ششمین همایش بین‌المللی افق های نویـن در
مهندسی برق، کامپیوتـر و مکانیک
6th International Conference on the New Horizons in
Electrical Engineering, Computer and Mechanical
www.mhconf.ir

تعبیه‌سازی صفاتی شبکه مبتنی بر حفظ ساختار و تحلیل معنایی گراف

محدثه طاهرپرور[۱]، فاطمه احمدی آبکناری[۲] و پیمان بیات[۳]

[۱]دانشجوی دکتری ، گروه مهندسی کامپ یوتر و فناوری اطلاعات، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی واحد رشت
mtaherparvar@phd.iaurasht.ac.ir

[۲] استادیار ، گروه مهندسی کامپیوتر و فناوری اطلاعات، گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام نور
Fateme.Abkenari@pnu.ac.ir

[۳] استادیار ، گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشکده فنی و مهندسی ، دانشگاه آزاد اسلامی واحد رشت bayat@iaurasht.ac.ir

چکیده

هدف از تعبیه‌سازی شبکه[۱] این است که رئوس یک شبکه یا گراف پیچیده به یک فضای برداری با ابعاد کم به صورت پیوسته نگاشت شوند و ساختار اصلی و خصوصیات ذاتی شبکه حفظ گردد. بیشتر روش‌های موجود صرفاً بر حفظ ویژگی‌های ساختاری محلی رئوس تمرکز می کنند و از این رو تا حد زیادی از صفات یا اطلاعات غنی هر گره در شبکه چشم‌پوشی می‌کنند. عنوان مقالات در مجموعه داده ای از شبکه استنادی مقالات می‌تواند حاکی از جهت‌گیری‌های پژوهشی آنها باشد که به صورت بالقوه در جستجوهای معنادارتر از این مقالات تاثیرگذار است. در این مقاله، یک چارچوب تعبیه‌سازی صفاتی شبکه ملقب به DCB[2] طراحی و پیشنهاد شده است که می‌تواند اطلاعات و محتوای متن رئوس را در فرایند تعبیه‌سازی شبکه لحاظ کند. با الهام از قیاس بین توالی راس‌ها و اسناد، مدل موضوعی جفت‌کلمه[۳] برای استخراج ویژگی‌ها در شبکه اتخاذ شده است. علاوه بر این، DCB در ضمن حفظ همبستگی ساختاری، مفاهیم متنی[۴] را در نمایش‌ها از طریق مدل موضوعی جفت‌کلمه ادغام می‌کند. آزمایشات انجام شده بر روی گراف هایی از شبکه‌های دنیای واقعی نشان داده است که روش پیشنهادی این مقاله از سه روش برتر تعبیه‌سازی شبکه با نام‌های پیاده‌روی عمیق، CARE و CONE در تجسم شبکه[۵]، طبقه‌بندی رئوس[۶] و پیش‌بینی یال[۷] بهتر است.

واژه‌های کلیدی

شبکه‌های اجتماعی، یادگیری عمیق، مدل موضوعی جفت‌کلمه، تعبیه‌سازی صفاتی شبکه[۸]

---

[1] Network embedding
[2] Deep-CONTEXT-BTM
[3] Biterm topic modeling (BTM)
[4] Textual semantics
[5] Network visualization
[6] Node classification
[7] Link prediction
[8] Attributed network embedding (ANE)

۱. مقدمه

امروزه شبکه‌های اطلاعاتی[9] در انواع سیستم‌های واقعی و در همه جا وجود دارند. چگونگی استخراج موثر ویژگی‌های یک شبکه، امری پیچیده و چالش برانگیز است. به دلیل پراکنده بودن مقادیر در ماتریس همجواری، ویژگی‌های ساختاری اساسی به خوبی منعکس نمی‌شوند. اخیراً بحث تعبیه‌سازی شبکه، به صورت فزاینده‌ای توجه پژوهشگران را به خود جلب نموده است. تعبیه‌سازی شبکه قصد دارد با حفظ خواص گراف، بردارهای متراکم کم بعد رئوس را در شبکه را بیاموزد [۱]. زمینه‌های پژوهشی مختلف در تجزیه و تحلیل شبکه، همانند پیش‌بینی یالها [۲] و طبقه‌بندی رئوس [۳] و ... می‌توانند از طریق الگوریتم‌های یادگیری ماشین موجود به صورت کارآمد براساس درون یافته‌های این حوزه بالنده شوند.

در چند سال گذشته، روش‌های بسیاری برای یادگیری نمایش شبکه ارائه شده است. اکثریت قریب به اتفاق مدل‌ها تمایل دارند اطلاعات ساختاری را به صورت متن استخراج کرده و سپس همسایگی‌های مرتبه پایین یا بالا را در بین آن‌ها بدست آورند. روش پیاده‌روی عمیق[10] [۴] پیاده‌روی‌های تصادفی را برای هر راس ایجاد می‌کند و از آن‌ها به عنوان اطلاعات زمینه‌ای برای یادگیری نمایش رئوس استفاده می‌کند. Node2Vec [۵] با استفاده از دو پارامتر از پیش تعیین شده برای کنترل روش پیاده‌روی تصادفی، که یک معامله بین جستجوهای اول سطح[11] و اول عمق[12] است پیاده روی عمیق را گسترش می‌دهد. در این مقاله، ارتباط مفهومی گره‌ها بوسیله‌ی مدل‌های موضوعی به ارتباط ساختاری بدست آمده از روش پیاده‌روی تصادفی اضافه شده است. AANE[13] [۶] تعبیه‌سازی شبکه را براساس تجزیه صفت وابستگی و تفاوت تعبیه‌سازی بین رئوس مجاور یاد می‌گیرد. SemNE[14] [۷] اطلاعات ترتیب رئوس را در فرایند تعبیه‌سازی ادغام می‌کند و سپس نمایش‌ها را اصلاح کرده و آن‌ها را به وسیله داده‌های تفسیری[15] متناسب‌تر می‌کند.

شبکه‌های دنیای واقعی معمولاً حاوی اطلاعات غنی هستند که از اهمیت ویژه‌ای در برنامه‌های سطح جامعه مانند تجسم شبکه برخوردار هستند. اکثر روش‌ها مستقیماً کاهش ابعاد را با استفاده از ماتریس‌های TF-IDF انجام می‌دهند [۸]، در حالی که این روش ساده فقط می‌تواند تشابه متن را اندازه‌گیری کند تا تشابه معنایی کلمات. علاوه بر این، راس‌ها در همان جامعه معمولاً برخی از اطلاعات ویژگی مشترک را به اشتراک می‌گذارند. به عنوان مثال، مقالات در یک زمینه تحقیقاتی خاص دارای عناوین و چکیده‌های مشابه هستند. انتظار می‌رود این شباهت‌ها برای تسهیل کاوش در ساختار شبکه و به دست آوردن نمایش‌های موثر بکار گرفته شود. بنابراین، استفاده از مفاهیم متنی و روش‌هایی که بتوانند این مفاهیم را در تعبیه‌سازی شبکه لحاظ کنند بسیار ضروری است.

برای پرداختن به چالش‌های ذکر شده در این مقاله، یک روش جدید تعبیه‌سازی شبکه مبتنی بر تحلیل متن (DCB) ارائه شده است. در روش پیاده‌روی عمیق، کلمات موجود در متن و رئوس موجود در دنباله‌های پیاده‌روی تصادفی دارای یک توزیع قانونمند هستند و تکنیک‌های یادگیری عمیق [۹] مانند Skip-Gram که در پردازش زبان طبیعی موفق بوده را برای اولین بار در تجزیه و تحلیل شبکه بکار برده است[۴]. الگوریتم پیاده‌روی عمیق، پیاده‌روی‌های تصادفی کوتاه را ایجاد می‌کند که نمایش‌های اجتماعی رئوس شبکه بر این اساس یاد گرفته می‌شوند. این نمایش‌های نهفته روابط اجتماعی را در یک فضای برداری با ابعاد نسبتاً کمی رمزگذاری می‌کند. در این مقاله بوسیله‌ی نمونه‌برداری تصادفی، دنباله‌های پیاده‌روی ساخته می‌شوند سپس محتوای متن هر راس و همسایگان آن در یک سند واحد جمع‌آوری شده و بوسیله مدل موضوعی جفت‌کلمه، الگوی موجود در متون تجمیع شده محاسبه می‌گردد. رئوس مجاور توزیع جامعه مبتنی بر متن مشابهی دارند. پس روند مدل‌سازی برای توالی‌های اسناد با ایجاد توالی‌های متنوع تولید شده از طریق نمونه‌برداری تصادفی می‌تواند به مدل موضوعی کمک کند تا توزیع موضوع با ثبات‌تری را یاد بگیرد. به طور خاص، نوآوری مقاله به شرح زیر است:

یک روش جدید تعبیه‌سازی شبکه با نام DCB ارائه شده است که در آن، پیاده‌روی‌های تصادفی بوسیله‌ی نمونه‌برداری تصادفی تولید می‌شوند. به ازای هر دنباله‌ای از پیاده‌روی‌های تصادفی تولید شده در مرحله قبلی، محتوای متن هر گره با گره‌های همسایه تجمیع شده و

---

[9] Information network
[10] Deep-Walk
[11] Breadth-first Search
[12] Depth-first Search
[13] Accelerated attributed network embedding
[14] Semi-Supervised network embedding
[15] Annotation

بوسیله مدل موضوعی جفت‌کلمه، الگوی موجود در آن‌ها به ازای هر گره در شبکه یاد گرفته می‌شود. بردارهای یادگیری تولید شده از مراحل اول و دوم بهم می‌پیوندند و بردار نمایش نهایی را تولید می‌کنند. روش ارائه شده در این مقاله با نام DCB در چند شبکه دنیای واقعی ارزیابی شده است. نتایج تجربی نشان داده است که که الگوریتم پیشنهادی در زمینه تجسم شبکه، طبقه‌بندی رئوس و پیش‌بینی یال با پیچیدگی زمانی رقابتی، از دیگر روش‌های مبتنی بر جامعه و مبتنی بر متن پیشی گرفته است.

ادامه مقاله به شرح زیر سازماندهی شده است: بخش ۲ به طور خلاصه پیش زمینه‌ها و الگوریتم‌های مربوط به استراتژی پیاده‌روی تصادفی، جوامع موجود در شبکه و مدل موضوعی را ارائه داده است. در بخش ۳، به طور رسمی نگاشت‌ها و توضیحات تعریف شده است. در بخش ۴ آزمایش‌های گسترده‌ای برای تأیید اثربخشی روش DCB نشان داده شده است و حساسیت پارامتری روش پیشنهادی تجزیه و تحلیل شده است. در پایان، با بحث در مورد چارچوب DCB نتیجه‌گیری شده و زمینه‌های آتی پژوهش در بخش ۵ خلاصه شده است.

۲. پژوهش‌های پیشین

یادگیری ارائه[16] [۱۰] به طور گسترده‌ای در زمینه‌های مختلف مانند بینایی رایانه‌ای[17] و پردازش زبان طبیعی[18] ورد استفاده قرار گرفته است. پروزی[19] و همکاران [۴] تشابه بین رئوس در توالی پیاده‌روی تصادفی و کلمات در متن را بیان کرده‌اند. آن‌ها مدل پیاده‌روی عمیق را پیشنهاد کردند که با استفاده از معماری Skip-Gram بردارهای ویژگی را از توالی پیاده‌روی‌های تصادفی استخراج می‌کند. Skip-Gram، به عنوان یک مدل زبانی قادر است با بهینه‌سازی حفظ احتمال محلی با استفاده از نزول شیب تصادفی همراه با نمونه‌گیری منفی، بردارهای نمایشی کم بعدی از کلمات را یاد می‌گیرد. به طور خاص، یک دنباله راس $s = (v1, v2, \ldots\ldots, v|s|)$ بدست آمده با یک پیاده‌روی تصادفی از طریق شبکه به عنوان یک توالی کلمه در نظر گرفته می‌شود و هر راس در دنباله به عنوان یک کلمه در نظر گرفته می‌شود. در مرحله بعدی، پیاده‌روی عمیق می‌تواند نمایش‌های شبکه را به وسیله Skip-Gram بدست آورد که هدف به حداکثر رساندن میانگین ورود به سیستم مشاهدات یک راس است:

$$\max_{\emptyset} = \frac{1}{|s|}\sum_{i=1}^{|s|} logPr(Vi-w, \ldots.Vi-1, Vi+1, \ldots..Vi+w)|\,Vi \qquad (1)$$

در فرمول (۱) متن از کلمات موجود در هر دو طرف کلمه مورد نظر به اندازه‌ی طول پنجره w تشکیل شده است. گروور[20] و همکاران [۵] توالی رئوس را با استفاده از استراتژی‌های نمونه‌برداری اول عمق و اول سطح از طریق دو پارامتر p و q ارائه کرده‌اند تا بتواند از نظر هموفیلی[21] یا معادل ساختاری شبکه ورودی را بیاموزند. یانگ[22] و همکاران [۳۰] ثابت کردند که پیاده‌روی عمیق معادل فاکتورسازی ماتریسی است که عنصر M_{ij} آن را می‌توان به عنوان اطلاعات متقابل مثبت[23] (PMI) از یک جفت متن (Vi, Vj) تفسیر کرد. آنها همچنین محتوای متنی رئوس را در چارچوب فاکتوراسیون ماتریس گنجانیدند. چن[24] و همکاران [۱۱] تعبیه‌سازی شبکه اطلاعات جانبی را ارائه داده‌اند که همسایگی معنایی[25] را برای مدل‌سازی شکل هر گره تعریف می‌کند سپس پیاده‌روی تصادفی را برای کاوش در این همسایگی اعمال می‌کند. وانگ[26] و همکاران [۱۲] یک مدل عمیق با یک معماری نیمه نظارت به نام SDNE ارائه داده‌اند که داده‌ها را به یک فضای پنهان غیرخطی نگاشت می‌کند و قادر است همزمان همسایگی مرتبه اول و مرتبه دوم را بهینه کند. این روش‌ها تنها سعی در حفظ ساختار میکروسکوپی شبکه‌ها دارند. از آنجا که آن‌ها عملکرد احتمال شرطی راس‌های موجود در متن را به حداکثر می‌رسانند نمایش رأس‌ها فقط به این مساله مربوط می‌شوند که در یک پنجره از یک دنباله متنی از رئوس ظاهر شوند اما اطلاعات جامعه و خصوصیات موجود در هر گره

---

[16] Representation Learning
[17] Computer vision
[18] Natural language processing
[19] Perozzi
[20] Grover
[21] Homophily
[22] Yang
[23] Pointwise mutual information
[24] Chen
[25] Semantical neighborhood
[26] Wang

که الگوهای سراسری شبکه را حفظ می‌کند نادیده گرفته می‌شود. چن و همکاران [۱۳] روشی را با اطلاعات گروهی با ارزش برای شبکه‌های در مقیاس بزرگ با در نظر گرفتن ساختارهای داخلی گروه‌ها و اطلاعات موجود در بین گروه‌ها، پیشنهاد کردند.

در سال‌های اخیر، مدل‌های موضوعی آماری، که هدف آن‌ها استفاده از متن مشاهده شده برای استنباط توزیع موضوع پنهان است با موفقیت برای استخراج مباحث در متون پیچیده استفاده شده است. مای[۲۷] و همکاران [۱۴] یک راه‌حل کلی از متن‌کاوی با ساختار شبکه به نام NetPLSA ارائه داده‌اند که بهینه‌سازی موضوع را در شبکه نشان می‌دهد. آن‌ها یال‌های بین اسناد را به عنوان یک قانون ایجاد کننده‌ی شبکه در نظر گرفته‌اند به گونه‌ای که اسناد مرتبط می‌توانند توزیع‌های مشابه موضوعی را به اشتراک بگذارند. NetPLSA هدفش استفاده از اطلاعات متنی شبکه از طریق مدل موضوعی است. این مدل بر استخراج ویژگی‌های جامعه از ویژگی‌های متن متمرکز است در حالی که ساختار توپولوژیکی شبکه فقط به عنوان کمکی استفاده می‌شود از این رو هنگام مدیریت اطلاعات متنی نویزی یا ناقص، نمایش‌های خوبی را نمی‌توان بدست آورد.

شی[۲۸] و همکاران [۱۵] یک روش تعبیه انتشار[۲۹] شبکه ارائه داده‌اند که هدفشان حل محدودیت‌های همچون تمایل به انتخاب گره‌ها با درجه‌ی بالا و همچنین بی‌توجهی به ساختار سراسری در شبکه‌های بسیار پیچیده در پیاده‌روی تصادفی است. چن[۳۰] و همکاران [۱۶] یک مدل قابل تعمیم ارائه داده‌اند که هم از اطلاعات یال و هم از اطلاعات مرکزیت[۳۱] گره‌ها برای یادگیری نمایش‌های برداری با ابعاد کم استفاده می‌کند که می‌تواند اطلاعات مختلف مرکزیت راس‌ها را حفظ کند. ژائو[۳۲] و همکاران [۱۷] یک چارچوب یکپارچه برای توصیه‌های اجتماعی و رفتاری با تعبیه‌سازی شبکه ارائه داده‌اند و یک رویکرد تعبیه‌سازی شبکه مشترک به عنوان یک مرحله قبل از آموزش برای نمایش‌های پنهان کاربران معرفی کرده‌اند. لی[۳۳] و همکاران [۱۸] یک مدل تعبیه‌سازی شبکه بدون نظارت برای رمزگذاری اطلاعات رابطه یال‌ها ارائه داده‌اند بنابراین نمایش ویژگی‌های رئوس می‌تواند بیشتر ضبط شود. وو[۳۴] و همکاران [۱۹] برای یادگیری نمایش‌های شبکه برای کارهای خاص، یک مدل LSTM توجه دوگانه چند وظیفه‌ای[35] ارائه کرده‌اند. این مدل می‌تواند ساختار، محتوا و اطلاعات برچسب را ضبط کند سپس نمایش‌های رئوس را با توجه به وظیفه پایین دست تنظیم کند. یوان[۳۶] و همکاران [۲۰] الگوریتمی با نام CONE ارائه کرده‌اند که از مدل‌سازی موضوعی برای تشخیص جامعه گره‌های موجود در شبکه و ساخت دنباله پیاده‌روی‌های تصادفی استفاده کرده‌اند. کیخا[۳۷] و همکاران [۲۱] الگوریتمی با نام CARE ارائه کرده‌اند که از روش تشخیص جامعه لوویان[۳۸] برای تشخیص جوامع گره‌های موجود در شبکه و ساخت دنباله پیاده‌روی‌های تصادفی استفاده کرده است.

بر خلاف تحقیقات فوق، این مقاله یک روش پیاده‌روی تصادفی مبتنی بر تحلیل متن گره‌ها پیشنهاد داده است. از مدل موضوعی جفت کلمه برای تحلیل مفهومی ارتباط گره‌ها و تشخیص جوامع در پیاده‌روی‌های تصادفی ایجاد شده استفاده می‌کند.

۳. متدولوژی

در این بخش، شرحی بر مدل موضوع آماری و الگوریتم پیشنهادی ارائه می‌شود. قبل از معرفی کل روش، با بحث در مورد بیان مساله تعبیه‌سازی شبکه شروع می‌شود.

۳.۱. بیان مساله

با فرض وجود یک گراف ویژگی[۳۹] $G = (V, E, T)$ طوریکه V مجموعه‌ای از رئوس است؛ $E \subseteq V \times V$ یال‌ها نشان‌دهنده‌ی روابط بین رئوس هستند و T نشان‌دهنده‌ی محتوای متن رئوس است؛ به طور خاص، اطلاعات متنی هر راس $v \in V$ مربوط به توالی کلمه $T_v =$

---

[27] Mei
[28] Shi
[29] Diffusion
[30] Chen
[31] Centrality
[32] Zhao
[33] Li
[34] Wu
[35] Multi-task dual attention LSTM model
[36] Yuan
[37] Keikha
[38] Louvian
[39] Attribute network

$(W1, W2, \dots\dots, W_{n_v})$ می‌باشد که در آن $n_v = |T_v|$ است. تعبیه‌سازی شبکه سعی در ایجاد یک ماتریسی از ویژگی‌ها با بعد کم با عنوان $\Phi \in R^{|V|\times d}$ برای شبکه دارد طوریکه $d \ll |V|$ بعد فضای نمایش پنهان را مشخص می‌کند. از نمایش پنهان بدست آمده می‌توان در برنامه‌های کابردی تجریه و تحلیل شبکه استفاده کرد.

۳.۲. پیاده روی تصادفی

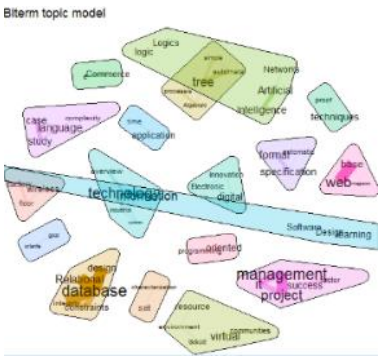
یک پیاده‌روی تصادفی از گره‌ی V_i با عنوان W_{V_i} را نشان داده می‌شود. یک فرایند تصادفی با متغییرهای تصادفی $W_{V_i}^1, W_{V_i}^2, \dots\dots, W_{V_i}^k$ است به گونه‌ای که $W_{V_i}^{k+1}$ راسی است که به طور تصادفی از همسایگان گره V_k انتخاب می‌شود. پیاده‌روی‌های تصادفی به عنوان یک اندازه‌گیری شباهت برای انواع مسائل در توصیه به محتوا[40] و تشخیص جامعه استفاده شده است. آنها همچنین پایه و اساس الگوریتم‌های حساس به خروجی هستند که از آن‌ها برای محاسبه اطلاعات ساختار محلی استفاده میشود [۴].

این ارتباط دارای ساختار محلی است که با استفاده از جریانی از پیاده‌روی‌های تصادفی کوتاه به عنوان ابزار اصلی روش برای استخراج اطلاعات از یک شبکه گراف محور است. علاوه بر گرفتن اطلاعات از جامعه، استفاده از پیاده‌روی‌های تصادفی به عنوان پایه‌ای برای الگوریتم، دو ویژگی مطلوب دیگر را نیز ارائه می‌دهد. در مرحله اول، موازی‌سازی اکتشافات محلی را آسانتر می‌کند. چندین پیاده‌روی تصادفی (در موضوعات مختلف، فرایندها یا ماشین‌های مختلف) می‌توانند همزمان بخش‌های مختلفی از یک گراف را مورد تجزیه قرار دهد. ثانیا، با تکیه بر اطلاعات به دست آمده از پیاده‌روی‌های تصادفی کوتاه، امکان ایجاد تغییرات کوچک در ساختار گراف بدون نیاز به اعتبار سراسری فراهم می‌شود [۴].

۳.۲. توزیع جامعه مبتنی برساختار و متن براسا مدل موضوعی جفت‌کلمه

مدل‌سازی موضوعی برای طبقه‌بندی و تجزیه و تحلیل اسناد مورد استفاده قرار می‌گیرند اما نشان داده شده است که مدل‌های موضوعی مانند LDA[41]، CTM[42] و STM[43] هنگامی که متن اسناد از سایت‌های رسانه‌های اجتماعی جمع‌آوری شده باشد با دردسر مواجه می‌شوند زیرا اغلب این اسناد دارای محدودیت کاراکتری هستند و طول سند آن‌ها به طور قابل توجهی کوتاه‌تر است. در این شرایط، روش‌های مدل‌سازی موضوعی نام برده شده دارای اثربخشی کمتری هستند.

رویکردی با نام مدل موضوعی جفت‌کلمه BTM پیشنهاد شده است. BTM به طور مستقیم فرآیند رخداد مشترک را مدل می‌کند. مدل موضوعی جفت‌کلمه ارتباط بین کلمات را با استفاده از مفهوم ترتیب کلمات در روند مدل‌سازی در نظر می‌گیرد. شکل (۱)، خروجی مدل موضوعی جفت‌کلمه بر روی مجموعه داده‌ای متن کوتاه را نشان می‌دهد [۲۲].



شکل ۱. **خروجی مدل موضوعی** جفت‌کلمه

همانطور که قبلاً ذکر شد جامعه نقشی اساسی در تعبیه‌سازی شبکه دارد. برای استخراج الگوهای سراسری جامعه، یک ترکیبی از مدل موضوع آماری و یک پیاده‌روی تصادفی شده است. مدل موضوعی جفت‌کلمه، یک مدل موضوعی آماری است که در این مقاله برای تخمین

[40] Content recommendation
[41] Latent dirichlet allocation
[42] Correlated topic model
[43] Structural topic model

توزیع‌های جامعه و تحلیل مفهومی ارتباط هر گره با گره‌های همسایه‌اش از روش مدل موضوعی جفت‌کلمه استفاده شده است. به منظور توضیح بهتر نحوه کارکرد BTM در شناسایی جامعه شبکه، مدل به جای کلمات و موضوعات از طریق رئوس و اجتماعات توصیف می‌شود. [۲۲].

روند مدل‌سازی BTM مربوط به توزیع مشروط متغیرهای پنهان و متغیرهای مشاهده شده در زیر است:

$$P_r(v_i|S) = \sum_{j=1}^{k} P_r(v_i|c_i = j)P_r(c_i = j|S) \quad (2)$$

که در آن $P_r(v_i|S)$ احتمال مشاهده راس v_i در توالی S است. $P_r(v_i|c_i = j)$ نمایانگر احتمال راس v_i در جامعه نهفته j است و $P_r(c_i = j|S)$ نشانگر احتمال انتخاب یک راس از جامعه j در توالی S است. BTM توزیع جامعه-راس $P_r(v|c)$ و توزیع جامعه-اجتماع $P_r(v|S)$ را از تعداد معینی از جوامع k با استفاده از نمونه‌گیری گیبس[۴۴] تخمین می‌زند. این روش به طور تصادفی یک جامعه محلی به هر راس v_i از توالی فعلی S اختصاص می‌دهد. سپس هر راس v_i مورد بررسی قرار می‌گیرد و جامعه آن j براساس احتمال $P_r(C_i = j|v_i, S, C_{-i})$ با استفاده از فرمول (۳) تا زمانی که پارامترهای مدل BTM همگرا شوند به روز می شود. [۲۲].

$$P_r(c_i = j|v_i, S, c_{-i}) \propto \frac{N_{vc}(v_i,j)+\beta}{\sum_v N_{vc}(v,j)+|v|\beta} \cdot \frac{N_{sc}(S,j)+\alpha}{\sum_c N_{sc}(S,c)+k\alpha} \quad (3)$$

در اینجا $N_{vc}(v_i,j)$ شماری از همه انتسابات راس-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 به جز انتساب فعلی t_i برای راس v_i ، و α و β پارامترهای هایپر هستند که به عنوان فاکتورهای هموارسازی[۴۵] برای BTM 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:

$$p_r(v_i|c_i = j) = \frac{N_{vc}(v_i,j) + \beta}{\sum_v N_{vc}(v,j) + |v|\beta} \quad (4)$$

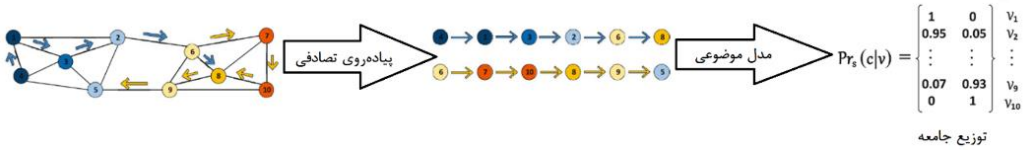
$$p_r(c_i = j|S) = \frac{N_{sc}(S,j) + \alpha}{\sum_c N_{sc}(S,c) + k\alpha} \quad (5)$$

با استفاده از این دو توزیع شرطی، این روش می‌تواند احتمال اینکه یک راس در دنباله‌ی گره‌ها برای تجمیع اسناد متعلق به هر جامعه باشد را تخمین می‌زند بدین ترتیب ساختار جامعه شبکه در این خصوص شناسایی می‌شود.

۳.۳. تعبیه‌سازی شبکه‌ی ویژگی مبتنی بر تحلیل متن گره‌ها

در این مقاله، یک استراتژی پیاده‌روی تصادفی مبتنی بر تحلیل متن گره‌ها ارائه شده است. برای یادگیری توزیع جامعه مبتنی بر ساختار[۴۶] و متن[۴۷] به ترتیب از دو مدل BTM با بیش از پارامترهای مختلف استفاده می‌شود. 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 راس‌ها تولید شده در طی این روش برای به روزرسانی مداوم، به مدل BTM مبتنی بر ساختار وارد می‌شوند. پس از تعداد مشخصی از تکرارها، الگوهای مقدماتی سراسری و جوامع مبتنی بر ساختار آموخته می‌شوند. در این فرآیند، از Skip-Gram برای به حداکثر رساندن تابع احتمال شرطی رئوس استفاده می‌شود بنابراین بردارهای نمایشی رأس از این توالی‌های پیاده‌روی تصادفی بدست می‌آیند.



شکل ۲. **توصیف گرافیکی از مدل موضوعی ساختاری برای تخمین احتمال گره‌ها در خصوص تجمیع اسناد هر گره**

[44] Gibbs
[45] smoothing
[46] structure-based
[47] text-based

محتوای متن یک راس با چندین همسایه در یک سند واحد جمع می‌شود و در نهایت به مدل BTM مبتنی بر متن وارد می‌شود. شبکه استنادی را به عنوان مثال در نظر بگیرید هر مقاله به عنوان یک راس و عنوان آن به عنوان محتوای متن در نظر گرفته می‌شود. فرض شده است که یک رابطه ذاتی بین عنوان مقاله و مقالاتی که به آن استناد می‌کند وجود دارد. این عناوین مرتبط جمع می‌شوند سپس متن مبتنی بر BTM می‌تواند برای کشف موضوعات صریح یا جوامع موجود در آن استفاده شود. بردار بازنمایی اطلاعات متنی توسط BTM مبتنی بر متن ارائه می‌شود که در نهایت با نمایش‌های مبتنی بر ساختار اصلی که بوسیله‌ی Skip-Gram تولید شده است می‌پیوندد و خروجی نهایی را تولید می‌کند. چارچوب DCB در شکل ۳ آورده شده است.

Algorithm 1 Framework of DCB

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.
Output: matrix of network representations: $\phi \in R^{|v| \times d}$
1: Sample ϕ from $u^{|v| \times d}$
2: **for** i = 0 to γ **do**
3: $\vartheta = Shuffle(V)$
4: **for** each vertex $v \in \vartheta$ **do**
5: S_v = RandomWalk(G, v, l)
6: $D_v = ContextAggregation(G, v, P_{r_t})$
7 SkipGram(ϕ, S_v, w)
8: **end for**
9: $P_{r_s} = BTM_s(S)$
10: $P_{r_t} = BTM_t(D)$
11: **end for**
12: $\phi = \phi \oplus P_{r_t}$
13: **return** ϕ

شکل ۳. چارچوب روش DCB

ماتریس U به طور تصادفی تولید می‌شود تا بردارهای راس را در خط ۱ الگوریتم ۱ مقداردهی اولیه کند. پیاده‌روی تصادفی معادل الگوریتم سنتی پیاده‌روی عمیق خواهد بود. از طریق پیاده‌روی تصادفی، به تفصیل در الگوریتم ۲ شرح داده شده است توالی‌های رئوس تولید شده و به مدل BTM مبتنی بر ساختار ارسال می‌شوند. به طور مشابه، ویژگی‌های متن یک راس با همسایگان آن جمع می‌شود و آن‌ها را به مدل BTM مبتنی بر متن وارد می‌شوند. پس از آن، احتمال رئوس متعلق به جامعه مبتنی بر ساختار P_{r_s} و متعلق به جامعه مبتنی بر متن P_{r_t} را می‌توان یاد گرفت. اولی برای هدایت فرآیند تولید توالی گره‌ها در خصوص تخمین احتمال هر گره برای تجمیع اسناد استفاده می‌شود در حالی که دومی در آخر به بردارهای نمایشی اصلی بدست آمده توسط Skip-Gram در خط ۱۲ ارسال می‌شود.

Algorithm 2 RandomWalk

Input: graph:G(V, E, T); Walk length: l.
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: $p(u|v) = 1$
6: **end for**
7: select a vertex u from neighbors based on random selection
8: **else**
9: backtrack in the sequence and select a vertex at random
10: **end if**
11: **end while**
13: **return** S_v

شکل ۴. الگورتیم پیاده‌روی تصادفی

دنباله رئوس با راس v آغاز می‌شود و چندین بار تکرار می‌شود تا اینکه طول دنباله به طول پیاده‌روی از پیش تعیین شده l برسد. خطوط ۴-۷ در شکل ۴ هسته اصلی روند پیاده‌روی تصادفی را نشان می‌دهد. درصورتی که راس فعلی v باشد احتمال شرطی راس بعدی بصورت تصادفی از بین همسایگان همان گره انتخاب می‌شود. اگر این راس فعلی همسایگی نداشته باشد این ترتیب دنباله پیاده‌روی را بسمت عقب برمی‌گردد و یک راس را به طور تصادفی انتخاب می‌کند.

Algorithm 3 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)< γ. log |v| **do**
3: **if** current vertex has neighbors **then**
4: **for** each neighbor vertex u of v **do**
5: $p(u|v) = 1 - \frac{1}{2} \cdot \sum_{i=1}^{k} |p_{r_s}(c_i|v) - p_{r_s}(c_i|u)|$
6: **end for**
7: select a vertex u from neighbors based on Roulette Wheel
8: $D_v = D_v \oplus T_u$
8: **else**
9: $D_v = D_v \oplus T_v$
10: **end if**
11: **end while**
13: **return** D_v

شکل ۵. الگوریتم جمع‌بندی متن

شکل ۵ 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 [۴۹].

$$\gamma . \log |v| \ll length(D_v) \quad (6)$$

خط شماره ۶ نحوه اختصاص احتمالات به ازای گره‌های همسایه را نشان می‌دهد. اگر گره‌ی همسایه در همان جامعه‌ای باشد که گره‌ی مبدا قرار دارد آنگاه دارای احتمالات نزدیک بهم هستند که مقدار تفریق آن‌ها مقدار کمی را ایجاد می‌کند و در نهایت احتمال گره‌هایی که با گره‌ی مبدا در یک جامعه قرار دارند بیشتر از گره‌هایی خواهد بود که در یک جامعه‌ی مشترک با گره‌ی مبدا نیستند. . انتخاب رئوس بعدی متکی به استراتژی انتخاب براساس انتخاب چرخ رولت[48] الگوریتم ژنتیک است [۲۳]. به طوری که رئوس در یک جامعه قرار دارند احتمال انتخاب بالاتری را خواهند داشت.

با این حال، رئوس ممکن است حاوی اطلاعات مشخصه کمی، حتی فقط چند برچسب باشند. با توجه به این واقعیت که رئوس در همان جامعه کاملاً به هم پیوسته‌اند و ویژگی‌های آن‌ها بسیار شبیه به هم است ویژگی‌های متنی یک رأس به همسایگان خود در یک سند واحد تجمع می‌یابد که ضمن حفظ مجاورت مرتبه اول، محدودیت طول اسناد برآورده می‌شود. براساس الگوریتم ۳ همسایگان با توزیع جامعه با ثبات‌تری را برای تجمیع متن اسناد انتخاب می‌شوند.

۳.۴. تحلیل پیچیدگی

در الگوریتم پیشنهادی، فرآیند آموزش شامل دو بخش است که به ترتیب مربوط به مدل Skip-Gram و مدل BTM است. به طور خاص، پیچیدگی Skip-Gram در پیاده‌روی عمیق برابر با $O(n\,\gamma\,l\,logn)$ است [۴]، که در آن n تعداد کل راس‌ها، γ تعداد پیاده‌روی‌های تصادفی به ازای هر گره و l طول پیاده‌روی تصادفی است. برای مدل BTM مبتنی بر متن، پیچیدگی آن $O(n\,\gamma\,k\,t)$ است. در اینجا t میانگین طول اطلاعات متنی هر رأس است که بزرگتر از $\gamma \cdot \log n$ است. در مجموع، پیچیدگی محاسباتی DCB $O\,(n\,\gamma\,(l\log n + kt))$ است.

۴. آزمایشات

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

۴.۱. روش‌های پایه بکار گرفته شده برای مقایسه

پیاده‌روی عمیق الگوریتم پیشقدم در خصوص تعبیه‌سازی شبکه است که از پردازش زبان طبیعی برای تعبیه‌سازی شبکه استفاده می‌کند، که پیاده‌روی‌های تصادفی روی شبکه‌ها ایجاد می‌کند و از Skip-Gram برای به دست آوردن نمایش هر رأس استفاده می‌کند. علاوه بر این، پیاده‌روی عمیق را می‌توان به عنوان بخش خاصی از الگوریتم DCB در نظر گرفت [۴]. CARE الگوریتمی در خصوص تعبیه‌سازی شبکه آگاه به جامعه است و اطلاعات مربوط به جامعه را بوسیله‌ی روش لوویان بدست می‌آورد. و در نهایت پیاده‌روی‌های حاصله بوسیله‌ی روش Skip-Gram به بردارهای نمایشی با بعد کم تبدیل می‌شوند [۲۱]. CONE الگوریتمی در خصوص تعبیه‌سازی شبکه است که اطلاعات جامعه را بوسیله‌ی مدل‌های موضوعی بدست آورده و پیاده‌روی تصادفی تولید شده را با بکار بردن روش Skip-Gram به بردارهای نمایشی با بعد کم تبدیل می‌کند. [۲۰]

۴.۲. مجموعه داده‌ها

Cora شامل ۲۷۰۸ مقاله یادگیری ماشین از ۷ کلاس و ۵۴۲۹ یال در بین مقالات است. هر رأس یک مقاله را نشان می‌دهد و روابط استنادی بین اسناد یک شبکه پیچیده معمولی را تشکیل می‌دهد [۲۴]. DBLP V12 شامل ۴ میلیون مقاله و ۴۵ میلیون یال بین مقالات است و تاریخ این مجموعه داده برابر با ۲۰۲۰/۰۴/۰۹ است. در این مقاله از دو زیرگراف با تعداد ۲۰۰۰ و ۵۰۰۰ گره برای اجرا استفاده شده است [۲۵]. در این مقاله از محتوی عنوان هر مقاله بعنوان اطلاعات ویژگی استفاده شده است. در خصوص مجموعه داده Cora عناوین استخراج شده از مجموعه داده‌ی اصلی دارای مقادیر گم شده بودند که در این خصوص از لینک مقالات بعنوان جایگزینی در این باره استفاده شده است [۲۴].

---

[48] Roulette wheel

جدول شماره ۱. آمار مجموعه داده‌های استفاده شده برای آزمایشات

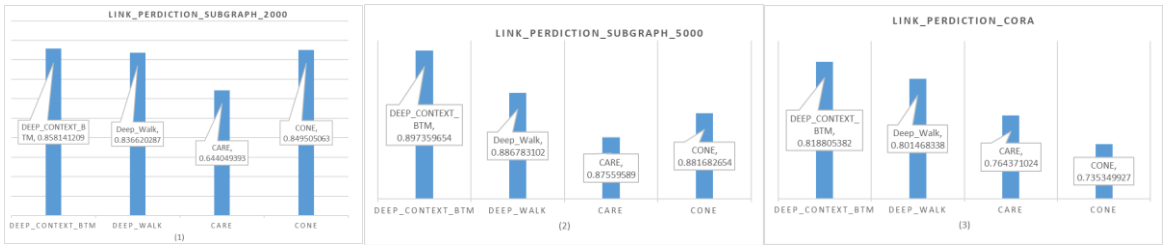
مجموعه داده‌ها	گره‌ها	یال‌ها	برچسب‌ها
Cora	۲۷۰۸	۵۴۲۹	۷
Dblp_2000	۲۰۰۰	۴۰۱۳	۴
Dblp_5000	۵۰۰۰	۱۱۵۸۷	۴

۴.۳. پارامترها

بعد بردار نمایش d = 128 برای همه مجموعه داده‌های بالا تنظیم شده 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 استفاده شده برای DCB، CARE، CONE با مقادیر استفاده شده برای پیاده‌روی عمیق مطابقت دارد. علاوه بر این، Mm = 0.2 و Ms = 0.3 در الگوریتم‌های CONE مقداردهی شده است. در تمام موارد بالا مقدار متغییر k برابر با ۱۴ در نظر گرفته شده است.

۴.۴. پیش‌بینی یال

پیش‌بینی یال یک وظیفه در یادگیری بانظارت است که تلاش می‌کند پیش‌بینی کند که آیا یالی بین رئوس وجود دارد یا خیر. پنجاه درصد از یال‌های شبکه‌ی اصلی بصورت تصادفی حذف می‌شود تا عملکرد مدل‌ها مورد ارزیابی قرار گیرد. با این فرض که مابقی شبکه متصل می‌ماند. یک معیار ارزیابی استاندارد، منطقه زیر منحنی[۴۹] (AUC) اتخاذ شده است [۲۶] که نشان‌دهنده احتمال شبیه‌تر بودن رئوس بالقوه متصل نسبت به موارد نامربوط است.



شکل ۶. امتیازات ACU بر روی معیار پیش‌بینی یال

در شکل ۶ مشاهده می‌شود که روش DCB به طور مداوم از تمام روش‌های پایه دیگر در این سه مجموعه داده بهتر عمل می‌کند. CARE عملکرد ضعیفی دارد و CONE نیز در مقایسه با سایر الگوریتم‌های مرتبط با متن، عملکرد ضعیفی دارد. با این حال، پیاده‌روی عمیق الگوریتمی است که از گراف نمونه‌برداری می‌کند که دارای مساله نمونه‌برداری ناکافی در شبکه‌های متراکم دارد. به دلیل سادگی استراتژی پیاده‌روی تصادفی اصلی و محدودیت مدل کیسه‌ای از کلمات[۵۰]، پیاده‌روی عمیق زمانی که با شبکه‌های متراکم با اطلاعات متنی غنی سروکار دارند در پیش‌بینی پیوند چندان مؤثر نمی‌تواند عمل کند. نتایج تجربی عملکرد برتر مدل ما را نشان می‌دهد که به طورکامل از ساختار توپولوژیکی شبکه و ویژگی‌های گره‌ها استفاده می‌کند به خصوص زمانی که برای شبکه‌های متراکم اتخاذ می‌شود.

۴.۶. طبقه‌بندی گره‌ها

طبقه‌بندی راس برای ارزیابی کیفیت نمایش‌های به‌دست‌آمده استفاده می‌شود طوریکه رگرسیون لجستیک[۵۱] تنظیم‌شده با L2 [۲۷] به‌عنوان طبقه‌بندی‌کننده نظارت‌شده استفاده می‌شود و نسبت آموزش از ۱۰٪ تا ۹۰٪ متغیر است. در آزمایش‌ها، رئوس برچسب‌گذاری‌شده را به‌طور تصادفی به عنوان داده‌های آموزشی و رئوس باقی‌مانده را به عنوان داده‌های آزمایشی انتخاب می‌شود. دقت[۵۲]، فراخوانی[۵۳]، امتیازات Micro-F1 و Macro-F1 را به عنوان معیارهای ارزیابی در آزمایشات در نظر گرفته شده است. فرض کنید تعداد نمونه‌هایی که به طور صحیح در یک کلاس طبقه‌بندی شده‌اند مثبت واقعی (TP)، تعداد نمونه‌هایی که به اشتباه در این کلاس طبقه‌بندی شده‌اند مثبت کاذب

[49] Area Under Curve
[50] bag-of-words
[51] Logistic regression
[52] precision
[53] recall

(FP) و تعداد نمونه‌های طبقه‌بندی شده در این کلاس در کلاس‌های دیگر، منفی واقعی (TN) است. دقت و فراخوانی به ترتیب به عنوان TP/(TP+FP) و TP/(TP+TN) تعریف می‌شوند. - امتیاز F1 ترکیب وزنی از دقت و یادآوری است که به صورت زیر تعریف می‌شود:

$$F1 - score = 2.\frac{precision.recall}{precision + recall} \quad (7)$$

جدول ۲.

precision (%) of vertex classification on subset of Cora									
% Labled vertices	10%	20%	30%	40%	50%	60%	70%	80%	90%
Deep_Walk	0.620267	0.735636	0.751928	0.776447	0.779706	0.796268	0.78179	0.810533	0.81112
DEEP_CONTEXT_BTM	0.674852	0.743308	0.775674	0.782438	0.804739	0.803491	0.805691	0.827026	0.855566
CARE	0.740948	0.743416	0.740367	0.757877	0.741338	0.769592	0.774685	0.759104	0.750292
CONE	0.401248	0.60314	0.622189	0.637459	0.661083	0.67522	0.67428	0.673484	0.670309

جدول ۳.

Recall (%) of vertex classification on subset of Cora									
% Labled vertices	10%	20%	30%	40%	50%	60%	70%	80%	90%
Deep_Walk	0.664069	0.731426	0.754219	0.776	0.781388	0.79428	0.776138	0.802583	0.808118
DEEP_CONTEXT_BTM	0.678835	0.740655	0.770042	0.782154	0.804284	0.802583	0.801968	0.824723	0.856089
CARE	0.719032	0.736041	0.738924	0.746462	0.739291	0.75738	0.762608	0.741697	0.723247
CONE	0.460623	0.603599	0.616561	0.627077	0.655835	0.668819	0.664207	0.662362	0.667897

جدول ۴.

Micro-F1 (%) of vertex classification on subset of Cora									
% Labled vertices	10%	20%	30%	40%	50%	60%	70%	80%	90%
Deep_Walk	0.664069	0.731426	0.754219	0.776	0.781388	0.79428	0.776138	0.804428	0.808118
DEEP_CONTEXT_BTM	0.678835	0.740655	0.770042	0.782154	0.804284	0.802583	0.801968	0.824723	0.856089
CARE	0.719032	0.736041	0.738924	0.746462	0.739291	0.75738	0.762608	0.741697	0.723247
CONE	0.460623	0.603599	0.616561	0.627077	0.655835	0.668819	0.664207	0.662362	0.667897

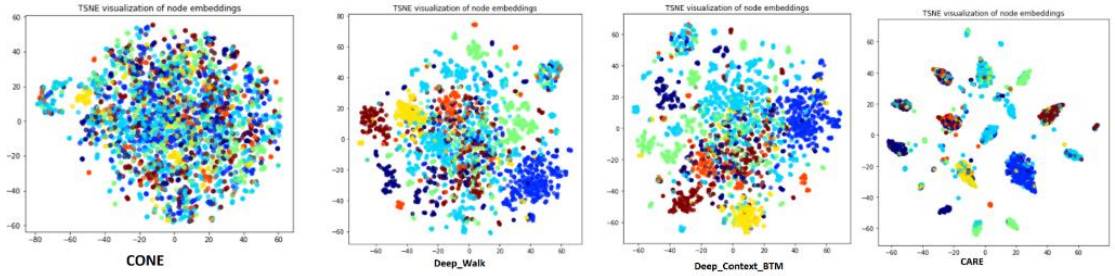
جدول ۵.

Macro-F1 (%) of vertex classification on subset of Cora									
% Labled vertices	10%	20%	30%	40%	50%	60%	70%	80%	90%
Deep_Walk	0.583304	0.71307	0.742073	0.756243	0.763717	0.777506	0.764953	0.796252	0.798318
DEEP_CONTEXT_BTM	0.651602	0.71335	0.748199	0.76854	0.790261	0.794474	0.79461	0.816848	0.839823
CARE	0.701254	0.711586	0.716973	0.720391	0.703544	0.733319	0.738022	0.720704	0.707439
CONE	0.359383	0.581234	0.559166	0.612954	0.643715	0.659506	0.656507	0.657173	0.638063

بدین ترتیب، امتیاز Macro-F1 با میانگین امتیاز F1 همه دسته‌ها محاسبه می‌شود در حالی که امتیاز Micro-F1 با میانگین دقت و یادآوری همه نمونه‌ها محاسبه می‌شود. آزمایشات برای ۱۰ بار تکرار شده است و میانگین دقت طبقه‌بندی با نسبت آموزشی متفاوت در مجموعه داده Cora در جداول ۲، ۳، ۴ و ۵ آمده است. طووریکه DCB به طور قابل توجهی بهتر از پیاده‌روی عمیق، CARE، CONE است. DCB در هنگام استفاده از ۹۰ درصد رئوس برچسب‌گذاری شده برای آموزش از پیاده‌روی عمیق بهتر عمل می‌کند که ضرورت استفاده از اطلاعات متنی یا ویژگی‌های موجود در گره‌ها در تعبیه‌سازی شبکه را نشان می‌دهد.

۴.۷. 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 بوسیله مدل‌های تعبیه‌سازی متفاوت با بردارهای با کم بعد نشان داده می‌شوند و سپس این بردارهای رئوس با استفاده از t-SNE در یک فضای دو بعدی نگاشت می‌شوند [۲۸]. تجسم شبکه Cora هر نقطه نشان دهنده یک مقاله است و توسط الگوریتم‌های مختلف تعبیه‌سازی شبکه به فضای دو بعدی نگاشت شده است. رنگ رئوس نشان دهنده تقسیم‌بندی‌های مختلف انتشارت است.



شکل ۷. تجسم شبکه‌های استنادی Cora بوسیله الگوریتم‌های مورد آزمایش

شکل۷ تجسم شبکه به دست آمده از الگوریتم‌های مختلف تعبیه‌سازی را نشان می‌دهد. از آنجایی که عناوین مقالات، که معمولاً از کمتر از ۱۰ کلمه تشکیل شده‌اند و 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 روش CONE و DCB می‌توان مشاهده کرد

که روش DCB بهتر از CONE عمل کرده است و این تاثیر مدل موضوعی بهتری است در روش DCB نسبت به CONE بکار برده شده است. برای پیاده‌روی عمیق، مقالات تقریباً براساس تقسیمات انتشار آنها گروه‌بندی شده‌اند اگرچه آنها به صورت خطی قابل تفکیک نیستند. با توجه به استفاده از اطلاعات جامعه، عملکرد CARE بهتر است. با این حال، مقالات در همان بخش‌ها به خوشه‌های کوچک تبدیل شده‌اند و فشردگی در یک بخش بسیار زیاد است.

۴.۸. حساسیت پارامترها

تأثیر تعداد جوامع k در جدول ۶ نشان داده شده است. پارامتر k از ۹ تا ۲۴ متغیر بوده و مشاهده می‌شود که منحنی برای روش DCB نسبتاً پایدار است. این امر نشان دهنده‌ی این است که تعداد جوامع k تنها تأثیر کمی بر عملکرد پیش‌بینی پیوند دارد. اگرچه باید خاطر نشان کرد که تعداد جوامع (موضوعات) k بر عملکرد مدل‌های موضوعی تأثیر زیادی دارد.

جدول ۶. تاثیر پارامتر k بر روی امتیاز AUC از وظیفه‌ی پیش‌بیتی یال

K	Link_Perdiction_SUBGRAPH_2000
9	0.868418727
14	0.858141209
19	0.854366051
24	0.871499457

۵ نتیجه‌گیری

در این مقاله، یک روش جدید جامعه‌گرا و مبتنی بر تحلیل متن در خصوص تعبیه‌سازی شبکه توسعه داده شده است. یک استراتژی پیاده‌روی تصادفی مبتنی بر تحلیل متنی گره‌ها پیشنهاد شده است تا به طور مشترک اطلاعات مربوط به مجاورت، جامعه و خصوصیات گره‌ها را حفظ کند. روش DCB بر کاستی‌های پژوهشی این حوزه غلبه کرده و به کارایی بالایی در تعبیه‌سازی شبکه مبتنی بر تحلیل مفهومی متن دست یافته است. نتایج تجربی در شبکه‌های مبتنی بر مفاهیم دنیای واقعی، اثربخشی و استحکام DCB را در مقایسه با سه روش پایه پیاده‌روی عمیق، CARE و CONE نشان می‌دهد. روش DCB بروی یک نوع گره تمرکز دارد. 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.

منابع

[1] Goyal, P., Ferrara, E., 2018. Graph embedding techniques, applications, and performance: A survey, Knowl.Based Syst. 151 78–94.

[2] Liben-Nowell, D., Kleinberg, J., 2007. The link-prediction problem for social networks, J. Am. Soc. Inf. Sci. Technol. 58 (7) 1019–1031.

[3] Bhagat, S., Cormode, G., Muthukrishnan, S., 2011. Node classification in social networks, in: Social Network Data Analytics, pp. 115–148.

[4] B. Perozzi, R. AlRfou, S. Skiena, 2014. Deepwalk: Online learning of social representations, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 701–710.

[5] Grover, A., Leskovec, J., 2016. node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855–864.

[6] Huang, X., Li, J., Hu, X., 2017. Accelerated attributed network embedding, in: Proceedings of the SIAM International Conference on Data Mining, pp. 633–641.

[7] Li, C., Li, Z., Wang, S., Yang, Y., Zhang, X., Zhou, J., 2017. Semi-Supervised Network Embedding, in: Proceedings of the 22nd International Conference on Database Systems for Advanced Applications, pp. 131–147.

[8] Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval, Inf. Process. Manag. 24 (5) 513–523.

[9] Brugere, I., Gallagher, B., BergerWolf, T.Y., 2018. Network structure inference, a survey: Motivations, methods, and applications, ACM Comput. Surv. 51 (2) 24.

[10] Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013)1798–1828.

[11] Chen, Z., Cai, T., Chen, C., Zheng, Z., Ling, G., 2019. SINE: Side Information Network Embedding, in: Proceedings of the 24th International Conference on Database Systems for Advanced Applications , pp. 692–708.

[12] Wang, D., Cui, P., Zhu W., 2016, Structural deep network embedding, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1225–1234.

[13] Chen, J., Zhang, Q., Huang, X., 2016. Incorporate group information to enhance network embedding, in: Proceedings of the 25th ACM International Conference on Information and Knowledge Management, pp. 1901–1904.

[14] Mei, Q., Cai, D., Zhang, D., Zhai, C., 2008. Topic modeling with network regularization, in: Proceedings of the 17th International Conference on World Wide Web, pp. 101–110.

[15] Shi, Y., Lei, M., Yang, H., Niu, L., 2019. Diffusion network embedding, Pattern Recognition. 88 518–531.

[16] Chen, H., Yin, H., Chen, T., Nguyen, Q.V.H., Peng, W.-C., Li, X., 2019. Exploiting centrality information with graph convolutions for network representation learning, in: Proceedings of the 35th IEEE International Conference on Data Engineering, pp. 590–601.

[17] Zhao, W., Ma, H., Li, Z., Ao, X., Li, N., 2019. SBRNE: An Improved Unified Framework for Social and Behavior Recommendations with Network Embedding, in: Proceedings of the 24th International Conference on Database Systems for Advanced Applications, pp. 555–571.

[18] Li, Q., Zhong, J., Li, Q., Cao, Z., Wang, C., 2019. Enhancing network embedding with implicit clustering, in: Proceedings of the 24th International Conference on Database Systems for Advanced Applications, pp. 452–467.

[19] Wu, L., Wang, D., Feng, S., Zhang, Y., Yu, G., 2019. MDAL: Multi-task Dual Attention LSTM Model for Semi supervised Network Embedding, in: Proceedings of the 24th International Conference on Database Systems for Advanced Applications, pp. 468–483.

[20] Yuan, Gao., Maoguo, Gong., Yu, Xie., Hua, Zhong, 2019. Community-oriented attributed network embedding, Knowledge-Based Systems, 0950-7051.

[21] Keikha, M.M., Rahgozar, M., Asadpour, M., 2018. Community aware random walk for network embedding, Knowl. Based Syst. 148 47–54.

[22] Yan, X., Guo, J., Lan, Y., and Cheng, X., *A Biterm Topic Model for Short Texts*, in Proceedings of the 22^{nd} international conference on World Wide Web, Rio de Janeiro, Brazil, pp. 1445–1455, 2013.

[23] Whitley, D., 1994. A genetic algorithm tutorial, Stat. Computer. 4 (2) 65–85.

[24] McCallum, A.K., Nigam, K., Rennie, J., Seymore, K., 2000. Automating the construction of internet portals with machine learning, Inf. Retr. 3 (2) 127–163.

[25] https://www.aminer.org/citation

[26] Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) 29–36.

[27] Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C., 2008. LIBLINEAR: A library for large linear classification, J. Mach. Learn. Res. 9 (8) 1871–1874.

[28] Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) 2579–2605.