شبکه آگاه به جامعه و متن (CCNE) قبل از فرایند پیاده‌روی تصادفی رخ می‌دهد و همچنین علاوه بر لوویان از روش‌هایی دیگر با نام‌های اگونت‌اسپلیتر و ادموت برای تخصیص جامعه استفاده شده است. اگونت‌اسپلیتر به صورت هم‌پوشان و ادموت به صورت غیرهم‌پوشان به تشخیص جوامع می‌پردازند. در تخصیص جامعه در روش تعبیه‌سازی شبکه مبتنی بر جامعه و ویژگی‌های معنایی (CSNE) پیاده‌روی تصادفی با استفاده از روش مدل موضوعی جفت‌کلمه صورت می‌پذیرد. این روش در چندین مرحله ساختار محلی و سراسری گره‌ها در شبکه را با ارائه مرزبندی جوامع حفظ می‌کند.

در هر دو روش CCNE و CSNE، مسأله نمونه‌برداری ناکافی از شبکه‌های متراکم- که از چالش‌های مهم در تعبیه‌سازی شبکه است- با استفاده از نمونه‌برداری‌های محلی و سراسری در شبکه تعدیل می‌گردند. در بخش بعد الگوریتم‌های پیشنهادی این پژوهش شرح داده می‌شوند.

1. Discriminative
2. Li
3. Chen
4. Wang
5. Mesoscopic
6. Modularized Non-Negative Matrix Factorization
7. Co-Author Network
8. Mei
9. Shi
10. Diffusion
11. Chen
12. Centrality
13. Zhao
14. Li
15. Wu
16. Multi-Task Dual Attention LSTM Model
17. Yuan
18. Keikha



شکل ۱: فلوچارت روش‌های پیشنهادی مبتنی بر تشخیص جوامع قبل از تولید پیاده‌روی تصادفی.



شکل ۲: فلوچارت روش پیشنهادی مبتنی بر تشخیص جوامع حین تولید پیاده‌روی تصادفی.

۳- تعبیه‌سازی شبکه

اطلاعات جامعه و تحلیل معنایی خصوصیات متنی مختص به هر گره، جزو ویژگی‌های مهم شبکه‌های اجتماعی است که ساختار محلی و سراسری شبکه را حفظ می‌کند. حفظ هر دو ساختار محلی و سراسری گره‌ها در شبکه در پژوهش‌های انجام‌شده تا کنون، کمتر مورد توجه قرار گرفته است. در این مقاله، الگوریتم‌های جدیدی برای تعبیه‌سازی ساختار شبکه معرفی می‌گردد که مبتنی بر اطلاعات جوامع و تحلیل معنایی بردارهای ویژگی گره‌های شبکه عمل می‌کنند.

فرض کنید که $G=(V,E,T)$ یک گراف ویژگی[1] است و در آن V مجموعه‌ای از رئوس، $E\subseteq V\times V$ یال‌ها روابط بین رئوس و T محتوای متن رئوس است. به طور خاص، اطلاعات متنی هر رأس $v\in V$ مربوط به توالی کلمه $T_v=(W_1,W_2,\ldots,W_n)$ است که در آن $n_v=|T_v|$ می‌باشد. تعبیه‌سازی شبکه، سعی در ایجاد یک ماتریس از ویژگی‌ها با ابعاد کم با عنوان $\Phi\in R^{|V|\times d}$ برای شبکه دارد به طوری که $d\ll|V|$ ابعاد فضای نمایش پنهان را مشخص می‌کند. از بردارهای نمایش به‌دست‌آمده می‌توان در برنامه‌های کاربردی تجزیه و تحلیل شبکه استفاده کرد. هدف این پژوهش، یافتن یک تابع نگاشت $\Phi\in R^{|V|\times d}$ است به نحوی که d بعد نمایش هر گره در شبکه است. همچنین برای به‌دست‌آوردن بهترین عملکرد نگاشت تابع Φ از مدل Skip-Gram استفاده شده است.

تشخیص جوامع

همان طور که قبلاً ذکر شد جوامع نقش تعیین‌کننده‌ای در تعبیه‌سازی شبکه دارند. اولویت الگوریتم‌های کلاسیک برای انتخاب گره بعدی در پیاده‌روی تصادفی، همسایگی مرتبه اول و دوم گره‌ها است. اگر به معیارهای جامعه‌محور در شبکه‌ها توجه شود مشاهده می‌گردد که گره‌های عضو یک جامعه دارای خط مشی و طرز فکر یکسانی هستند. از این رو باید علاوه بر بررسی همسایگی گره‌ها در شبکه به گره‌هایی که در یک جامعه یا جوامع نزدیک به هم هستند، اولویت بالاتری برای انتخاب به عنوان گره‌های بعدی در پیاده‌روی تصادفی تعلق گیرد.

در این پژوهش در فرایند تولید پیاده‌روی‌های تصادفی، دو روش برای تشخیص جوامع در شبکه در نظر گرفتیم: ۱) تشخیص جوامع قبل از تولید پیاده‌روی تصادفی و ۲) تشخیص جوامع حین تولید پیاده‌روی تصادفی. در شکل‌های ۱ و ۲ فلوچارت روش‌های ارائه‌شده آمده است.

1. Attributed Network

Algorithm 1: Framework of CCNE
Input: graph: $G(V,E,T)$; Window size: W ;
representation dimension d ;
walks per vertex: γ ; walk length: l .
Output: matrix of network representations: $\phi \in R^{|v|\times d}$
1: Com = Community Detection(G)
2: Sample ϕ from $u^{|v|\times d}$
3: **for** $i = 0$ **to** γ **do**
4: $v = shuffle(v)$
5: **for** each vertex $v \in V$ **do**
6: $W_{v_i} = RandomWalk(G, v, Com, l)$
7: $D_v = ContextAggreation(G, v)$
8: $SkipGram(\phi, S_v, W)$
9: **end for**
10: $\Pr_t = BTM_t(D)$
11: **end for**
12: $\phi = \phi \oplus \Pr_t$
13: **return** ϕ

شکل ۳: الگوریتم CCNE.

Algorithm 2: RandomWalk
Input: graph: $G(V,E,T)$; Source node of RW v_i ;
Nodes belong to the same community with v_i Com_{Vi} ;
walk length: l ; Random variable to Select from
neighbors or same community members: α .
Output: A path with max length l .
1: Initialize RW with v_i
2: **While** $length(path) < l$
3: **if** current node has neighbors
4: **if** $random(0,1) < \alpha$
5: select v_i at random from v_j 's neighbors
6: **else**
7: select v_i at random from members v_j 's communities
8: **else**
9: backtrack in the path and select the last node which has neighbors that are in the path
10: **end While**.

شکل ۴: پیاده‌روی تصادفی برای الگوریتم CCNE.

۳-۱ تشخیص جوامع قبل از تولید پیاده‌روی تصادفی

نظر به اینکه الگوریتم‌های موجود در تشخیص جوامع در شبکه به دو دسته کلی تشخیص جوامع هم‌پوشان و تشخیص جوامع غیرهم‌پوشان تقسیم‌بندی می‌گردند، از این رو در ادامه سه روش تشخیص جوامع، شامل روش‌های لوویان و ادموت از نوع هم‌پوشان و اگونت‌اسپلیتر از نوع غیر هم‌پوشان در الگوریتم پیشنهادی مورد استفاده قرار گرفته است.

در ادامه به معرفی الگوریتم CCNE می‌پردازیم که فرایند تشخیص جوامع را قبل از تولید پیاده‌روی تصادفی انجام داده و به صورت موازی با تولید دنباله‌های پیاده‌روی تصادفی، تحلیل معنایی متنی گره‌های موجود در دنباله‌ها را نیز انجام می‌دهد و نهایتاً بردارهای به‌دست‌آمده از دنباله‌های پیاده‌روی تصادفی و تحلیل معنایی را با هم برای تحلیل نهایی شبکه پیوند می‌دهد.

CCNE: الگوریتم تعبیه‌سازی شبکه آگاه به جامعه و متن

در این بخش، الگوریتم تعبیه‌سازی شبکه آگاه به جامعه و متن (CCNE) ارائه گردیده که هدف الگوریتم، ایجاد بردارهای نمایش برای گره‌های شبکه است. در CCNE تشخیص جوامع قبل از تولید پیاده‌روی تصادفی انجام می‌شود و انتخاب گره‌ها برای دنباله پیاده‌روی تصادفی به صورت هموفیلی صریح و ضمنی است. همچنین تحلیل معنایی متن گره‌ها به صورت موازی با ساخت دنباله‌های پیاده‌روی انجام می‌شود و در نهایت بردارهای تولیدشده از بخش‌های پیاده‌روی تصادفی و تحلیل معنایی متن به هم پیوند خورده و بردار نهایی برای تعبیه‌سازی شبکه به دست می‌آید. الگوریتم ۱ مراحل CCNE را نشان می‌دهد.

همان طور که در شکل ۳ آمده است، ابتدا جوامع گراف ورودی تشخیص داده می‌شود. از روش اگونت‌اسپلیتر در الگوریتم CCNE-Ego، روش ادموت در الگوریتم CCNE-EdMot و روش لوویان در الگوریتم CCNE-Louvain برای کشف جوامع استفاده می‌شوند. قبل از یادگیری بردارهای نمایش برای گره‌های شبکه، ماتریس U به طور تصادفی تولید می‌شود تا بردار نمایش گره‌های بخش بعدی تولید گردد. اکنون الگوریتم، قادر به یادگیری بردارهای نمایش نهایی در خطوط ۳ تا ۱۱ است. قبل از تکرار گره‌های شبکه در خط ۴ برای جلوگیری از تأثیر گره‌های رؤیت‌شده در ارائه نهایی، گره‌ها جابه‌جا می‌گردند. هسته اصلی تعبیه‌سازی شبکه در ارائه نهایی، گره‌ها جابه‌جا می‌گردند. هسته اصلی تعبیه‌سازی شبکه در خط ۶ انجام می‌شود که در آنجا پیاده‌روی تصادفی برای گره انتخاب‌شده تولید می‌شود. خط ۷ بخش تحلیل معنایی متن گره‌ها را نشان می‌دهد که توضیح آن در بخش الگوریتم ۳ آمده است. سرانجام از مسیرهای تولیدشده و تحلیل معنای متن گره‌ها برای به‌روزرسانی ارائه گره‌ها در خط ۱۲ استفاده می‌شود.

شکل ۴ الگوریتم پیاده‌روی تصادفی را برای الگوریتم CCNE نشان می‌دهد. پیاده‌روی تصادفی که از گره v شروع می‌شود توسط w_V نشان داده شده است. می‌توان یک دنباله پیاده‌روی تصادفی برای گره v را با برخی از متغیرهای تصادفی $w_V^1, w_V^2, \ldots, w_V^K$ نشان داد به طوری که در آن w_V^{k+1} گرهی است که به طور تصادفی از بین همسایگان فوری یا گره‌هایی انتخاب می‌شود که در یک جامعه با گره k قرار دارد.

برای ایجاد یک پیاده‌روی تصادفی سفارشی که از گره v شروع شده است ابتدا همه همسایگان آن گره استخراج می‌شوند. سپس یک متغیر تصادفی r بین ۰ و ۱ ایجاد می‌گردد. اگر r از α کمتر باشد یک گره به طور تصادفی از همسایگان نزدیک انتخاب می‌شود و در غیر این صورت، گره از بین گره‌هایی انتخاب می‌شود که در همان جامعه با گره kام از w_V قرار دارند.

روش پیشنهادی، پیاده‌روی تصادفی را با درنظرگرفتن اطلاعات محلی و سراسری از شبکه داده‌شده استخراج می‌کند. پیاده‌روی‌های تصادفی به صورت مستقل ساخته می‌شوند و از این رو الگوریتم حاضر، توانایی موازی‌سازی دارد تا روند تعبیه‌سازی، سرعت یابد. علاوه بر این اگر برخی از گره‌های جدید به شبکه اضافه یا حذف شود، پیاده‌روی تصادفی فقط برای گره‌های جدید اجرا می‌گردد.

شکل ۵ الگوریتم جمع‌بندی متن رئوس را نشان می‌دهد. برای کاهش انحراف بین توزیع جامعه خلفی و توزیع واقعی جامعه، باید این نکته را در نظر گرفت که طول یک سند از تعداد کل اسناد کمتر نیست [۴۸]. برای این بخش، دو نوع الگوریتم جمع‌بندی متن ارائه شده است. نوع اول در شکل ۵ با عنوان تجمیع متن نوع اول و نوع دوم در شکل ۶ با نام تجمیع متن نوع دوم نشان داده شده است.

در شکل ۵ در خط ۵، انتخاب رئوس بعدی متکی به استراتژی انتخاب تصادفی از بین همسایگان هر گره است. در خط ۷، انتخاب گره بعدی به صورت تصادفی از بین گره‌هایی صورت می‌گیرد که در یک جامعه با گره قبلی قرار داشته باشند. در شکل ۶ خط ۵، هسته اصلی این نوع است و

Algorithm 3: Context Aggregation_type1
Input: graph $G(V,E,T)$;
number of communities: k;
probability of vertices belonging to structure-based communities: P_{r_s}.
Output: the contextual text information: D_v

```
Initialize D_v with T_v
While length(D_v) < γ.log|v| do
    if current vertex has neighbors, then
        if ( 0 < random <= alpha ) then
            u = select a vertex u from neighbors based on random selection
        else if ( alpha < random < 1 ) then
            u = select a vertex u at random from members current vertex's communities
    else
        break
    D_v = D_v ⊕ T_u
    end if
end while
return D_v
```

شکل ۵: جمع‌بندی متن رئوس نوع اول برای الگوریتم CCNE.

Algorithm 3: Context Aggregation_type2
Input: graph: $G(V,E,T)$; number of communities: k;
probability of vertices belonging to structure-based communities: P_{r_s}.
Output: the contextual text information: D_v.

```
Initialize D_v with T_v
While length(D_v) < γ log|v| do
    if current vertex has neighbors, then
        u = select a vertex u from neighbors based on random selection if both vertexes are in same community
    else
        break
        D_v = D_v ⊕ T_u
    end if
end while
return D_v
```

شکل ۶: جمع‌بندی متن رئوس نوع دوم برای الگوریتم CCNE.

گره بعدی را بر اساس همسایگی و این نکته که هر دو گره به یک جامعه تعلق داشته باشند، انتخاب می‌کند. اگر بخواهیم این دو نوع انتخاب را با یکدیگر مقایسه کنیم در نوع اول، زمانی که شبکه بسیار بزرگ باشد بالطبع، جوامع نیز دارای اندازه‌های بزرگ‌تری خواهند بود و چون انتخاب گره‌ها به صورت تصادفی صورت می‌گیرد، ممکن است که گره‌های انتخاب‌شده از لحاظ متنی تشابه کمتری نسبت به هم داشته باشند. اما در نوع دوم، هرچه اندازه شبکه بزرگ باشد به دلیل انتخاب گره‌ها برای تجمیع متن از همسایه‌های درجه یک و دو و به شرط تعلق به یک جامعه، گره‌های انتخاب‌شده از لحاظ متنی دارای تشابه بیشتری خواهند بود.

۳-۲ تشخیص جوامع حین تولید پیاده‌روی تصادفی

در تشخیص جوامع حین تولید پیاده‌روی از آنجا که هر رأس در شبکه با یک احتمال خاص به جوامع منفرد یا چندگانه تعلق دارد، یک دنباله کامل را می‌توان ترکیبی از جوامع مختلف در نظر گرفت. در این بخش از روش تعبیه‌سازی شبکه مبتنی بر ویژگی‌های معنایی و جامعه استفاده می‌کنیم که در ادامه به شرح آنها خواهیم پرداخت.

۳-۲-۱ توزیع جامعه مبتنی بر ساختار و متن بر اساس مدل موضوعی جفت‌کلمه

در این بخش برای استخراج الگوهای سراسری جامعه، ترکیبی از مدل موضوعی جفت‌کلمه و یک دنباله پیاده‌روی تصادفی مبتنی بر مرزبندی را ارائه می‌کنیم. به منظور توضیح بهتر نحوه عملکرد مدل موضوعی جفت‌کلمه در شناسایی جامعه و تولید پیاده‌روی تصادفی در شبکه، مدل به جای کلمات و موضوعات از طریق رئوس و اجتماعات توصیف می‌شود [۵۲]. روند مدل‌سازی مدل موضوعی جفت‌کلمه مربوط به توزیع مشروط متغیرهای پنهان و متغیرهای مشاهده‌شده در (۲) آمده است

$$p_r(v_i|S)=\sum_{j=1}^{k}P_r(v_i|c_i=j)P_r(c_i=j|S) \qquad (۲)$$

که $P_r(v_i|S)$ احتمال مشاهده رأس v_i در توالی S، $P_r(v_i|c_i=j)$ احتمال رأس v_i در جامعه نهفته j و $P_r(c_i=j|S)$ احتمال انتخاب یک رأس از جامعه j در توالی S است. مدل موضوعی جفت‌کلمه یا BTM توزیع جامعه- رأس $P_r(v|c)$ و توزیع جامعه- اجتماع $P_r(v|S)$ را از تعداد معینی از جوامع k با استفاده از نمونه‌گیری گیبس[1] تخمین می‌زند. این روش به طور تصادفی، یک جامعه محلی را به هر رأس v_i از توالی فعلی S اختصاص می‌دهد. سپس هر رأس v_i مورد بررسی قرار می‌گیرد و جامعه آن j بر اساس احتمال $P_r(c_i=j|v_i,S,c_{-i})$ با استفاده از (۳) تا زمانی که پارامترهای مدل موضوعی جفت‌کلمه همگرا شوند به‌روز می‌شود [۵۲]

$$P_r(c_i=j|v_i,S,c_{-i})\propto \frac{N_{vc}(v_i,j)+\beta}{\sum_v N_{vc}(v,j)+|v|\beta}\cdot\frac{N_{vc}(v,j)+\alpha}{\sum_c N_{sc}(S,c)+k\alpha} \qquad (۳)$$

در اینجا $N_{vc}(v_i,j)$ شماری از همه انتسابات رأس- جامعه است و $N_{sc}(S,c)$ شماری از انتسابات جامعه- توالی را نشان می‌دهد. علاوه بر این، c_{-i} تمام انتسابات فوق را نشان می‌دهد. به‌جز انتساب فعلی t_i برای رأس v_i و پارامترهای هایپر α و β که به عنوان فاکتورهای هموارسازی[2] عمل می‌کنند. بنابراین توزیع شرطی $P_r=(v_i|c_i=j)$ و $P_r(c_i=j|S)$ را می‌توان به شرح زیر برآورد کرد

$$P_r(v_i|c_i=j)=\frac{N_{vc}(v_i,j)+\beta}{\sum_v N_{vc}(v,j)+|v|\beta} \qquad (۴)$$

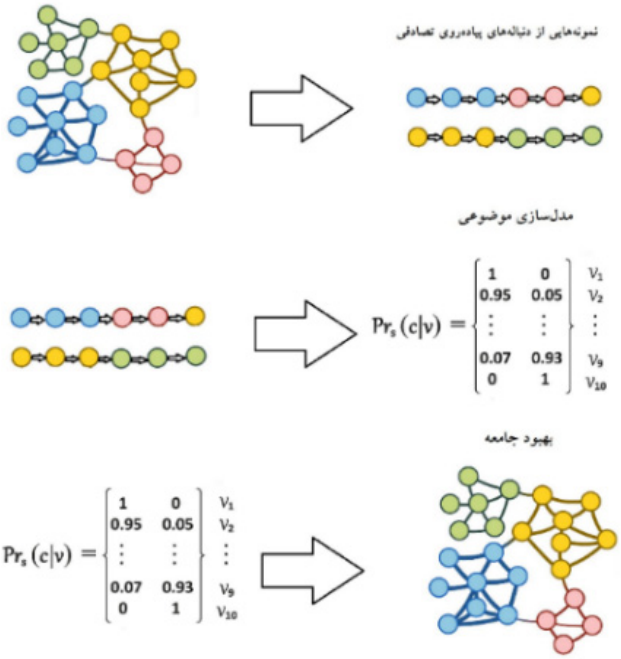
$$P_r(c_i=j|S)=\frac{N_{sc}(S,j)+\alpha}{\sum_c N_{sc}(S,c)+k\alpha} \qquad (۵)$$

با استفاده از دو توزیع شرطی بالا، این روش می‌تواند احتمال آن را که یک رأس در دنباله گره‌ها برای تجمیع اسناد متعلق به هر جامعه باشد تخمین بزند. بدین ترتیب ساختار جامعه شبکه در این خصوص شناسایی می‌شود.

۳-۲-۲ الگوریتم تعبیه‌سازی شبکه مبتنی بر ویژگی‌های معنایی و جامعه

در الگوریتم تعبیه‌سازی شبکه مبتنی بر ویژگی‌های معنایی و جامعه CSNE، برای یادگیری توزیع جامعه مبتنی بر ساختار[3] و متن[4] به ترتیب

1. Gibbs
2. Smoothing
3. Structure-Based
4. Text-Based



شکل ۷: توصیف گرافیکی پیاده‌روی تصادفی مبتنی بر مرزبندی بر اساس مدل موضعی جفت‌کلمه.

از دو مدل موضوعی جفت‌کلمه با بیش از پارامترهای مختلف استفاده می‌شود و پس از آن، نمایش‌های رئوس را می‌توان از طریق مدل Skip-Gram آموخت.

از پیاده‌روی تصادفی اصلی در شبکه برای ساختن دنباله‌های پیاده‌روی تصادفی استفاده می‌شود و توالی‌های رئوس تولیدشده در طی این روش برای به‌روزرسانی مداوم به مدل موضوعی جفت‌کلمه مبتنی بر ساختار وارد می‌شوند. پس از تعداد مشخصی از تکرارها، الگوهای مقدماتی سراسری و جوامع مبتنی بر ساختار آموخته می‌شوند. در این فرایند از Skip-Gram برای حداکثرکردن تابع احتمال شرطی رئوس استفاده می‌شود و بنابراین بردارهای ارائه رئوس از این توالی‌های پیاده‌روی تصادفی به دست می‌آیند.

اکثر روش‌های قبلی برای مدل‌سازی ساختار همسایگی یک گره فقط از همسایگی مرتبه اول و دوم استفاده می‌کردند. در مقابل، در این بخش علاوه بر حفظ ساختار همسایگی گره‌ها به مرور با تشکیل مرزبندی‌هایی بین گره‌ها با هدف ایجاد جوامع در شبکه، گره‌هایی دارای اولویت بالاتری هستند که هم در همسایگی گره مبدأ باشند و هم پارامترهای مختص جوامع آنها به هم نزدیک باشد.

بر اساس شکل ۷ در چند مرحله اول، از پیاده‌روی عمیق در شبکه استفاده می‌شود و دنباله‌های رئوس تولیدشده در طی این روش برای به‌روزرسانی مداوم، به مدل موضوعی جفت‌کلمه مبتنی بر ساختار وارد می‌شوند. پس از تعداد مشخصی از تکرارها، الگوهای مقدماتی سراسری و جوامع با به‌کارگیری مدل موضوعی جفت‌کلمه مبتنی بر ساختار یاد گرفته می‌شوند و مرزبندی‌هایی در رئوس شبکه برای نشان‌دادن جوامع مختلف به وجود می‌آید. سپس پیاده‌روی تصادفی مبتنی بر جامعه در مراحل بعدی اتخاذ می‌گردد و در طی فرایند تکراری، مرزبندی‌ها دقیق‌تر و جوامع آشکارتر می‌شوند. بنابراین پیاده‌روی‌های تصادفی تمایل دارند که در یک جامعه خاص به رئوس دسترسی داشته باشند. وقتی رئوس در یک دنباله به یک جامعه تعلق داشته باشند، یادگیری توزیع واقعی جامعه مطابق با (۳) تسهیل می‌شود. از Skip-Gram برای به‌حداکثررساندن تابع احتمال شرطی رئوس در جامعه استفاده می‌شود و بنابراین بردارهای نمایشی

Algorithm 4: Framework of CSNE
Input: graph: $G(V,E,T)$; Window size: w;
representation dimension d; walks per vertex: γ;
walk length: l; number of topics or communities: k;
margin appear moment: M_m; margins enlarge speed: M_s.
Output: matrix of network representations: $\phi \in R^{|v|\times d}$.

Sample ϕ from $u^{|v|\times d}$
$M_{\max} = 0$
for $i = 0$ **to** γ **do**
  if $i > \gamma \times M_m$ **then**
    $M_{\max} = \min(1, M_{\max} + \frac{1}{\gamma \times M_s})$
  end if
  $v = shuffle(V)$
  for each vertex $v \in V$ **do**
    $S_v = RandomWalk(G, v, l)$
    $D_v = ContextAggregation(G, v, P_{r_s})$
    $SkipGram(\Phi, S_v, W)$
  end for
  $P_{r_s} = BTM_s(S)$
  $P_{r_t} = BTM_t(D)$
end for
$\phi = \phi \oplus p_{r_t}$
return ϕ

شکل ۸: الگوریتم CSNE.

رئوس از دنباله‌های پیاده‌روی تصادفی به دست می‌آیند. علاوه بر این، محتوای متن یک رأس با چندین همسایه در یک سند واحد جمع می‌شود و به مدل موضوعی جفت‌کلمه مبتنی بر متن وارد می‌گردد و خروجی تولیدشده در نهایت به بردارهای ارائه مبتنی بر ساختار می‌پیوندد.

الگوریتم CSNE در شکل ۸ آمده است. بر اساس شکل، ماتریس U به طور تصادفی تولید می‌شود تا بردارهای رأس را در خط ۱ الگوریتم ۱ مقداردهی اولیه کند. $M_{\max}$ حداکثر مقدار پارامتر مرزبندی فعلی را بین رئوس نشان می‌دهد که در ابتدا با مقدار صفر تنظیم شده و حداکثر مقدار آن، ۱ است. هنگامی که پارامتر مرزبندی برابر با صفر باشد، پیاده‌روی تصادفی معادل الگوریتم سنتی پیاده‌روی عمیق خواهد بود. با افزایش $M_{\max}$، احتمال دستیابی به رئوس در جوامع مختلف به تدریج کاهش می‌یابد. زمانی که پارامتر $M_{\max}$ به مقدار ۱ می‌رسد، پیاده‌روی تصادفی فقط در یک جامعه می‌تواند انجام شود. به این ترتیب می‌توان پارامتر $M_{\max}$ را با عنوان میزان اطمینان از ایجاد جامعه در نظر گرفت. این فرایند توسط دو پارامتر M_m و M_s بین ۰ و ۱ کنترل می‌شود. در میان آنها، M_m نشان‌دهنده لحظه ظاهرشدن مرز است و M_s زمان لازم برای رسیدن $M_{\max}$ به ۱ را نشان می‌دهد. از طریق پیاده‌روی تصادفی مبتنی بر مرزبندی که به تفصیل در شکل ۹ شرح داده شده است، توالی‌های رئوس تولید به مدل موضوعی جفت‌کلمه مبتنی بر ساختار ارسال می‌شوند و به طور مشابه، ویژگی‌های متن یک رأس را با همسایگان آن تجمیع می‌شود و متون جمع‌آوری‌شده به مدل موضوعی جفت‌کلمه مبتنی بر متن وارد می‌شود.

احتمال رئوس متعلق به جامعه مبتنی بر ساختار P_{r_s} و متعلق به جامعه مبتنی بر متن P_{r_t} آموخته می‌شود. P_{r_s} در شکل‌های ۹ و ۱۰، در خط ۵ برای انتخاب گره‌ها در دنباله پیاده‌روی تصادفی استفاده می‌شود؛ در حالی که P_{r_t} در شکل ۸ در انتها به بردارهای ارائه اصلی به‌دست‌آمده توسط

Algorithm 5: RandomWalk
Input: graph: $G(V,E,T)$; Walk length: l ; number of communities: K ; probability of vertices belonging to structure-based communities: P_{r_s} ; max margin among vertices: $M_{\max}$.
Output: a random walk sequence: S_v .
1: Initialize random walk
2: **While** $length(S_v) < l$ **do**
3: **if** current vertex has neighbors, **then**
4: **for** each neighbor vertex u of v **do**
5: $p(u|v) = 1 - \frac{1}{2}.\sum_{i=1}^{k}\left|p_{r_s}(C_i|v) - p_{r_s}(C_i|u)\right|$
6: **end for**
7: select a vertex u from neighbors of v based on Roulette Wheel.
8: **else**
9: backtrack in the sequence and select a vertex at random
10: **end if**
11: **end while**
12: **return** S_v

شکل ۹: پیاده‌روی تصادفی و تشخیص مروری جامعه.

Algorithm 6: Context Aggregation
Input: graph $G(V,E,T)$; number of communities: k ; probability of vertices belonging to structure-based communities: P_{r_s} .
Output: the contextual text information: D_v .
1: Initialize D_v with T_v
2: **While** $length(D_v) < \gamma \log|v|$ **do**
3: **if** current vertex has neighbors, **then**
4: **for** each neighbor vertex u of v **do**
5: $1 - \frac{1}{2}.\sum_{i=1}^{k}\left|P_{r_s}(C_i|v) - P_{r_s}(C_i|u)\right|$
6: **end for**
7: select a vertex u from neighbors based on Roulette Wheel
8: $D_v = D_v \oplus T_u$
9: **else**
10: $D_v = D_v \oplus T_v$
11: **end if**
12: **end while**
13: **return** D_v

شکل ۱۰: الگوریتم تلفیق متن هر گره با همسایگانش.

Skip-Gram در خط ۱۶ پیوند می‌خورد.

در ارائه‌شده در شکل ۹، دنباله رئوس با رأس v آغاز می‌شود و چندین بار تکرار می‌گردد تا اینکه طول دنباله به طول پیاده‌روی ازپیش‌تعیین‌شده l برسد. خطوط ۲ تا ۱۱ در شکل ۹ هسته اصلی رویکرد پیاده‌روی تصادفی را نشان می‌دهند. در صورتی که رأس فعلی v باشد، احتمال شرطی رأس بعدی u یا $P(u|v)$ بزرگ‌تر خواهد بود اگر P_{r_s} متعلق به دو گره v و u دارای مقادیر نزدیک به هم باشند. زمانی این دو مقدار احتمالی به هم نزدیک می‌شوند که هر دو در یک جامعه یا جوامع نزدیک به هم واقع شده باشند.

پارامتر $M_{\max}$ میزان تأثیر جامعه را در احتمال نهایی مختص به گره‌ها نشان می‌دهد. هرچه این مقدار به عدد یک نزدیک‌تر باشد، انتخاب گره‌های دنباله پیاده‌روی تصادفی از یک جامعه صورت می‌گیرد. انتخاب رئوس بعدی متکی به استراتژی انتخاب بر اساس انتخاب چرخ رولت‌ویل الگوریتم ژنتیک است [۵۳]. اگر رأس فعلی همسایگی نداشته باشد، ترتیب دنباله پیاده‌روی تصادفی به سمت عقب برمی‌گردد و یک رأس را به طور تصادفی انتخاب می‌کند.

مدل Skip-Gram و مدل موضوعی جفت‌کلمه، رویکردهای یادگیری آنلاین هستند؛ به این معنی است که توالی رئوس به جای اینکه به‌صورت کلی مورد تحلیل قرار بگیرند می‌توانند در دسته‌های کوچک‌تر نیز وارد مدل‌ها شوند. روش این بخش مانند بخش قبلی قادر به ایجاد تغییرات کوچک در شبکه است و پیاده‌روی‌های تصادفی جدید می‌توانند بر روی منطقه تغییریافته بدون نیاز به محاسبه رئوس قبلی تمرکز کنند.

شکل ۱۰ الگوریتم روند جمع‌بندی متن را نشان می‌دهد. برای کاهش انحراف بین توزیع جامعه خلفی و توزیع واقعی جامعه باید این نکته را در نظر گرفت که طول یک سند از تعداد کل اسناد کمتر نیست [۵۴]

$$v \log|v| \ll length(D_v) \qquad (۶)$$

خط پنج شکل ۱۰، نحوه اختصاص احتمالات به ازای گره‌های همسایه را نشان می‌دهد. اگر گره همسایه در همان جامعه‌ای باشد که گره مبدأ قرار دارد آن گاه دارای احتمالات نزدیک به هم هستند که تفریق آنها مقدار کمی را ایجاد می‌کند. در نهایت احتمال گره‌هایی که با گره مبدأ در یک جامعه قرار دارند بیشتر از گره‌هایی خواهد بود که در یک جامعه مشترک با گره مبدأ نیستند. انتخاب رئوس بعدی متکی به استراتژی انتخاب بر اساس انتخاب چرخ رولت‌ویل الگوریتم ژنتیک است؛ به طوری رئوسی که در یک جامعه قرار دارند دارای احتمال انتخاب بالاتری هستند.

٤- آزمایش‌ها

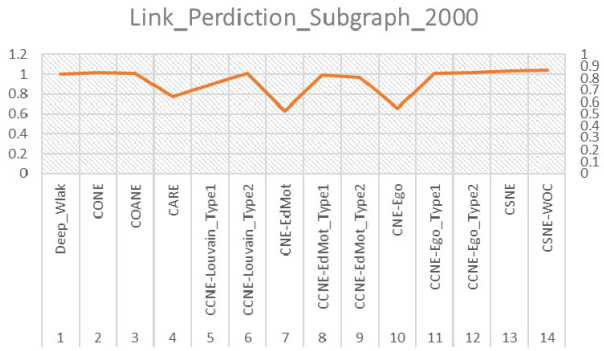
در این بخش ابتدا روش‌های مبنا، مجموعه داده‌های آزمایشی و تنظیمات پارامتر را مورد بررسی قرار می‌دهیم. سپس الگوریتم‌های جدید ارائه‌شده در مقاله بر روی یادگیری بدون نظارت مانند تجسم شبکه و دو وظیفه یادگیری بانظارت مانند طبقه‌بندی رئوس و پیش‌بینی یال‌ها مورد ارزیابی قرار گرفته است.

روش پیاده‌روی عمیق که از پردازش زبان طبیعی برای تعبیه‌سازی شبکه استفاده می‌کند الگوریتمی پیشقدم در حوزه تعبیه‌سازی شبکه است. روش CARE الگوریتمی آگاه به جامعه در خصوص تعبیه‌سازی شبکه است. همچنین روش‌های CARE، CONE و COANE الگوریتم‌های آگاه به جامعه در خصوص تعبیه‌سازی شبکه هستند. روش CARE با استفاده از الگوریتم لوویان و روش‌های CONE و COANE با استفاده از مدل‌های موضوعی اطلاعات جامعه را استخراج کرده و پیاده‌روی تصادفی را تولید می‌نمایند. در نهایت پیاده‌روی‌های تولیدشده را با استفاده از روش Skip-Gram به بردارهای نمایشی با بعد کم تبدیل می‌کنند.

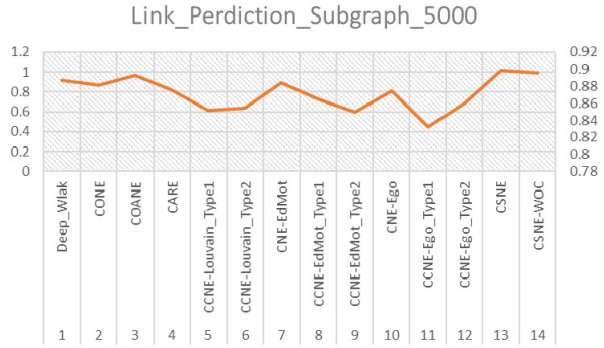
در این ادامه این بخش، به مقایسه روش‌های جدید ارائه شده در این مقاله با نام‌های CNE، CCNE_Type1، CCNE_Type2 و CSNE با روش‌های کلاسیک پیاده‌روی عمیق، CARE، CONE و COANE بر روی مجموعه داده‌های معرفی‌شده پرداخته شده است.

٤-١ مجموعه داده‌ها

دیتاست Cora شامل ۲۷۰۸ مقاله یادگیری ماشین از ۷ کلاس و ۵۴۲۹ یال در بین مقالات است. هر رأس یک مقاله را نشان داده و روابط استنادی بین اسناد یک شبکه پیچیده معمولی را تشکیل می‌دهد [۵۵]. DBLP V۱۲ شامل ۴ میلیون مقاله و ۴۵ میلیون یال بین مقالات است و تاریخ این مجموعه داده برابر با ۲۰۲۰/۰۴/۰۹ می‌باشد (جدول ۱). در این مقاله از دو زیرگراف با تعداد ۲۰۰۰ و ۵۰۰۰ گره برای اجرا استفاده شده است [۵۶].



شکل ۱۱: امتیازات ACU بر روی معیار پیش‌بینی یال برای زیرگراف ۲۰۰۰تایی.



شکل ۱۲: امتیازات ACU بر روی معیار پیش‌بینی یال برای زیرگراف ۵۰۰۰تایی.

در این نوشتار از عنوان هر مقاله به برای اطلاعات ویژگی استفاده شده است. در خصوص مجموعه داده Cora عناوین استخراج‌گردیده از مجموعه داده اصلی دارای مقادیر گم‌شده بودند که در این خصوص از لینک مقالات به عنوان جایگزینی در این باره استفاده شده است.

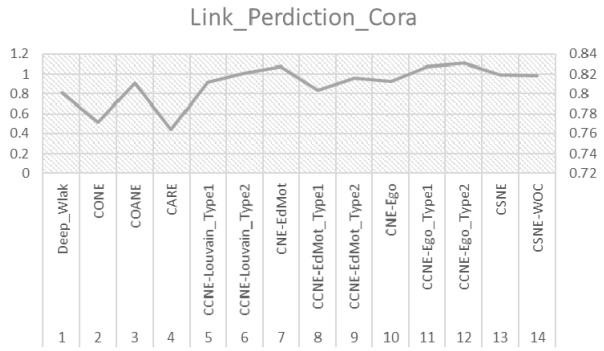
بعد بردار نمایش $d=۱۲۸$ برای همه مجموعه داده‌های بالا تنظیم گردیده است. برای پیاده‌روی عمیق تعداد پیاده‌روی برابر با ۲۰، طول پیاده‌روی برابر با ۲۰ و اندازه پنجره w برابر با ۱۰ تنظیم شده است. به منظور ارائه یک مقایسه منصفانه، تنظیمات پارامتر استفاده‌گردیده برای CNE، CCNE_Type۱&۲، CSNE، CARE، CONE و COANE با مقادیر استفاده‌شده برای پیاده‌روی عمیق مطابقت دارد. علاوه بر این، $M_m=۰/۲$ و $M_s=۰/۳$ در الگوریتم‌های CONE، COANE و CSNE مقداردهی گردیده و در تمام موارد بالا مقدار متغیر k برابر با ۱۴ در نظر گرفته شده است.

٤-۲ پیش‌بینی یال

پیش‌بینی یال یک وظیفه در یادگیری بانظارت است. معیار ارزیابی استاندارد، منطقه زیر منحنی (AUC)[1] اتخاذ شده که نشان‌دهنده احتمال شبیه‌تربودن رئوس بالقوه متصل نسبت به موارد نامربوط است.

در شکل‌های ۱۱ تا ۱۳ مشاهده می‌شود که دقت نتایج به‌دست‌آمده برای مجموعه داده‌ای ۵۰۰۰عضوی بهتر از بقیه مجموعه داده‌هاست. مجموعه داده بزرگ‌تر دارای ارتباطات بیشتری هستند و روابط موجود در جامعه را بهتر نشان می‌دهند. دقت نتایج برای مجموعه داده Cora از مجموعه داده‌های دیگر مخصوصاً در روش‌هایی که از مدل‌های موضوعی برای تشخیص جوامع استفاده کرده‌اند بدتر عمل کرده است. در خصوص مجموعه داده Cora می‌توان این مورد را ذکر کرد که این مجموعه داده در خصوص محتوای متنی یعنی عنوان مقاله‌ها، دارای موارد گم‌شده

1. Area Under Curve



شکل ۱۳: امتیازات ACU بر روی معیار پیش‌بینی یال برای مجموعه داده Cora.

جدول ۱: مجموعه داده‌های استفاده‌شده برای آزمایش‌ها.

مجموعه داده‌ها	گره‌ها	یال‌ها	برچسب‌ها
Cora	۲۷۰۸	۵۴۲۹	۷
Dblp_۲۰۰۰	۲۰۰۰	۴۰۱۳	۴
Dblp_۵۰۰۰	۵۰۰۰	۱۱۵۸۷	۴

بسیاری بوده که در این حالت از لینک مقالات به‌جای عناوین گم‌شده استفاده شده است؛ اما مجموعه داده‌های ۲۰۰۰ و ۵۰۰۰تایی قسمت عناوین آنها کامل بوده و روش‌های مرتبط با مدل‌های موضوعی یا تشخیص جوامع حین پیاده‌روی تصادفی بهتر از روش‌های مربوط به تشخیص جوامع قبل از پیاده‌روی تصادفی عمل کرده‌اند. بنابراین مدل‌های موضوعی برای عملکرد بهتر به مجموعه داده متنی تقریباً کامل و درستی نیازمند هستند.

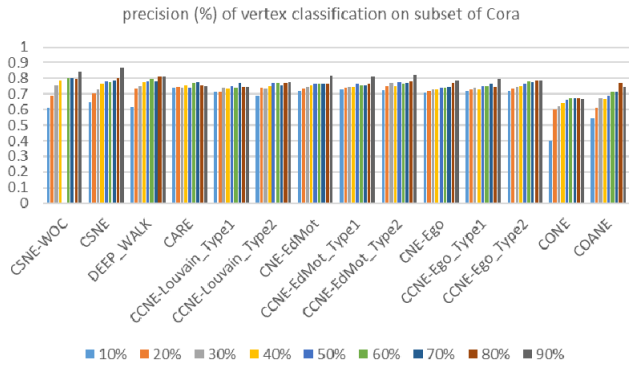
بر اساس این تحلیل اگر مجموعه داده دارای بخش متنی کامل و درستی باشد، بهتر است که از روش‌های تشخیص جامعه حین پیاده‌روی تصادفی و در غیر این صورت از روش‌های تشخیص جامعه قبل از روند پیاده‌روی تصادفی استفاده کرد.

با بررسی نمودارهای ارائه‌شده در شکل‌های ۱۱ تا ۱۳ در خصوص الگوریتم‌های تخصیص جامعه قبل از پیاده‌روی تصادفی مشاهده می‌شود که روش‌های ادموت و اگونت‌اسپلیتر در هر دو حالت بدون استفاده از متن و با استفاده از متن بهتر از روش لوویان عمل کرده‌اند. در روش غیرهم‌پوشان ادمونت با بررسی زیرگراف‌ها موتیف در شبکه و همچنین روش هم‌پوشان اگونت‌اسپلیتر با کاهش مسأله هم‌پوشان به یک مسأله غیرهم‌پوشان بهتر از روش تعبیه‌سازی به وسیله روش غیرهم‌پوشان لوویان (CARE) عمل کرده‌اند. نتایج تجربی، عملکرد بهتر مدل‌های ارائه‌شده در دو فاز تشخیص جوامع قبل و حین پیاده‌روی تصادفی را نشان می‌دهد.

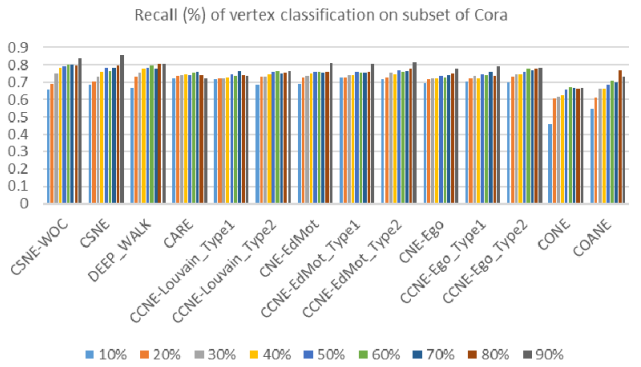
٤-۳ طبقه‌بندی گره‌ها

طبقه‌بندی رأس برای ارزیابی کیفیت نمایش‌های به‌دست‌آمده استفاده می‌شود؛ به طوری که رگرسیون لجستیک[2] تنظیم‌شده با L۲ به‌عنوان طبقه‌بندی‌کننده نظارت‌شده استفاده می‌شود و نسبت آموزش از ۱۰٪ تا ۹۰٪ متغیر است. پارامترهای دقت[3]، فراخوانی[4]، امتیازات Micro-F۱ و Macro-F۱ به عنوان معیارهای ارزیابی در آزمایش‌ها در نظر گرفته شده است.

2. Logistic Regression
3. Precision
4. Recall



شکل ۱۴: امتیاز دقت بر روی مجموعه داده Cora.



شکل ۱۵: امتیاز فراخوانی بر روی مجموعه داده Cora.

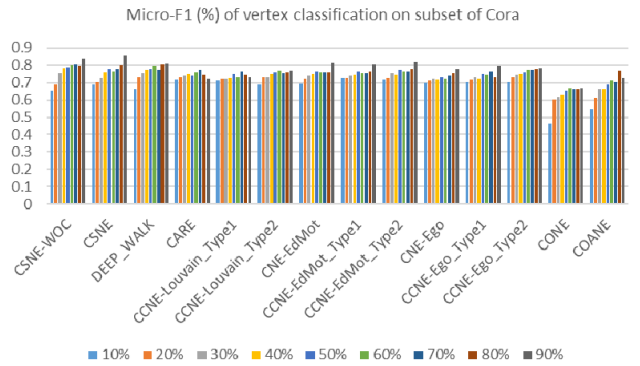
آزمایش‌ها برای ۱۰ بار تکرار شده و میانگین دقت طبقه‌بندی با نسبت آموزشی متفاوت در مجموعه داده Cora در شکل‌های ۱۴ تا ۱۷ آمده است. دلیل انتخاب مجموعه داده Cora در این مرحله، تعداد ۷ برچسبی است که مقالات بر اساس این برچسب‌ها طبقه‌بندی شده‌اند؛ به طوری که CNE، CCNE_Type۱&۲، CSNE و CSNE-WOC[1] به طور قابل توجهی بهتر از روش‌های دیگر عمل کرده‌اند. CSNE و CSNE-WOC در هنگام استفاده از ۹۰ درصد رئوس برچسب‌گذاری‌شده برای آموزش از روش‌های دیگر مورد مطالعه بهتر عمل می‌کنند که ضرورت استفاده از اطلاعات متنی یا ویژگی‌های موجود در گره‌ها در تعبیه‌سازی شبکه را نشان می‌دهد.

٤-٤ تجسم شبکه

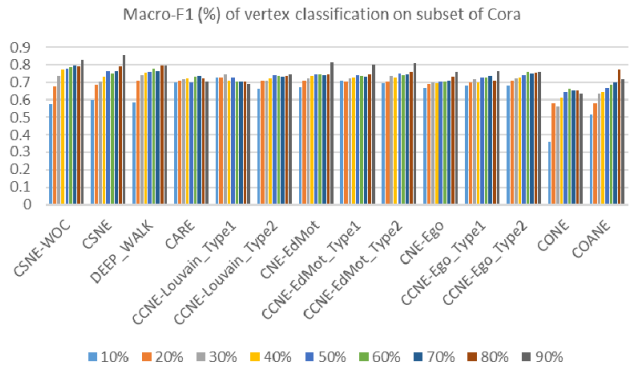
تجسم شبکه برای تجزیه و تحلیل داده‌های با ابعاد بالا ضروری است که می‌تواند به طور مستقیم، ساختار ذاتی داده‌ها را آشکار کند. شبکه DBLP با به‌کارگیری مدل‌های تعبیه‌سازی متفاوت با بردارهای با بعد کم نشان داده می‌شوند و سپس این بردارها با استفاده از t-SNE در یک فضای دوبعدی نگاشت می‌گردند. تجسم شبکه برای مجموعه داده Cora، هر نقطه نشان‌دهنده یک مقاله است و توسط الگوریتم‌های مختلف تعبیه‌سازی شبکه به فضای دوبعدی نگاشت گردیده است. رنگ رئوس نشان‌دهنده تقسیم‌بندی‌های مختلف انتشارات است.

شکل ۱۸ تجسم شبکه به‌دست‌آمده از الگوریتم‌های مختلف تعبیه‌سازی را نشان می‌دهد. از آنجایی که عناوین مقالات که معمولاً از کمتر از ۱۰ کلمه تشکیل شده‌اند به عنوان اطلاعات ویژگی در نظر گرفته می‌شوند، استخراج ویژگی‌ها از این متون کوتاه برای الگوریتم‌های تعبیه‌سازی شبکه یک چالش بزرگ است. در همین راستا با توجه به تفکیک رنگ‌ها در دو

1. CSNE-Without Context



شکل ۱۶: امتیاز Micro-F۱ بر روی مجموعه داده Cora.



شکل ۱۷: امتیاز Macro-F۱ بر روی مجموعه داده Cora.

CSNE روش COANE و CSNE می‌توان مشاهده کرد که روش CSNE بهتر از COANE عمل کرده و این برتری به دلیل مدل موضوعی بهتری است که در روش CSNE نسبت به روش COANE استفاده شده است. برای پیاده‌روی عمیق، مقالات تقریباً بر اساس تقسیمات انتشار آنها گروه‌بندی شده‌اند؛ اگرچه آنها به صورت خطی قابل تفکیک نیستند. با توجه به استفاده از اطلاعات جامعه، روش‌های تفکیک جامعه قبل از پیاده‌روی تصادفی بهتر است. با این حال، مقالات در همان بخش‌ها به خوشه‌های کوچک تبدیل شده‌اند و فشردگی در یک بخش بسیار زیاد است.

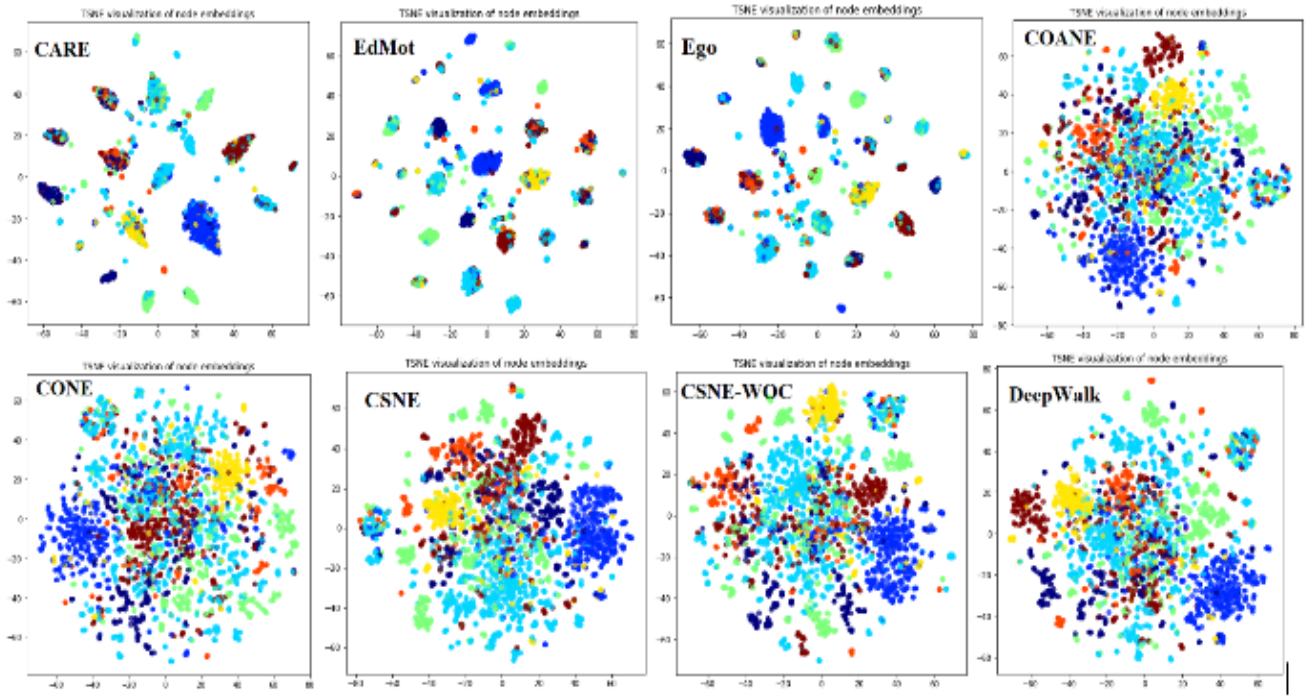
٤-٥ حساسیت پارامترها

تأثیر تعداد جوامع k در شکل ۱۹ نشان داده شده است. پارامتر k از ۹ تا ۲۴ متغیر بوده و مشاهده می‌شود که مقادیر برای روش CSNE نسبتاً روبه‌رشد و پایدار هستند. این امر نشان‌دهنده آن است که تعداد جوامع k تنها تأثیر کمی بر عملکرد پیش‌بینی پیوند دارد. اگرچه باید خاطرنشان کرد که تعداد جوامع (موضوعات) k بر عملکرد مدل‌های موضوعی تأثیر زیادی دارد.

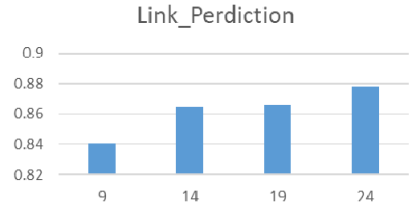
تأثیر تعداد قدم‌ها در پیاده‌روی تصادفی در شکل ۲۰ نشان داده شده است. پارامتر Number of Walk از ۱۰ تا ۴۰ متغیر بوده و مشاهده می‌شود که مقادیر برای روش CSNE برای مقدار ۲۰ گام دقت بهتری را ارائه کرده است. بر اساس همین نتایج، مقدار تعداد گام‌ها در تمام مقایسات برابر با عدد ۲۰ مقداردهی اولیه شده است.

٥- نتیجه‌گیری و پژوهش‌های آینده

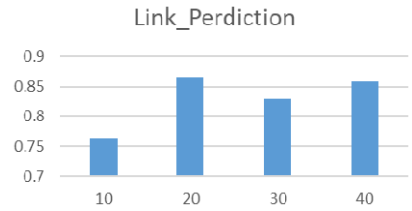
در این مقاله، روش‌های جدید مبتنی بر تشخیص جامعه قبل و حین پیاده‌روی تصادفی با درنظرگرفتن تحلیل متن ارائه شده و مقایسه بر روی تأثیر این روش‌ها در مجموعه داده‌های مختلف مورد بررسی قرار گرفته



شکل ۱۸: تجسم شبکه‌های استنادی Cora در خصوص الگوریتم‌های مورد آزمایش.



شکل ۱۹: تأثیر تعداد جوامع بر روی معیار پیش‌بینی یال برای الگوریتم CSNE.



شکل ۲۰: تأثیر تعداد قدم‌ها در پیاده‌روی تصادفی بر روی معیار پیش‌بینی یال برای الگوریتم CSNE.

است. استراتژی‌های پیاده‌روی تصادفی مبتنی بر تحلیل متنی گره‌ها پیشنهاد شده تا به طور مشترک اطلاعات مربوط به مجاورت، جامعه و خصوصیات گره‌ها را حفظ کند. روش‌های CNE، CCNE_Type۱، CCNE_Type۲ و CSNE بر کاستی‌های پژوهشی این حوزه غلبه کرده و به کارایی بالایی در تعبیه‌سازی شبکه مبتنی بر تحلیل جامعه و تحلیل مفهومی متن دست یافته است. نتایج تجربی در شبکه‌های مبتنی بر مفاهیم دنیای واقعی، اثربخشی و استحکام CNE، CCNE_Type۱، CCNE_Type۲ و CSNE را در مقایسه با چهار روش پایه پیاده‌روی عمیق، CARE، CONE و COANE نشان می‌دهد. روش‌های CNE، CCNE_Type۱، CCNE_Type۲ و CSNE بر روی یک نوع گره تمرکز دارند. با این حال، شبکه‌های دنیای واقعی معمولاً از انواع مختلفی از رئوس، روابط و اطلاعات صریح تشکیل شده‌اند. بنابراین در ادامه این پژوهش می‌توان روش پیشنهادی را بر روی شبکه‌های ناهمگن گسترش داد. همچنین می‌توان از نتایج به‌دست‌آمده در خصوص تعبیه‌سازی شبکه برای به‌دست‌آوردن گره‌های تأثیرگذار و بیشینه‌سازی تأثیر در شبکه‌های اجتماعی، استفاده بهینه کرد. در کارهای آینده قصد داریم این کار را در خصوص به‌دست‌آوردن گره‌های تأثیرگذار و بیشنه‌سازی شبکه بسط دهیم.

٦- سپاسگزاری

این مقاله مستخرج از رساله دکترای تخصصی محدثه طاهرپرور در واحد رشت دانشگاه آزاد اسلامی است.

مراجع

[1] P. Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: a survey," *Knowl.-Based Syst.*, vol. 151, pp. 78-94, Jul. 2018.

[2] H. Cai, V. W. Zheng, and K. Chang, "A comprehensive survey of graph embedding: problems, techniques and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 9, pp. 1616-1637, Sept. 2018.

[3] I. Brugere, B. Gallagher, and T. Y. BergerWolf, "Network structure inference, a survey: motivations, methods, and applications," *ACM Comput. Surv.*, vol. 51, no. 2, Article ID: 24, 39 pp., Mar. 2019.

[4] F. Huang, X. Zhang, J. Xu, C. Li, and Z. Li, "Network embedding by fusingmultimodal contents and links," *Knowl.-Based Syst.*, vol. 171, pp. 44-55, May 2019.

[5] J. Liao, S. Wang, D. Li, and X. Li, "FREERL: fusion relation embedded representation learning framework for aspect extraction," *Knowl. Based Syst.*, vol. 135, pp. 9-17, Nov. 2017.

[6] L. Boratto, S. Carta, G. Fenu, and R. Saia, "Using neural word embeddings to model user behavior and detect user segments," *Knowl.-Based Syst.*, vol. 108, pp. 5-14, Sept. 2016.

[7] M. Ji, J. Han, and M. Danilevsky, "Ranking-based classification of heterogeneous information networks," in *Proc. of the 17th ACM SIGKDD In. Conf. on Knowledge Discovery and Data Mining*, pp. 1298-1306, San Diego, CA, USA, 21-24 Aug. 2011.

[8] R. A. Sinoara, J. Camachocollados, R. Rossi, R. Navigli, and S. O. Rezende, "Knowledge-enhanced document embeddings for text classification," *Knowl.-Based Syst.* vol. 163, pp. 955-971, Jan. 2019.

[9] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019-1031, May 2007.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. of the 26th Int. Conf. on Neural Information Processing Systems,* vol. 2, pp. 3111-3119, Lake Tahoe, NA, USA, 5-10 Dec. 2013.

[11] B. Perozzi, R. AlRfou, and S. Skiena, "Deepwalk: online learning of social representations," in *Proc. of the 20th ACM SIGKDD Int.*

24th Int. Joint Conf. on Artificial Intelligence, pp. 2111-2117, Buenos Aires, Argentina, 25-31 Jul. 2015.

[36] Z. Chen, T. Cai, C. Chen, Z. Zheng, and G. Ling, "SINE: side information network embedding," in *Proc. of the 24th Int. Conf. on Database Systems for Advanced Applications*, pp. 692-708, Chiang Mai, Thailand, 22-25 Apr. 2019.

[37] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 1225-1234, San Francisco, CA, USA, 13-17 Aug. 2016.

[38] X. Wang, D. Jin, X. Cao, L. Yang, and W. Zhang, "Semantic community identification in large attribute networks," in *Proc. of the 30th AAAI Conf. on Artificial Intelligence*, pp. 265-271, Phoenix, AZ, USA, 12-17 Feb. 2016.

[39] M. Li, J. Liu, P. Wu, and X. Teng, "Evolutionary network embedding preserving both local proximity and community structure," *IEEE Trans. Evol. Comput.*, vol. 24, no. 3, pp. 523-535, Jun. 2019.

[40] J. Chen, Q. Zhang, and X. Huang, "Incorporate group information to enhance network embedding," in *Proc. of the 25th ACM Int. Conf. on Information and Knowledge Management*, pp. 1901-1904, Indianapolis, IN, USA, 24-28 Oct. 2016.

[41] X. Xia, *et al*., "Improving automated bug triaging with specialized topic model," *IEEE Trans. Softw. Eng.*, vol. 43, no. 3, pp. 272-297, Mar. 2017.

[42] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. of the 22nd Annual Int. ACM SIGIR Conf. on Research and Revelopment in Information Retrieval*, pp. 50-57, Berkeley, CA, USA, 15-19 Aug. 1999.

[43] M. Steyvers, P. Smyth, M. RosenZvi, and T. Griffiths, "Probabilistic author-topic models for information discovery," in *Proc. of the 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 306-315, Seattle, WA, USA, 22-25 Aug. 2004.

[44] M. RosenZvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proc. of the 20th Conf. on Uncertainty in Artificial Intelligence*, pp. 487-494, Banff, Canada, 6-11 Jul. 2004.

[45] Q. Mei, D. Cai, D. Zhang, and C. Zhai, "Topic modeling with network regularization," in *Proc. of the 17th Int. Conf. on World Wide Web*, pp. 101-110, Beijing, China, 21-25 Apr. 2008.

[46] Y. Shi, M. Lei, H. Yang, and L. Niu, "Diffusion network embedding," *Pattern Recognit.*, vol. 88, pp. 518-531, Apr. 2019.

[47] H. Chen, *et al*., "Exploiting centrality information with graph convolutions for network representation learning," in *Proc. of the 35th IEEE Int. Conf. on Data Engineering*, pp. 590-601, Macao, China ,8-11 Apr. 2019.

[48] W. Zhao, H. Ma, Z. Li, X. Ao, and N. Li, "SBRNE: an improved unified framework for social and behavior recommendations with network embedding," in *Proc. of the 24th Int. Conf. on Database Systems for Advanced Applications*, pp. 555-571, Chiang Mai, Thailand, 22-22 Apr. 2019.

[49] Q. Li, J. Zhong, Q. Li, Z. Cao, and C. Wang, "Enhancing network embedding with implicit clustering," in *Proc. of the 24th Int. Conf. on Database Systems for Advanced Applications*, pp. 452-467, Chiang Mai, Thailand, April 22-25, 2019.

[50] L. Wu, D. Wang, S. Feng, Y. Zhang, and G. Yu, "MDAL: multi-task dual attention LSTM model for semi-supervised network embedding," in *Proc. of the 24th Int. Conf. on Database Systems for Advanced Applications*, Chiang Mai, Thailand, April 22-25, 2019.

[51] Y. Gao, M. Gong, Y. Xie, and H. Zhong, "Community-oriented attributed network embedding," *Knowledge-Based Systems*, vol. 193, Article ID: 105418, Apr. 2019.

[52] X. Cheng, X. Yan, Y. Lan, and J. Guo, "BTM: topic modeling over short texts," *IEEE Trans. on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 2928 - 2941, Dec. 2013.

[53] D. Whitley, "A genetic algorithm tutorial," *Stat. Comput.*, vol. 4, no. 2, pp. 65-85, Jun. 1994.

[54] J. Tang, Z. Meng, X. Nguyen, Q. Mei, and M. Zhang, "Understanding the limiting factors of topic modeling via posterior contraction analysis," in *Proc. of the 31st Int. Conf. on Machine Learning*, pp. 190-198, Beijing China 21-26 Jun. 2014.

[55] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore, "Automating the construction of internet portals with machine learning," *Information Retrieval.* vol. 3, no. 2, pp. 127-163, Jun. 2000.

[56] -, *DBLP Citation Network*, https://www.aminer.org/citation

Conf. on Knowledge Discovery and Data Mining, pp. 701-710, New York, NY, USA, 24-27 Aug. 2014.

[12] A. Grover and J. Leskovec, "node2vec: scalable feature learning for networks," in *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 855-864, San Francisco, CA, USA, 13-17 Aug. 2016.

[13] J. Tang, *et al.*, "LINE: large-scale information network embedding," in *Proc. of the 24th Int. Conf. on World Wide Web*, pp. 1067-1077, Florence Italy, 18-22 May 2015.

[14] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. of the 31st Int. Conf. on Neural Information Processing Systems*, pp. 1024-1034, Long Beach, CA, USA, 4-9 Dec. 2017.

[15] H. Gao and H. Huang, "Deep attributed network embedding," in *Proc. of the 27th Int. Joint Conf. on Artificial Intelligence*, pp. 3364-3370, Stockholm Sweden, 13-19 Jul. 2018.

[16] X. Huang, J. Li, and X. Hu, "Label informed attributed network embedding," in *Proc. of the 10th ACM Int. Conf. on Web Search and Data Mining*, pp. 731-739, Cambridge, UK, 6-10 Feb. 2017.

[17] J. Liang, P. Jacobs, J. Sun, and S. Parthasarathy, "Semi-supervised embedding in attributed networks with outliers," in *Proc. of the SIAM Int. Conf. on Data Mining*, pp. 153-161, 2018.

[18] L. Yang, *et al.*, "Modularity based community detection with deep learning," in *Proc. of the 25th Int. Joint Conf. on Artificial Intelligence*, pp. 2252-2258, New York, NY, USA, 9-15 Jul. 2016.

[19] X. Wang, P. Cui, J. Wang, J. Pei, W. Zhu, and S. Yang, "Community preserving network embedding," *Proc. of the 31st AAAI Conf. on Artificial Intelligence*, pp. 203-209, Washington, DC, USA, 4-7 Feb. 2017.

[20] S. Ismail, and R. Ismail, "Modularity approach for community detection in complex networks," in *Proc. the 11th Int, Conf. Ubiquitous Information Management and Communication*, 6 pp., Beppu, Japan, 5-7 Jan. 2017.

[21] S. Fortunato and M. Barthelemy, "Resolution limit in community detection," in *Proc. Natl. Acad. Sci. USA*, vol. 104, no. 1, pp. 36-41, Jan. 2007.

[22] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513-523, 1988.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. 1, pp. 993-1022, 2003.

[24] M. M. Keikha, M. Rahgozar, and M. Asadpour, "Community aware random walk for network embeding," *Knowl. Based Syst.*, vol. 148, pp. 47-54, 2018.

[25] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.*, vol. 10, Article ID: P10008, 2008.

[26] P. Li, L. Huang, C. Wang, and J. Lai, "EdMot: an edge enhancement approach for motif-aware community detection," in *Proc. of the 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, pp. 479-487, Anchorage, AK, USA, 4-8 Aug. 2019.

[27] A. Epasto, S. La. anzi, R. Leme, "Ego-Spli.ing framework: from non-overlapping to overlapping clusters," in *Proc. of the 23th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, Halifax, Canada, 13-17 Aug. 2017.

[28] L. Tang and H. Liu, "Leveraging social media networks for classification," *Data Min. Knowl. Discov.*, vol. 23, no. 3, pp. 447-478, Nov. 2011.

[29] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 250, no. 5500, pp. 2319-2323, 22 Dec. 2000.

[30] A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A. J. Smola, "Distributed large-scale natural graph factorization," in *Proc. of the 22nd Int. Conf. on World Wide Web*, pp. 37-48, Rio de Janeiro, Brazil, 13-17 May 2013.

[31] T. F. Cox and M. A. Cox, *Multidimensional Scaling*, CRC Press, 2000.

[32] S. Yan, D. Xu, B. Zhang, H. J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40-51, Jan. 2007.

[33] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, no. 3-5, pp. 75-174, Feb. 2010.

[34] K. Henderson, *et al.*, "RolX: structural role extraction & mining in large graphs," in *Proc. of the 18th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, Beijing, China, 12-16 Aug. 2012.

[35] C. Yang, Z. Liu, D. Zhao, M. Sun, and E. Y. Chang, "Network representation learning with rich text information," in *Proc. of the*

محدثه طاهرپرور در سال ۱۳۸۷ مدرک کارشناسی مهندسی کامپیوتر نرم‌افزار از دانشگاه آزاد اسلامی واحد لاهیجان و در سال ۱۳۹۲ مدرک کارشناسی ارشد مهندسی کامپیوتر نرم‌افزار از دانشگاه آزاد اسلامی واحد قزوین دریافت نمود و دانشجوی دکتری تخصصی مهندسی کامپیوتر، واحد رشت، دانشگاه آزاد اسلامی، رشت، ایران است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: داده کاوی، یادگیری عمیق و الگوریتم‌های بهینه‌سازی است. او همچنین در پایتون و جاوا تجربه دارد.

فاطمه احمدی آبکناری در سال ۱۳۸۶ مدرک کارشناسی ارشد رشته فناوری اطلاعات از دانشگاه صنعتی امیر کبیر (پلی‌تکنیک) تهران و در سال ۲۰۱۲ مدرک دکتری تخصصی مهندسی کامپیوتر از دانشگاه UTM مالزی دریافت نمود. در سال‌های ۱۳۹۰ تا ۱۴۰۱ استادیار دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام نور، واحد رشت، ایران بوده و در حال حاضر عضو هیأت علمی دانشگاه Northeastern ونکور کانادا است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: یادگیری ماشینی، داده کاوی، متن کاوی، احساسات و عقاید کاوی، شبکه های عصبی مصنوعی، یادگیری عمیق و پردازش زبان طبیعی است.

پیمان بیات مدرک کارشناسی ارشد از دانشگاه آزاد اسلامی واحد اراک و مدرک دکتری مهندسی کامپیوتر از دانشگاه UCSI مالزی دریافت نموده. وی در حال حاضر استادیار، دانشکده مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد رشت، ایران است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: سیستم‌های توزیع‌شده، پردازش تصویر و داده کاوی است.