7th International Conference on Web Research
May 19-20, 2021; Tehran, IRAN ICWR2021

Conformance Evaluation of Topic Modeling Approaches on Web-Based Short Text Dynamic Graph Databases

Mohadeseh Taherparvar[1], Fatemeh Ahmadi Abkenari[2], Pyman Bayat [3]

[1] Ph.D. Candidate, Department of Computer Engineering and Information Technology, Islamic Azad University of Rasht, Iran
mtaherparvar@phd.iaurasht.ac.ir

[2] Associate Professor, Department of Computer Engineering and Information Technology, Payam-E-Noor University of Rasht, Iran
Fateme.Abkenari@pnu.ac.ir

[3] Associate Professor, Department of Computer Engineering and Information Technology, Islamic Azad University of Rasht, Iran
bayat@iaurasht.ac.ir

Abstract

Topic modeling is a popular analytical approach for extracting topics from textual data and clustering them. There are many methods for topic modeling that consider the types of relationships and constraints for different types of datasets. Many researchers have been interested in Dirichlet's Latent Analysis (LDA) modeling method because of its flexibility and compatibility. But choosing this method on complex and dynamic datasets brings many challenges. Due to the rapid development of social networks and the existence of dynamic and short text databases, the aim of this research is to investigate the feasibility of using the best topic modeling appeoach based on evaluation criteria such as topic coherence, runtime, exclusivity and perplexity. In this paper, various approaches for topic modeling on dynamic short text datasets are analyzed. Such datasets can have a variety of applications. Article-related datasets, social media datasets and user feedback databases about a product offered by a business are among them. Due to the graph-based nature of the dataset used in this study (DBLP dataset), the results obtained from our experiments, helps a lot in the challenging problem of identifying communities in the field of graph analysis. Proper identification of communities can be effective in finding appropriate influential nodes in research areas based on customer based graph analysis such as viral marketing. The results of employing various topic modeling methods on the DBLP database and the node type of article title and the evaluation of the results with the mentioned topic evaluation criteria show the stability and compatibility of the Biterm method on this database.

Keywords: Community detection, Content analysis, Machine learning, Text mining, Topic modeling.

7th International Conference on Web Research
May 19-20, 2021; Tehran, IRAN ICWR2021

ارزیابی میزان تطابق راهکارهای مدلسازی موضوعی بر پایگاه‌های داده تحت وب گراف‌محور متن‌کوتاه پویا

محدثه طاهرپرور[1*]، فاطمه احمدی آبکناری[2] و پیمان بیات[3]

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

[2] استادیار، گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشکده فنی و مهندسی، دانشگاه پیام نور مرکز رشت
Fateme.Abkenari@pnu.ac.ir

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

چکیده

مدل‌سازی موضوعی یک ابزار تحلیلی محبوب برای استخراج موضوع از داده‌های متنی و خوشه بندی داده های پیکره های متنی است. روش‌های زیادی برای مدل‌سازی موضوعی وجود دارد که انواع روابط و محدودیت‌ها را در انواع مجموعه داده‌ها در نظر می‌گیرند. بسیاری از پژوهشگران به روش مدل‌سازی تحلیل پنهان دریکله[1] به دلیل انعطاف‌پذیری و سازگاری آن علاقه‌مند هستند. اما انتخاب این روش در خصوص مجموعه داده‌های پیچیده و خاص با چالش های بسیار همراه است. نظر به گسترش شبکه های اجتماعی و وجود پایگاه های داده پویا و متن کوتاه، بررسی امکان پذیر بودن استفاده از بهترین روش مدل سازی موضوعی بر اساس معیارهای ارزیابی همچون انسجام موضوع[2]، زمان اجرای مدل، انحصارطلبی[3] و میزان حیرت[4] مدل هدف پژوهش حاضر است. در این مقاله، رویکردی‌های مختلفی از روش های مدل‌سازی موضوعی در خصوص مجموعه داده متنی کوتاه پویا مورد تجزیه و تحلیل قرار گرفته است. مجموعه داده متن کوتاه پویا می‌تواند کاربردهای متنوعی داشته باشد به عنوان مثال، مجموعه داده‌های مربوط به موضوع مقالات، مجموعه داده‌های گردآوری شده از رسانه‌های اجتماعی، مجموعه داده‌های نظرات کاربران در خصوص محصول جدید ارائه شده توسط یک شرکت تجاری و موارد دیگر. با توجه به گراف محور بودن پایگاه داده مورد استفاده در این پژوهش، موضوع بدست آمده از خروجی اعمال روش های مدلسازی موضوعی، کمک شایانی در مساله چالش برانگیز تشخیص جوامع در حوزه تحلیل گراف می کند. تشخیص مناسب جوامع می‌تواند در یافتن گره‌های تاثیرگذار مناسب در بازاریابی ویروسی موثر باشد. نتایج بدست آمده از بررسی انواع روش های مدل سازی موضوعی بر پایگاه داده DBLP و نوع گره موضوع مقاله و ارزیابی نتایج با معیارهای ارزیابی موضوعی نشان از پایداری و تطابق روش بایترم بر روی این پایگاه داده دارد.

کلمات کلیدی

تحلیل محتوا، تشخیص جامعه، مدل سازی موضوعی، متن کاوی، یادگیری ماشینی.

۱- مقدمه

مدل‌سازی موضوع یک روش یادگیری ماشینی است که قادر به اسکن مجموعه ای از اسناد، تشخیص الگوهای کلمات و عبارات در آن‌ها و خوشه‌بندی خودکار گروه‌های کلمه‌ای و عبارات مشابه است که مجموعه اسناد را به بهترین وجه توصیف می‌کند. تجزیه و تحلیل داده‌ها در خصوص پست‌های شبکه‌های اجتماعی، ایمیل‌ها، چت‌ها و موارد دیگر، کاری چالش برانگیز است. تجزیه و تحلیل متن با استفاده از هوش‌مصنوعی از الگوریتم‌های

متنوعی برای پردازش زبان طبیعی استفاده می‌کند یکی از آن‌ها تجزیه و تحلیل موضوع است که برای تشخیص خودکار موضوعات از متن استفاده می‌شود. مدل‌سازی موضوع نوعی مدل‌سازی آماری برای کشف "موضوعات" انتزاعی موجود در مجموعه اسناد است.

مدل‌سازی موضوعی، یک ابزار آماری محبوب برای استخراج متغیرهای نهان در مجموعه داده‌های بزرگ است [۱]. به ویژه برای استفاده در داده‌های متنی بسیار مناسب است. با این حال، برای تجزیه و تحلیل داده‌های بیوانفورماتیک، داده‌های اجتماعی [۲] و داده‌های محیطی نیز مورد استفاده قرار می‌گیرد. اخیرا مدل‌سازی موضوعی در مبحث پیمایش گراف یا پیاده‌روی تصادفی و کمک به تشخیص اجتماعات بر روی گراف‌ها استفاده می شود. اگر به عنوان مثال گرافی در خصوص فعالیت افراد در شبکه‌های اجتماعی موجود باشد هر گره از گراف مربوط به فردی در شبکه اجتماعی است آن گره می‌تواند دارای ویژگی‌های دیگری به صورت متن مانند توضیحی در خصوص تخصص و حوزه‌ی فعالیت آن فرد در شبکه‌ی اجتماعی باشد. یکی از روش‌هایی که می‌تواند به یافتن حوزه‌ها یا اجتماعات مرتبط بین گره‌ها در گراف بیانجامد مدل‌سازی موضوع است. مدل‌سازی موضوعی با ارائه‌ی موضوعات مرتبط با هم بین گره‌ها می‌تواند به ساخت اجتماعات مجزا بین گره‌های یک گراف کمک کند.

جامعه نقش اساسی در درون‌سازی شبکه[۵] دارد. برای استخراج الگوهای سراسری جامعه و ارائه‌ی پیمایش مناسب برای گراف شبکه، ترکیبی از مدل موضوعی آماری و پیاده‌روی تصادفی می‌تواند بکار رود. به منظور توضیح بهتر نحوه‌ی کارکرد مدل موضوعی در شناسایی جوامع در شبکه، مدل به جای کلمات و موضوعات از طریق رئوس و اجتماعات توصیف می‌شود.

برخی از روش‌های مدل‌سازی موضوعی عبارت‌اند از: تحلیل معنایی نهفته[۶] LSA [۳]، تجزیه و تحلیل معنایی نهفته احتمالی[۷] PLSA [۴]، تخصیص پنهان دیریکله LDA [۵]، مدل موضوع همبسته[۸] CTM [۶]، مدل موضوع ساختاری[۹] STM [۷] و مدل موضوع بایترم[۱۰] BTM [۸]. هنوز پژوهش های زیادی برای بهبود الگوریتم‌ها در راستای درک متن کامل اسناد در جریان است.

در این مقاله، برخی از ادبیات علمی پر استناد راجع به فرآیندهای مدل‌سازی موضوعی برای مجموعه داده‌های متنی کوتاه مورد بررسی قرار خواهد گرفته است. بنابراین، هدف مقاله بررسی تطابق یا عدم تطابق روش های موجود مدلسازی موضوعی بر روی مجموعه داده گراف محور شامل متن های کوتاه است.

در ادامه، مقاله به صورت زیر سازماندهی شده است: در بخش ۲، پژوهش های مربوطه در خصوص مدل‌های موضوعی مورد بحث قرار می‌گیرد. در بخش ۳، آزمایشات مربوطه در این خصوص ارائه می‌شود و در بخش ۴، نتیجه‌گیری و زمینه های آتی پژوهش بیان می‌شود.

۲- مروری بر پژوهش‌های پیشین

مدل‌سازی موضوعی در ابتدا در دهه ۱۹۸۰ توسعه یافت و از حوزه موضوع "مدل‌سازی احتمالات مولد" مشتق شده است [۶].

مدل موضوعی نوعی مدل احتمالی است که در سال‌های اخیر در زمینه علوم کامپیوتر از متن‌کاوی و بازیابی اطلاعات منشعب شده است. سرآغاز مدل‌های موضوعی، روشی با نام تحلیل معنایی نهفته LSA است [۳]. LSA یک مدل احتمالی نیست و براساس این روش، تحلیل معنایی نهفته احتمالی PLSA توسط هافمن (۱۹۹۹) پیشنهاد شد[۴]. پس از PLSA، تخصیص پنهان دیریکله LDA توسط آقای بلای و همکاران پیشنهاد شده است [۵]. امروزه، تعداد مدل‌های احتمالی که مبتنی بر LDA هستند در حال رشد است.

۲-۱-رویکردهای مدل‌سازی موضوع

در این بخش برخی از روش‌های مدل‌سازی موضوعی که با کلمات، اسناد و موضوعات سروکار دارند مورد بررسی قرار می‌گیرد. علاوه بر این، ایده کلی هر یک از این روش‌ها، و در صورت وجود مثالی برای این روش‌ها ارائه شده است.

۲-۱-۱- تحلیل معنایی نهفته

تحلیل معنایی نهفته LSA یک روش یا تکنیک در زمینه پردازش زبان طبیعی است. هدف اصلی تحلیل معنایی نهفته LSA ایجاد نمایش مبتنی بر بردار برای متون و ساختن محتوای معنایی است [۳].

۲-۱-۲- تحلیل معنایی نهفته احتمالی

تجزیه و تحلیل معنایی نهفته احتمالی PLSA روشی است که بعد از روش LSA برای رفع برخی از معایب موجود در LSA معرفی شده است. هدف اصلی PLSA شناسایی و تمایز بین زمینه‌های مختلف استفاده از کلمات بدون مراجعه به فرهنگ لغت یا اصطلاح‌نامه است [۴].

۲-۱-۳- تخصیص پنهان دیریکله

یکی از قدیمی‌ترین و متداول‌ترین روش مدل‌سازی موضوعی تخصیص پنهان دیریکله یا LDA است، که توسط بلای، ان‌جی و جردن (۲۰۰۳) توسعه یافته است [۵].

LDA چندین فرضیه در محاسبات خود ارائه می‌دهد که ممکن است برای تولید مدل‌های موضوعی دقیق یا واقع‌گرایانه با توجه به مجموعه داده‌های دنیای واقعی موثر نباشند. فرض اول این است که مقدار k ثابت و معلوم است [۱]؛ برای اکثر برنامه‌های کاربردی در دنیای واقعی که از مدل‌سازی موضوعی استفاده می‌کنند، تعداد ایده‌آلی از موضوع برای یک گروه قبل از تولید مدل وجود ندارد. فرض دوم مطرح شده در LDA این است که

همه مباحث از یکدیگر مستقل هستند و بنابراین این مدل ارتباط بین مباحث را به حساب نمی‌آورد [۵].

۲-۱-٤- مدل موضوع همبسته

تحول عمده بعدی در روش‌های مدل موضوعی مدل موضوع همبسته CTM است که توسط بلای (۲۰۰۶) پیشنهاد شده است [۵]. مشابه روش قبلی، این رویکرد اسناد منحصربه‌فرد را به عنوان ترکیب نسبت موضوعات در نظر می‌گیرد و از یک روند تولید تقریباً یکسان پیروی می‌کند [۵].

روش CTM تناسب بهتری با داده‌ها دارد و توانایی پشتیبانی از تعداد بیشتری از موضوعات را دارد [۵]. مدل CTM همچنین فرض می‌کند k یک مقدار شناخته شده است و بنابراین روشی را برای تعیین k بهینه برای یک مجموعه ارائه نمی‌دهد [۶].

مدل CTM مبین‌تر از LDA است و روابط واقع بینانه‌تری را بین موضوعات مختلف درون یک سند ارائه می‌دهد. CTM در مقایسه با LDA نسبت به k حساسیت کمتری دارد. یکی از محدودیت‌های قابل توجه این روش این است که همبستگی‌ها فقط می‌توانند بین دو موضوع به طور همزمان انجام شوند. به طور کلی، این مدل یک روش مناسب برای داده‌هایی است که دارای موضوعات کاملاً همبسته هستند، به عنوان مثال روش CTM برای مجموعه‌ای که از مقاله‌های یک ژورنال جمع‌آوری شده، به خوبی کار می‌کند.

۲-۱-۵- مدل‌های موضوعی ساختاری

یکی از محبوب‌ترین تکنیک‌هایی که در سال‌های اخیر ظهور کرده است، مدل‌سازی موضوع ساختاری STM است. STM شباهت زیادی به LDA دارد، اما از متاداده ها مربوط به اسناد (مانند نام نویسنده، تاریخی که سند تولید شده است و ...) استفاده می‌کند تا انتساب کلمات به موضوعات پنهان در یک مجموعه را بهبود بخشد. به علاوه، برخلاف مدل LDA، موضوعات موجود در STM می‌توانند با هم مرتبط باشند. مدل STM پیچیده‌تر از LDA است، بنابراین استفاده از آن نیز دشوارتر است [۷].

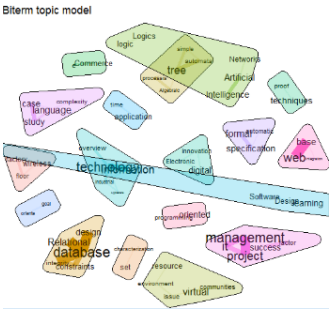
۲-۱-٦-مدل‌های موضوعی بهینه‌شده برای متن‌کوتاه

مدل‌سازی موضوعی برای طبقه‌بندی و تجزیه و تحلیل اسناد مورد استفاده قرار می‌گیرند اما نشان داده شده است که مدل‌های موضوعی مانند LDA، CTM و STM هنگامی که سند 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 [۲]. در این شرایط، مدل‌سازی موضوعی دارای اثربخشی کمتری است.

رویکردی با نام مدل موضوع بایترم BTM پیشنهاد شده است. BTM به طور مستقیم فرآیند رخداد مشترک را مدل می‌کند. مدل بایترم ارتباط بین کلمات را با استفاده از مفهوم ترتیب کلمات در روند مدل‌سازی در نظر می‌گیرد [۸].

شکل (۱)، خروجی مدل موضوعی بایترم بر روی مجموعه داده‌ای متن کوتاه را نشان می‌دهد.

در ادامه جداول (۱) و (۲) بیان‌گر ویژگی‌ها و محدودیت‌های روش‌های مدل‌سازی موضوع است.



شکل (۱) : خروجی مدل موضوعی بایترم

جدول (۱) : مشخصات روش‌های مدل‌سازی موضوعی

نام روش	ویژگی‌ها
Latent Semantic Analysis (LSA) **تحلیل معنایی نهفته**	–در نظر گرفتن کلمات مترادف برای تعیین موضوع. – توجه به معنای کلمات. – دارای پیش زمینه آماری قوی [۳].
Probabilistic Latent Semantic Analysis (PLSA) **تحلیل معنایی نهفته احتمالی**	– تولید کردن هر کلمه از یک موضوع واحد. – بخوبی کنترل کردن کلماتی که دارای چندین معنا هستند. – آشکارسازی شباهت‌های موضعی به وسیله کلاس‌بندی واژه‌هایی که یک زمینه مشترک دارند [۹–۱۰–۱۱].
Latent Dirichelet Allocation (LDA) **تخصیص پنهان دریکله**	–نیاز به حذف دستی کلمات توقف. – در نظر نگرفتن همبستگی بین موضوعات [۵].
Correlated Topic Model (CTM) **مدل موضوع همبسته**	– استفاده از توزیع نرمال لجستیک برای ایجاد روابط بین عناوین. – مجاز دانستن وقوع کلمات در سایر عناوین و نمودار موضوع. – داشتن تتناسب بهتر روش CTM با داده‌ها. – توانایی پشتیبانی از تعداد بیشتری از موضوعات. – ارائه‌ی روابط واقع بینانه‌تری بین موضوعات مختلف درون یک سند [۶].
Structural Topic Model (STM) **مدل موضوع ساختاری**	– استفاده از فراداده‌های موجود درسند به عنوان متغیرهای مختلف در روش تخمین موضوع. – وجود ارتباط بین متغیرهای شناخته شده و موضوعات پنهان در کانون اصلی مطالعه [۱۳] .
Biterm topic model (BTM) **مدل موضوع بایترم**	– ازاستفاده از رخدادهای مشترک به عنوان آماری برای حمایت از تولید موضوع، به جای مدل‌سازی رخدادهای مشترک. – اهمیت نظم و ترتیب رخداد کلمات را بیشتر از رخداد مشترکشان می‌داند. –در خصوص متن کوتاه بخوبی عمل می‌کند [۱۴].

جدول (۲) : محدودیت‌های روش‌های مدل‌سازی

نام روش	محدودیت‌ها
Latent Semantic Analysis (LSA) **تحلیل معنایی نهفته**	–دشوار بودن تعیین تعداد مباحث. – استفاده از فرهنگ‌لغت برای تفسیر مقادیر اختصاص داده شده به کلمات. – نزدیک بودن کلمات مشابه از لحاظ معنایی. – ایجاد اختلال در کار توسط کلماتی که دارای بیش از یک معنا هستند [۳].
Probabilistic Latent Semantic Analysis (PLSA) **تحلیل معنایی نهفته احتمالی**	–رنج بردن از مساله‌ی بیش‌برازش[11] زمانیکه تعداد پارامترها با تعداد اسناد به صورت خطی رشد کنند [۱۲].
Latent Dirichelet Allocation (LDA) **تخصیص پنهان دریکله**	– آشکار کردن مشکلات پراکندگی هنگام مواجهه با واژگان بزرگ. – ثابت بودن مقدار k. – مشخص نبودن روش مناسب برای تعیین بهترین مقدار متغیر k برای اکثر روش‌های مدل‌سازی موضوعی، از جمله LDA. – ناتوانی LDA در نشان دادن روابط را در بین موضوعات [۵].
Correlated Topic Model (CTM) **مدل موضوع همبسته**	– نیازمند بودن به محاسبات زیاد. – بسیار زمانبر بودن. – حل نشدن مساله متغییر k بطور کامل [۶].
Structural Topic Model (STM) **مدل موضوع ساختاری**	– غیرقابل حل بودن مدل برای تعداد زیادی از متغیرها یا بیش از یک محتوای متغیر. – پیچیده بودن زیاد مدل [۱۳].
Biterm topic model (BTM) مدل موضوع بایترم	– با استفاده حداکثری از رخداد مشترک در متون کوتاه، اما اغلب روش بایترم نمی‌تواند بین اسناد کوتاه و بلند تفاوتی قائل شود [۱۴].

۳- آزمایشات

در این بخش، مجموعه داده مورد استفاده در این مقاله، تنظیم پارامترهای روش‌های مدل‌سازی و سپس نتایج حاصل از اجرای الگوریتم‌های مدل‌سازی موضوعی مانند LDA، CTM، STM و BTM بر روی مجموعه داده مورد استفاده ارائه شده است. در این سری از آزمایشات مقدار متغیر K (تعداد موضوعات بدست آمده از مجموعه داده‌ی متنی) در دو صورت ثابت و متغیر مقداردهی می‌شود.

۳-۱- روش‌های پایه

روش‌های پایه مورد استفاده برای مقایسه عبارتند از، مدل تخصیص پنهان دیریکله، مدل موضوع همبسته، مدل موضوع ساختاری، و مدل موضوع بایترم.

مدل تخصیص پنهان دریکله یکی از مدل‌های پرکاربرد در حوزه‌های مختلف هوش مصنوعی است. یک مورد از این حوزه‌ها، کاربرد مدل تخصیص پنهان دریکله در تعیین جوامع موجود در گراف‌های شبکه است [۱۵].

گره‌های موجود در شبکه می‌توانند به یک یا چند جامعه اختصاص داشته باشند. علاوه بر ارتباطات موجود بین گره‌ها در گراف می‌توان از ویژگی‌هایی همچون عناوین گره‌ها برای کمک به ایجاد جامعه استفاده کرد. به این معنا که گره‌های موجود در یک جامعه می‌تواند دارای عناوین نزدیک بهم باشند.

عناوین گره‌های شبکه می‌توانند شامل ویژگی‌های یک گره در شبکه های اجتماعی باشند مثلا اگر گره یک فرد در شبکه‌ی اجتماعی اینستاگرام یا توییتر است عنوان گره شامل توضیحات آن فرد در خصوص مهارتهایی است که خود را با آنها مطرح کرده است. این عناوین جزو متون کوتاه در نظر گرفته می‌شوند.

در این مقاله سعی شده است روش‌هایی برای مقاسیه انتخاب شوند که بعد از روش مدل‌سازی LDA ارائه شده‌اند تا بتوان انتخاب مناسبی در خصوص گراف‌های شبکه و تعیین جوامع موجود در آن‌ها داشت.

۳-۲- مجموعه‌ی داده

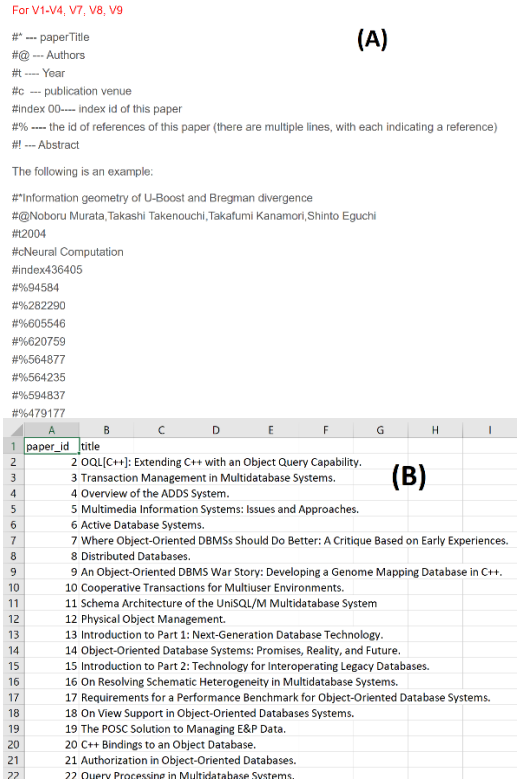
در این مقاله از مجموعه داده‌ی DBLP استفاده شده است. DBLP یک شبکه‌ی استنادی ناهمگن است. ناهمگن به این معنی است که گره‌ها و یال‌های موجود در این شبکه همه از یک نوع نیستند. مثالی از شبکه‌ی استنادی شامل سه نوع گره (به عنوان مثال، نویسندگان، مقالات و مجلات) و سه نوع رابطه می‌باشد (یک نویسنده مقاله را می‌نویسد، مقاله‌ای در یک مجله منتشر می‌شود و یک مقاله به مقاله دیگری استناد می‌کند). اگر دو مقاله در یک مجله منتشر شوند در یک کلاس قرار می‌گیرند. به همین ترتیب، در صورتی که بیشتر مقاله‌های آن‌ها به یک مجله اختصاص داشته باشد، دو نویسنده با یک برچسب، برچسب‌گذاری می‌شوند.

مجموعه داده DBLP می‌تواند به صورت پویا عمل کند. به عنوان مثال، گره‌های نویسنده در مجموعه داده DBLP می‌توانند بصورت پویا به مجموعه داده اضافه یا به‌روز شوند. در طی یک سال تعدادی مهر زمانی تعریف می‌شود و براساس این زمان‌ها مجموعه داده‌ی DBLP به‌روز می‌شود. اگر دو نویسنده به‌تازگی در یک مقاله با هم همکاری کرده باشند دو گره‌ی با نام‌های نویسندگان جدید با یال ارتباطی بین‌شان به گراف ارتباطات نویسندگان اضافه می‌شود و اگر نام نویسندگان از قبل در مجموعه داده وجود داشت فقط ارتباط جدید بین آن‌ها به‌روز می‌شود.

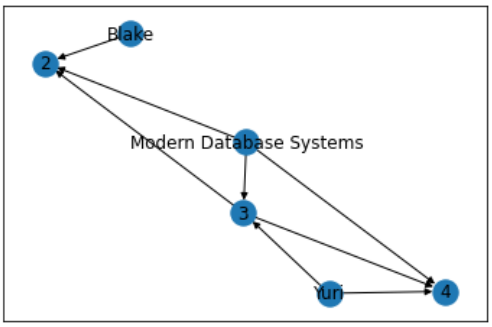
مجموعه داده مورد استفاده در این مقاله DBLP-Citation-network V9 است [16]. فرمت این مجموعه داده به صورت شکل (۲) است. فایل بدست آمده از مجموعه داده‌ی اصلی دارای بیش از یک میلیون داده‌ی موضوعی است که در این کار یک زیر مجموعه‌ی هزار عضوی مورد استفاده قرار گرفته است.

مجموعه‌داده‌ی DBLP یک مجموعه داده‌ی گراف‌محور از نوع NOSQL است که می‌توان گراف‌های متنوعی را توسط پایگاه داده‌های گراف‌محور مانند Neo4j و زبان‌های برنامه‌نویسی پایتون از آن استخراج کرد. در شکل (۳) می‌توان نمونه نمادینی از گراف استخراج شده را مشاهده کرد.

گرافی ک در این مقاله استفاده شده است از ارتباط استنادی مقالات استخراج شده است. موضوعات مرتبط با هر مقاله از این گراف استخراج شده



شکل (۲) : (A) فرمت مجموعه داده‌ی استنادی ناهمگن و (B) مجموعه داده متنی استخراج شده از مجموعه داده استنادی ناهمگن بوسیله کد پردازش متن.



شکل (۳) : گراف استخراج شده از مجموعه داده‌ی DBLP، Modern Database system **نام مجله،** Blake **و** Yuri **هر دو نویسنده مقاله و شماره‌های ۲، ۳ و ٤ شماره مقالات هستند.**

و بوسیله‌ی روش‌های مدل‌سازی موضوعی با هدف یافتن بهترین روش مدل‌سازی مورد تجزیه و تحلیل قرار گرفته است. برای استخراج این گراف از زبان‌برنامه‌سازی پایتون استفاده شده است. مجموعه‌ی داده‌ای استفاده شده در این مقاله شامل مجموعه‌ای از موضوعات مقالاتی است که به هم استناد شده‌اند و جزو متون کوتاه بحساب می‌آید.

۳-۳- تنظیم پارامترها

در این سری از آزمایشات انجام شده پارامتر K به دو صورت ثابت و متغییر مورد بررسی قرار گرفته است. K در حالت ثابت برابر با ۲۰ در نظر گرفته شده است و در حالت متغییر بین ۱۰ تا ۱۰۰ برای بررسی بهترین خروجی مدل تنظیم شده است. مقادیر ثابت و متغیر پارامتر k برای سنجش عملکرد مدل‌های موضوعی است. مقدار پارامتر بتا برابر با ۰.۰۱، پارامتر آلفا برابر با ۲.۵ و در ۱۰۰۰ تکرار اجرا شده است. مقادیر لحاظ شده برای پارمترهای آلفا و بتا برابر با مقدار پیش‌فرض در خصوص مدل استخراج موضوع بایترم است.

بستر اجرا زبان‌برنامه‌نویسی پایتون و R است. برای اجرای کدها از سیستمی با پردازنده‌ی اینتل Core i7 6500U با فرکانس ۲.۵GHz و حافظه‌ی اصلی ۸G استفاده شده است.

۳-٤- معیارهای ارزیابی

۳-٤-۱- انسجام موضوع

قبل از بیان معیار انسجام موضوع، توضیحی مختصری در خصوص معیار میزان حیرت مدل داده می‌شود. میزان حیرت یکی از معیارهای ارزیابی ذاتی است و به طور گسترده‌ای برای ارزیابی مدل‌های تحلیل متنی استفاده می شود. این متغییر نشان می‌دهد که چگونه یک مدل از داده‌های جدیدی که قبلاً مشاهده نکرده است متعجب می‌شود و احتمال میزان حیرت با استفاده از محاسبه احتمال-لاگ[۱۲] از یک مجموعه داده‌ی تست[۱۳] بدست می‌آید.

با تمرکز بر روی احتمال-لاگ می‌توان معیار میزان حیرت را با اندازه‌گیری احتمال بدست آمده از داده‌های جدید در مدلی که قبلا یادگرفته است، بدست آورد.

با این حال، احتمال بدست آمده از معیار میزان حیرت و قضاوت انسان اغلب به هم نزدیک نیستند. یعنی نتایج بدست آمده برای معیار میزان حیرت ممکن است موضوعات قابل تفسیر توسط انسان را به همراه نداشته باشد. این محدودیت انگیزه‌ای برای کار بیشتر در خصوص مدل‌سازی قضاوت انسان و در نتیجه معیار انسجام موضوع شد. معیار انسجام موضوع با اندازه‌گیری میزان شباهت معنایی بین کلمات، در خصوص یک موضوع واحد را بدست می‌آید.

براساس شکل (۴) با قرار دادن مقدار متغیر K=20 مقادیر معیار انسجام موضوع برای چهار روش مدل‌سازی موضوع مورد مقایسه قرار گرفته است و نتایج خروجی نشان می‌دهد که روش مدل‌سازی موضوع بایترم نسبت به بقیه روش‌ها خروجی بهتری از لحاظ انسجام موضوع ارائه داده است.

با توجه به شکل (۵) اکنون مقدار متغیر K را برای روش مدل‌سازی موضوعی بایترم متغییر در نظر گرفته می‌شود و با بررسی نتایج ارائه شده در شکل (۵) می‌توان دید که هر چقدر مقدار متغییر k بیشتر شود انسجام موضوعی بهتری بدست می‌آید.

۳-٤-۲- انحصار طلبی مدل

معیار انحصارطلبی معنایی اندازه انحصارطلبی کلمات برتر در موضوع را نشان می‌دهد. یعنی تا چه اندازه کلمات برتر یک موضوع به عنوان کلمات برتر سایر موضوعات ظاهر نمی‌شود یا به عبارتی اختصاصی بودن کلمات برتر در یک موضوع را بیان می‌کند. این مقدار متوسط میانگین کل کلمات برتر هر کلمه در مبحث را تقسیم بر مجموع احتمالات آن کلمه در همه مباحث می‌کند.

برای مقایسه مقدار انحصارطلبی معنایی در خروجی مدل‌ها، هر چقدر تفاوت بین مقادیر کمتر باشد، نشان می‌دهد کلمات برتر موجود در موضوعات می‌توانند در کلمات برتر سایر موضوعات ظاهر شوند. پس می‌توانند گفت خروجی بهتر خواهد بود که مقادیر انحصارطلبی بدست آمده برای موضوعات مختلف نسبت بهم دارای تفاوت مقداری بیشتری باشند.

براساس شکل (۶) مقادیر انحصارطلبی معنایی برای مقدار K=20 نشان داده شده است. همانطوری که از مقادیر بدست آمده در شکل شماره (۶) مشخص شده، نتایج انحصارطلبی بدست آمده در خصوص مدل موضوع بایترم نسبت بهم دارای تفاوت بیشتری هستند. پس در نتیجه می‌توان گفت مدل موضوعی بایترم نسبت به بقیه‌ی روش‌های مدل‌سازی موضوعی از لحاظ انحصارطلبی معنایی نتایج بهتری را ارائه داده است.

مقادیر بدست آمده برای میعار انحصارطلبی در مدل‌ها اکثرا دارای نتایج نزدیک بهم بوده. به همین دلیل برای درک بهتر نتایج از جدول استفاده شده است.

با توجه به شکل شماره (۷) اکنون مقدار متغیر K را برای روش مدل‌سازی موضوعی بایترم متغییر در نظر گرفته می‌شود و با بررسی نتایج ارائه شده در شکل (۷) می‌توان دید که هر چقدر مقدار متغییر k بیشتر شود معیار انحصارطلبی معنایی بهتری بدست می‌آید.

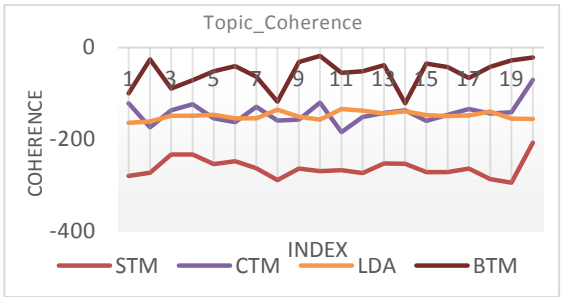
۳-٤-۳- زمان اجرا

زمان اجرا، به ازای یک بار اجرا هر مدل محاسبه شده است. در جدول ارائه شده در شکل (۸) زمان اجرای مدل‌های استخراج موضوع براساس ثانیه مورد بررسی قرار گرفته است. به دلیل تفاوت زیاد در زمان‌های بدست آمده از جدول برای نمایش بهتر استفاده شده است.

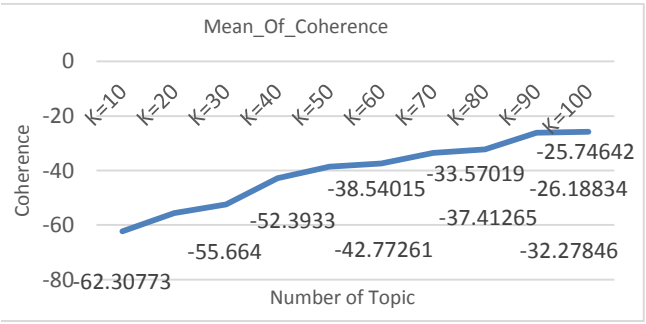
در نتیجه، از لحاظ زمانی مدل موضوع بیاترم بهترین عملکرد و مدل موضوع همبسته یا CTM دارای بدترین عملکرد است.

۳-٤-٤- تفسیر نتایج

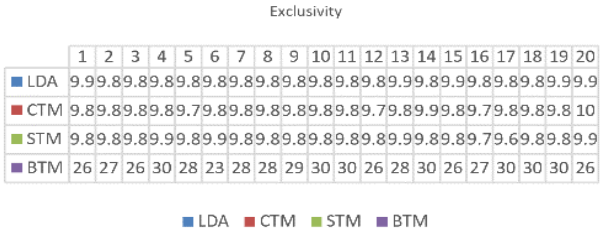
در حوزه‌ی تحلیل مدل‌های استخراج موضوع، مدلی بهتر است که موضوعات منسجم و منحصر به فردی را در زمان اجرای کوتاه ارائه دهد. با بررسی پارامترهای انسجام موضوع، انحصارطلبی و زمان اجرا برای روش‌های مدل استخراج موضوع تخصیص پنهان دریکله، مدل استخراج موضوع همبسته، مدل استخراج موضوع ساختاری و مدل استخراج موضوع بایترم می‌توان به



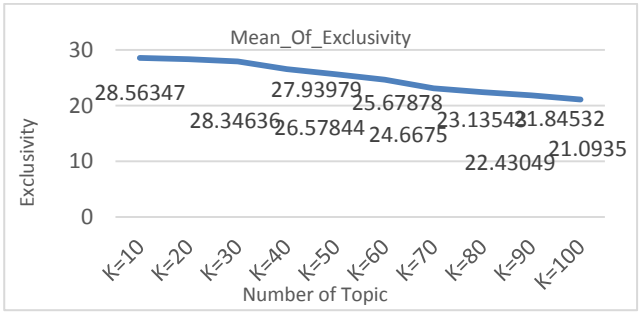
شکل (٤): مقایسه‌ی مدل‌های LDA، CTM، STM و BTM نسبت به معیار انسجام موضوع



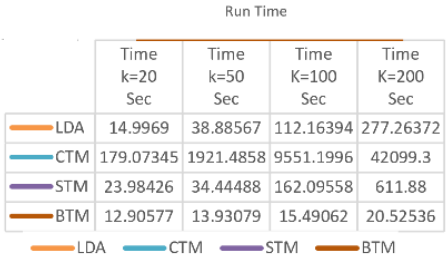
شکل (٥): میزان تغییر معیار انسجام موضوع نسبت به افزایش مقدار K مدل موضوعی بایترم



شکل (٦): مقایسه مدل‌های مدل‌های LDA، CTM، STM و BTM نسبت به معیار انحصارطلبی



شکل (۷): میزان تغییر معیار انحصارطلبی مدل نسبت به افزایش مقدار K در مدل موضوعی بایترم



شکل (۸): زمان اجرای مدل‌های موضوعی مدل‌های LDA، CTM، STM و BTM.

وضوح مشاهده کرد که مدل استخراج موضوع بایترم در خصوص مجموعه داده مرتبط با موضوع مقالات بدست آمده از مجموعه داده استنادی دارای بهترین عملکرد است.

۴- نتیجه‌گیری

این مقاله به بررسی میزان تطبیق‌پذیری استفاده از روش‌های مدل‌سازی موضوعی متناسب با پایگاه‌های داده متن‌کوتاه گراف‌محور می‌پردازد. از این رو ابتدا، مدل‌های موضوعی مورد نقد و بررسی قرار گرفت و ویژگی‌ها و محدویت‌های هر مدل بیان شد. مدل‌های موضوعی دارای کاربردهای مختلفی هستند یکی از مهمترین آن‌ها، استخراج معنایی بهترین موضوع از مجموعه‌ی داده‌ای متنی است. مجموعه داده DBLP در این مقاله مورد استفاده قرار گرفته است. براساس نتایج بدست آمده در این مقاله، روش مدل‌سازی موضوعی بایترم در خصوص مجموعه داده متن‌کوتاه دارای بهترین عملکرد بوده است.

در حیطه‌ی تحلیل گراف، تشخیص جوامع مساله‌ای چالش‌برانگیز است. گره‌های تاثیرگذار با ویژگی‌های نزدیک معمولا در یک جامعه قرار می‌گیرند. اگر گره‌های گراف مقالات باشند از روش‌های مدل‌سازی موضوعی می‌توان در حوزه پیمایش گراف و تشخیص جوامع بهره برد. دسته‌بندی گره‌ها براساس جوامع موجود در گراف شبکه می‌تواند به یافتن موثر گره‌های تاثیرگذار کمک شایانی کند.

در بازاریابی ویروسی انتخاب مناسب گره‌های تاثیرگذار می‌تواند در رشد صعودی سود شرکت‌های تجاری تاثیرگذار باشد. در پژوهش‌های آتی از مدل‌سازی موضوعی بایترم برای پیاده‌روی تصادفی بر روی گراف و تشخیص مناسب جوامع با هدف یافتن بهترین گره‌های تاثیرگذار برای رسیدن به بیشترین سود در بازاریابی ویروسی استفاده خواهیم کرد.

مراجع

[1] Blei, D. M., *Probabilistic topic models*, Commun ACM, vol. 55, no. 4, p. 77, Apr. 2012.

[2] Hong , L. and Davison, B. D., *Empirical study of topic modeling in Twitter*, in Proceedings of the First Workshop on Social Media Analytics - SOMA '10, Washington D.C., District of Columbia, pp. 80–88, 2010.

[3] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R., *Indexing by latent semantic analysis*, J Am Soc Inform Sci, vol. 41, no. 6, pp. 391–407, Sep. 1990.

[4] Hofmann, T., *Probabilistic Latent Semantic Indexing*, ACM SIGIR Forum, vol. 51, no. 2, p. 8, 2017.

[5] Blei, D. M., A. Ng, Y., and Jordan, M. I., *Latent Dirichlet Allocation*, J Mach Learn Res, vol. 3, no. 2003, pp. 993–1022, Jan. 2003.

[6] Lafferty , J. D., and Blei, D. M., *Correlated Topic Models*, in Advances in Neural Information Processing Systems 18, Y. Weiss, B. Schölkopf, and J. C. Platt, Eds. MIT Press, pp. 147–154, 2006.

[7] Roberts, M., Stewart, B., and Tingley, D, stm: R Package for Structural Topic Models, Journal of Statistical Software, Journal of Statistical Software , 91 (2), 1-40, 2019.

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

[9] Kakkonen, T., Myller, N., Sutinen, E., and Timonen, J., *Comparison of Dimension Reduction Methods for Automated Essay Grading*, Educational Technology &Society, 11 (3) , 275-288, 2008.

[10] Liu, S., Xia, C., and Jiang, X., *Efficient Probabilistic Latent Semantic Analysis with Sparsity Control*, IEEE International Conference on Data Mining, 905-910, 2010.

[11] Romberg, S., Hörster, E., and Lienhart, R., *Multimodal pLSA on visual features and tags*, *The Institute of Electrical and Electronics Engineers Inc.* , 414-417, 2009.

[12] Zhou, Z. Zhou, J., and Zhang, L., *Demand-adaptive Clothing Image Retrieval Using Hybrid Topic Model*, in Proceedings of the 2016 ACM on Multimedia Conference - MM '16, Amsterdam, The Netherlands, pp. 496–500, 2016.

[13] Wesslen, R., *Computer-Assisted Text Analysis for Social Science: Topic Models and Beyond* , arXiv:1803.11045v2 [cs.CL], 2018.

[14] Vorontsov, K., and Potapenko, A., *Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization*, in Analysis of Images, Social Networks and Texts, vol. 436, D. I. Ignatov, M. Y. Khachay, A. Panchenko, N. Konstantinova, and R. E. Yavorsky, Eds. Cham: Springer International Publishing, pp. 29–46, 2014.

[15] Gao, Y., Gong, M., Xie, Y., and Zhong, H., *Community-oriented attributed network embedding*, Knowledge-Based Systems, 2020.

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

زیرنویس‌ها

[1] Latent Dirichlet Allocation

[2] Topic coherence

[3] Exclusivity
[4] Perplexity
[5] Network embeding
[6] Latent semantic analysis
[7] Probabilistic Latent Semantic Analysis
[8] Correlated topic model
[9] Structural topic model
[10] Biterm topic model
[11] Overfitting
[12] Log-likelihood
[13] Held-out