Estimating Number of Topics in Topic Modeling on Persian Research Articles

Niloofar Mozafari
PhD in Artificial Intelligence; Assistant Professor; Regional Information Center for Science and Technology; Shiraz, Iran;
Email: mozafari@ricest.ac.ir

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 38 | No. 4 | pp. 1343-1366
Summer 2023
https://doi.org/10.22034/jipm.2023.701394

Received: 14, Nov. 2021 **Accepted: 06, Sep. 2022**

Abstract: This article presents a method to find the number of topics in Persian research articles, which is actually one of the main challenges in topic modeling. It is the process of automatically recognizing topics in a text with the aim of discovering hidden patterns.

This study has estimated the number of topics for Persian research articles using two approaches. The first is based on the greedy search and later uses Renormalization theory, which is a mathematical formalism to construct a procedure for changing the scale of the system so that the behavior of the system preserves. Also, the execution time of both algorithms on Persian academic articles has been compared with each other.

The findings indicate that the renormalization approach predicts the number of topics in Persian research articles with the lower time complexity in comparison to the greedy based approach.

The approach based on Renormalization has high efficiency for estimating the number of topics in Persian academic articles.

Keywords: Renormalization Theory, Rényi Entropy, Grid Search, Latent Dirichlet Allocation

تخمین تعداد موضوعات در مدل‌سازی موضوعی روی مقالات علمی فارسی

نیلوفر مظفری

دکتری هوش مصنوعی؛ استادیار؛ مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری؛ شیراز، ایران؛
mozafari@ricest.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۰/۰۸/۲۳ | پذیرش: ۱۴۰۱/۰۶/۱۵ | مقاله برای اصلاح به مدت ۷ ماه و ۲۸ روز نزد پدیدآور بوده است.

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۸ | شمارهٔ ٤ | ص‌ص ۱۳٤۳-۱۳٦٦
تابستان ۱٤۰۲
https://doi.org/10.22034/jipm.2023.701394

چکیده: این مقاله روشی را برای یافتن تعداد موضوعات در مقالات علمی فارسی ارائه می‌دهد که یکی از چالش‌های اصلی در مدل‌سازی موضوعی است و در واقع، فرایند تشخیص خودکار موضوعات در یک متن با هدف کشف الگوهای پنهان است.
پژوهش حاضر از نوع کاربردی است که با مقایسه دو روش، یکی مبتنی بر «گریدی» و دیگری مبتنی بر نظریه بازبهنجاری پارامتر تعداد موضوعات را برای مقالات نشریات فارسی تخمین می‌زند. روش «گریدی» با تعریف یک معیار برای ارزیابی مدل موضوعی و به‌دست آوردن این معیار با توجه به مقادیر مختلف تعداد موضوعات می‌تواند تعداد موضوعات بهینه را تخمین بزند. الگوریتم دیگر مبتنی بر نظریه بازبهنجاری است که در واقع، یک فرمولاسیون ریاضی برای ساخت یک رویّه برای تغییر مقیاس سیستم تحت بررسی است به‌صورتی که رفتار سیستم حفظ شود و تغییری در روند آن ایجاد نشود. با استفاده از این نظریه و استفاده از اطلاعات مرحله قبل می‌توان تعداد موضوعات را با سرعت تخمین زد. همچنین، مدت زمان اجرای هر دو الگوریتم روی مقالات نشریات مختلف فارسی، ارائه و با یکدیگر مقایسه شده است. یافته‌ها نشان‌دهنده کارایی روش مبتنی بر نظریه بازبهنجاری در تخمین تعداد موضوعات موجود در مقالات نشریات فارسی است.
نتایج نشان می‌دهد که روش مبتنی بر نظریه بازبهنجاری نسبت به روش «گریدی» با سرعت بالاتری می‌تواند تعداد موضوعات را تخمین بزند. از این روش می‌توان پارامتر تعداد موضوعات در مقالات نشریات فارسی را تخمین زد که در نهایت، به مدل‌سازی موضوعی نشریات فارسی با توجه به مقالات چاپ‌شده در آن منجر می‌شود.

کلیدواژه‌ها: نظریه بازبهنجاری، آنتروپی رونو، جست‌وجوی گریدی، توزیع دیریکله

۱. مقدمه

در گذشته اطلاعات روی رسانه‌های فیزیکی مانند کتاب ذخیره می‌شد. امروزه با گسترش و توسعه اطلاعات الکترونیکی از یک طرف، و رشد سریع اطلاعات الکترونیکی از طرف دیگر، با حجم اطلاعات قابل دسترس بسیاری روبه‌رو شده‌ایم. به‌منظور دسترس‌پذیری هرچه بهتر این اطلاعات، به روش‌های جدیدی نیاز هست (گیلوری ۱۳۷۹). سازماندهی اطلاعات از جمله روش‌های تجزیه‌وتحلیل اسناد به‌منظور دسترس‌پذیری هرچه بهتر آن است. یکی از چالش‌های پیش رو در جهت پردازش و سازماندهی اسناد با حجم بالا این است که به‌گونه‌ای بتوان این متون را نمایش داد که هم تا آنجا که ممکن است حجم داده کاهش پیدا کند تا ذخیره و پردازش راحت‌تر انجام گیرد و هم مفهوم به‌درستی منتقل گردد. مدل‌های موضوعی راه‌حلی برای این چالش هستند.

مدل‌سازی موضوعی به الگوریتم‌هایی گفته می‌شود که با پردازش متن، موضوعات مختلف موجود در آن متن (حتی به‌صورت پنهان) را استخراج می‌نمایند (Blei 2012). بنابراین، در مدل‌سازی موضوعی هر سند با توجه به موضوعات موجود که می‌تواند موضوعات پنهان هم باشد، تفسیر و سازماندهی می‌گردد. در این مدل‌ها، هر متن به‌صورت توزیعی از موضوعات و هر موضوع هم به‌صورت توزیعی از واژگان تعریف می‌گردد (Kherwa & Bansal 2020).

از مهم‌ترین چالش‌های موجود در مدل‌سازی موضوعی، به‌دست آوردن تعداد موضوعات موجود در یک متن است؛ به‌صورتی که عملکرد نهایی مدل به این پارامتر وابسته است. پژوهش‌های پیشین به‌طورعمده از روش جست‌وجوی «حریصانه»[1] برای به‌دست آوردن تعداد موضوعات موجود در یک متن استفاده کرده‌اند (Koltcov & Ignatenko 2020)؛ بدین‌صورت که با تعریف یک معیار، عملکرد مدل موضوعی را روی متن با توجه به پارامترهای مختلف سنجیده، و در نهایت، پارامتری که مدل با آن بهترین عملکرد را دارد، به‌عنوان تخمینِ پارامتر مورد بررسی در نظر گرفته‌اند (Röder, Both & Hinneburg 2015)، (Stevens et al., 2012). این روش با اینکه می‌تواند تخمین مناسبی از پارامتر مورد بررسی و یا همان تعداد موضوعات موجود در یک متن ارائه دهد، ولی پیچیدگی زمانی بسیار بالایی دارد و کُند است.

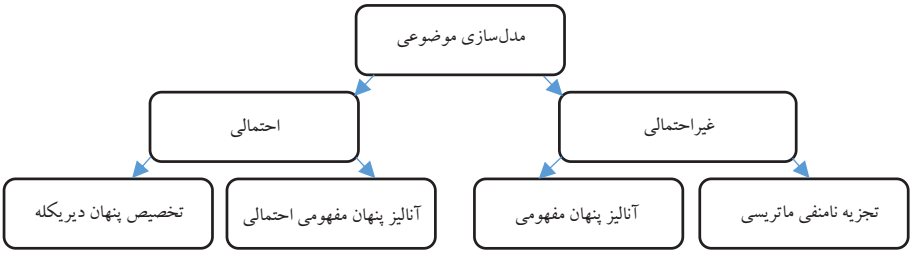
1. Greedy

نظر به اهمیت مدل‌سازی موضوعی در پردازش متون و بالاخص مقالات فارسی، هدف این پژوهش بررسی دو روش به‌دست آوردن تعداد موضوعات موجود در مقالات نشریات است. همچنین، مدت‌زمان اجرای هر دو الگوریتم روی مقالات نشریات مختلف فارسی ارائه و با یکدیگر مقایسه شده است. بنابراین، این پژوهش به‌دنبال پاسخ به سؤالات زیر است:

- ◇ تعداد موضوعات موجود در نشریات مورد بررسی چقدر است؟
- ◇ مقایسهٔ زمانی میان دو الگوریتم تخمین تعداد موضوعات به چه صورت است؟

۲. ادبیات پژوهش

مدل‌سازی موضوعی به‌صورت کلی به دو گروه مدل‌های احتمالی و غیراحتمالی دسته‌بندی می‌شوند (شکل ۱). مدل‌های احتمالی از یک توزیع احتمالی و به‌روز کردن پارامترهای آن در تکرارهای مختلف استفاده می‌کنند. روش‌های غیراحتمالی نیز در واقع، روش‌های جبری فاکتورگیری ماتریس[1] هستند (Kherwa & Bansal 2020). در ادامه، هر یک از این دسته‌ها و روش‌های آن مورد بررسی قرار خواهد گرفت.



شکل ۱. طبقه‌بندی مدل‌سازی موضوعی (Kherwa & Bansal 2020)

روش‌های غیراحتمالی همان روش‌های جبری فاکتورگیری ماتریس هستند که برای اولین‌بار با مفهوم آنالیز پنهان مفهومی[2] (Deerwester et al. 1990) و تجزیه نامنفی ماتریسی[3] (Lee & Seung 2001) مطرح شدند. از آنجا که این دو روش بر روش کیسه کلمات[4] مبتنی هستند؛ در هر دو پیکره به ماتریس اصطلاح-سند[5] تبدیل می‌شوند. مدل‌های احتمالی به‌منظور بهبود مدل‌های جبری ارائه گردیدند (Blei 2012).

1. matrix factorization
2. Latent Semantic Analysis (LSA)
3. Non-negative matrix factorization (NNMF)
4. Bag of words
5. term-document matrix

آنالیز پنهان مفهومی یک روش جبری بر اساس تجزیهٔ مقدارهای منفرد[1] است (Kherwa & Bansal 2017). این روش از فرضیهٔ توزیع[2] به‌منظور یافتن ارتباط معنایی میان اسناد و کلماتی که در آن‌ها وجود دارد، بهره می‌گیرد (Dudoit, Fridlyand & Speed 2002). در فرضیهٔ توزیع، اصطلاح‌ها با معنای مشابه در متون بسیاری در کنار یکدیگر به‌کار می‌روند. آنالیز پنهان مفهومی کاربردهای زیادی در حوزه‌هایی مانند بازیابی اطلاعات، آنالیز شبکه‌های اجتماعی[3] و خلاصه‌سازی متون[4] دارد.

تجزیهٔ نامنفی ماتریسی که کاربردهای بسیار زیادی در حوزه‌های کاهش ابعاد، تشخیص الگو، پردازش تصویر و مدل‌سازی زبان دارد (Kherwa & Bansal 2020)، برای تقریب داده‌های نامنفی ذخیره‌شده در یک ماتریس منفی، دو ماتریس نامنفی دیگر ایجاد می‌کند. هر تجزیه، تعبیرهای مختلفی را از ساختار ضمنی داده‌ها آشکار می‌کند که این تعبیرها از نظر ریاضی هم‌ارز هستند. بنابراین، از این روش می‌توان برای مدل‌سازی موضوعی استفاده کرد؛ به‌صورتی که با تجزیه ماتریس، موضوعات یا اطلاعات نهفته در متن کشف گردد.

همان‌طور که شکل ۱، نشان می‌دهد، از دسته روش‌های مبتنی بر مدل‌های احتمالی به دو مدل آنالیز پنهان مفهومی احتمالی[5] و تخصیص پنهان دیریکله[6] می‌توان اشاره کرد. مدل آنالیز پنهان مفهومی احتمالی یک تکنیک آماری نوین برای بررسی داده‌های هم‌رخداد به‌صورت احتمالی است (Hofmann 2013) که از مدل آماری منظر[7] بهره می‌برد. در این مدل فرض بر این است که بردار کلمات و اسناد شرطی مستقل هستند. لازم به ذکر است که تعداد متغیرهای پنهان که در اینجا موضوعات است، کمتر از تعداد متون و کلمات است.

تخصیص پنهان دیریکله یک روش مبتنی بر نظریه De Finetti (2017) است که ساختار آماری درونی و میانی سند را از طریق توزیع توأم در نظر می‌گیرد. در این روش فرض بر این است که هر سند حاوی چندین موضوع و هر موضوع، توزیعی روی واژگان است و کلماتی که مربوط به یک موضوع هستند، در آن موضوع دارای احتمال بالایی هستند. به‌عبارت دیگر، این روش، موضوعات را از داده‌ها یا پیکره یاد می‌گیرد.

---

1. (SVD)
2. distributional hypotheses
3. social network analysis
4. text summarization
5. Probabilistic Latent Semantic Analysis (PLSA)
6. Latent Dirichlet Allocation (LDA)
7. aspect model

پژوهش‌های دیگری هم در سال‌های اخیر برای مدل‌سازی موضوعی ارائه شده است که ارتباط میان واژگان را در سطحی محلی‌تر بررسی می‌کنند. در نخستین روشی که در این دسته ارائه گردید و به نحوی الهام‌بخش دیگر روش‌های این دسته بود، فرض بر این است که هر واژه افزون بر موضوع خود، به موضوع واژهٔ پیشین خود نیز وابسته است (Barbieri et al. 2013). تعمیمی بر این روش، توسط «ونگ» و همکارانش انجام گردید که در آن هر واژه بر مبنای موضوع خود می‌تواند تصمیم بگیرد که آیا با واژه قبلی یک ترکیب را تشکیل دهد یا خیر (Wang, McCallum & Wei 2007). یک مدل موضوعی احتمالاتی مبتنی بر روابط محلی واژگان در پنجره‌های همپوشان توسط «رحیمی، زاهدی و مشایخی» (۱۳۹۷) ارائه گردید. در پژوهشی دیگر فرض مشابهی در نظر گرفته شد و افزون بر آن، فرض شد که یک سلسله‌مراتب از موضوعات وجود دارد و هر واژه، مسیری مشخص را در این سلسله‌مراتب طی می‌کند تا توسط یک موضوع خاص تولید گردد (Yang et al. 2015).

پژوهش‌های دیگری نیز ارائه شده‌اند که بر خلاف روش‌های قبلی محدود به واژه پیشین نبوده و نتایج آن‌ها برای ترکیباتی با طول‌های متفاوت گزارش شده است (Noji, Mochihashi & Miyao 2013 و Sato & Nakagawa 2010). با توجه به تنک بودن داده‌ها، این پژوهش‌ها می‌بایست روی مجموعه داده‌های بسیار بزرگ آموزش داده شوند تا بتوان به نتایج ارائه‌شده اعتماد کرد. با توجه به بار محاسباتی بسیار زیاد این روش‌ها برای مجموعه داده‌های بسیار بزرگ با مشکل روبه‌رو شده و غیرعملی می‌شوند.

در داخل ایران نیز پژوهش‌های مختلفی روی مدل‌سازی موضوعی به‌عمل آمده است که به‌طور عمده از نتایج مدل‌سازی موضوعی استفاده شده است. از آن جمله می‌توان به «زمانی، دیانت و صادق‌زاده» اشاره کرد که با استفاده از روش آنالیز معنایی پنهان احتمالاتی که از منابع دانش محدود از محتویات متون و حذف کلمات زائد و غیرمفید حاصل شده است، به دسته‌بندی متون پرداختند (۱۳۹۳). آن‌ها همچنین از توابع ریاضی موجود در آنالیز پنهان مفهومی احتمالی استفاده کردند تا حجم بار محاسباتی را تا حدودی کاهش دهند. در پژوهشی دیگر از مدل تخصیص پنهان دیریکله به‌عنوان یک روش آنالیز معنایی برای استخراج ویژگی در دسته‌بندی اسناد استفاده شده است. از مشخصه‌های اصلی مدل ارائه‌شده، محاسبه احتمال عنوان بودن کلمات است که بر اساس تعداد تکرار کلمهٔ مورد نظر با دیگر کلمات محاسبه می‌شود. در این پژوهش در نهایت، از الگوریتم فرابتکاری

ژنتیک برای خوشه‌بندی نهایی استفاده گردیده است (شکری و معصومی ۱۳۹۵).

به‌عنوان پژوهشی دیگر که از مدل‌سازی موضوعی استفاده می‌کند، می‌توان به پژوهش انجام‌گرفته توسط «دامی و الیکایی» اشاره کرد. آن‌ها یک الگوریتم مدل‌سازی موضوعی برای رویدادهای خبری ارائه دادند که بر اساس یادگیری عمیق افزایشی عمل می‌کند. روش ارائه‌شده توسط آن‌ها یک چارچوب سه-مرحله‌ای و مقیاس‌پذیر مبتنی بر یادگیری عمیق ارائه می‌کند که برای یادگیری و اطلاع از یک سلسله‌مراتب از رویدادها در مورد یک موضوع به‌کار می‌رود و بر اساس رویدادهایی است که به محض وقوع، مرتبط با آن موضوع باشد (۱۳۹۶). لازم به ذکر است که روش ارائه‌شده توسط آن‌ها فقط می‌تواند روی رویدادهای خبری عمل کند.

«اسدی قادیکلایی» و همکاران نیز از نتایج مدل‌سازی موضوعی روی مقالات پژوهشگران ایران در حوزه غدد درون‌ریز و متابولیسم در پایگاه استنادی وب علوم استفاده کردند. نتایج آن‌ها نشان‌دهندهٔ عملکرد قابل قبول مدل تخصیص پنهان دیریکله در ارائه دسته‌های موضوعات حوزه غدد داشته است؛ به‌صورتی که دسته‌های موضوعی مستخرج از نظر ارتباط موضوعی در سطح بسیار بالایی بودند (۱۴۰۰).

پژوهش‌های انجام‌گرفته به‌طور عمده از نتایج مدل‌سازی موضوعی برای دسته‌بندی متون استفاده نمودند که یافته‌های آن‌ها نشان‌دهندهٔ موفقیت به‌کارگیری مدل‌سازی موضوعی در حوزه‌های مختلف است (Cheng, Cao & Liao 2022). این امر اهمیت تحلیل و پژوهش روی مدل‌سازی موضوعی را نشان می‌دهد (زرمهر، منصوری و کارشناس ۱۴۰۰). چالش بزرگی که در مدل‌سازی موضوعی وجود دارد، تخمین تعداد موضوعات در مدل موضوعی استفاده شده است که هدف این پژوهش بررسی و مقایسهٔ دو الگوریتم برای به‌دست آوردن تعداد موضوعات است.

۳. روش پژوهش

با توجه به اینکه هدف پژوهش، بررسی و تحلیل دو روش برای به‌دست آوردن تعداد موضوعات در مدل‌سازی موضوعی با توجه به داده‌های مقالات نشریات فارسی بود، شش نشریه به‌صورت تصادفی از نشریات «وزارت علوم» انتخاب شدند که عبارت‌اند از: «مطالعات مدیریت»، «مهندسی برق و مهندسی کامپیوتر ایران»، «روش‌های عددی در مهندسی»، «رهیافتی نو در مدیریت آموزشی»، «مکانیک هوافضا»، و «نشریه صفه». لازم به

ذکر است که اطلاعات مربوط به این نشریات در جدول ۱، نشان داده شده است.

از هر نشریه، ۲۰۰ مقاله که در سال‌های اخیر به چاپ رسیده‌اند، با نمونه‌گیری تصادفی انتخاب شده و اطلاعات کتابشناختی آن‌ها که شامل عنوان مقاله، چکیده و کلیدواژه است، به‌طور مستقیم از پایگاه داده «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری»[1] به کمک استراتژی جست‌وجو استخراج گردید. بنابراین در کل، عنوان، کلیدواژه، و چکیده ۱۲۰۰ مقاله مورد بررسی قرار گرفت.

جدول ۱. اطلاعات مربوط به نشریات مورد بررسی

نشریه	شاپا	ضریب تأثیر در ISC	ناشر	دوره انتشار	URL
مطالعات مدیریت	۲۵۳۸۳۲۱۳	۰/۱۸۴	پژوهشگاه تربیت بدنی و علوم ورزشی (وزارت علوم، تحقیقات و فناوری)	دوماهنامه	https://smrj.ssrc.ac.ir/
مهندسی برق و مهندسی کامپیوتر ایران	۱۶۸۲۳۷۴۵	۰/۰۵۳	پژوهشکده برق جهاد دانشگاهی	فصلنامه	http://ijece.saminatech.ir
روش‌های عددی در مهندسی	۲۲۲۸- ۷۶۹۸	۰/۰۲۹	دانشگاه صنعتی اصفهان	دوفصلنامه	http://jcme.iut.ac.ir/
رهیافتی نو در مدیریت آموزشی	۲۰۰۸-۶۳۶۹	۰/۲۷۳	دانشگاه آزاد اسلامی واحد مرودشت	دوماهنامه	http://jedu.marvdasht.iau.ir/
مکانیک هوافضا	۲۶۴۵-۵۳۲۳	۰/۰۹۰	دانشگاه جامع امام حسین (ع)	فصلنامه	https://maj.ihu.ac.ir/
صفه	۲۶۴۵-۵۹۰۰	۰/۲۵۵	دانشگاه شهید بهشتی	فصلنامه	https://soffeh.sbu.ac.ir/

از آنجا که داده‌های مورد بررسی در این پژوهش، مقالات نشریات فارسی است، ابتدا از هر مقاله اطلاعات کتابشناختی عنوان، کلیدواژه، و چکیده استخراج شد و عملیات پیش‌پردازش روی آن‌ها انجام گرفت. پس از تهیه این مجموعه، داده‌ها پردازش شده و کدگذاری کاراکترها از نظر کاراکترهای فارسی و عربی نرمال گردید. سپس، عملیات تبدیل متن به جمله و جمله به واژه انجام شد. برای این کار، ابتدا متن با توجه به

1. https://search.ricest.ac.ir/

کاراکترهای جداکننده[1] که شامل {، ؛ " ( ) : . < >} هستند، به مجموعه‌ای از توکن[2]ها تبدیل گردید. سپس، عملیات ریشه‌یابی[3] روی آن‌ها انجام گرفت. هر واژه با کد منحصربه‌فردی ذخیره گردید. افزون بر واژه، تعداد رخداد آن در مجموعه مقالات نیز محاسبه و ذخیره شد. همچنین، ایست‌واژه[4]ها، علائم و اعداد حذف گردیدند. خروجی این مرحله، واژگان پردازش‌شده‌ای بود که تعداد تکرار آن‌ها در هر مقاله نیز مشخص شده بود. لازم به ذکر است که تمامی این عملیات با زبان برنامه‌نویسی پایتون، نسخه ۳/۷ انجام گرفته است. کتابخانه‌های مورد استفاده شامل «پانداس»[5] که از آن برای تحلیل داده استفاده گردید؛ «نامپی»[6] برای کار با آرایه‌ها، و «گنسیم»[7] که مدل‌سازی موضوعی با کمک آن انجام گرفت، هستند.

در این پژوهش، دو روش یکی مبتنی بر معیار «گریدی» و دیگری مبتنی بر نظریه بازبهنجاری[8] برای به‌دست آوردن تعداد موضوعات مورد بررسی قرار گرفتند. روش جست‌وجوی «گریدی» که در اکثر پژوهش‌های پیشین برای تخمین تعداد موضوعات استفاده می‌شود، مقادیر مختلفی را برای پارامتر تعداد موضوعات اعمال کرده و کارایی مدل نهایی با توجه به آن پارامتر را بررسی کرده و سپس، پارامتری را که منجر به کارایی بیشتر می‌شود، به‌عنوان تخمین تعداد موضوعات در نظر می‌گیرد. بنابراین، این روش به یک معیار ارزیابی برای سنجهٔ میزان کارایی مدل نیاز دارد که در ادامه، در مورد معیار ارزیابی مدل موضوعی معرفی می‌شود.

در حالت کلی، ارزیابی یک مدل موضوعی بسیار مشکل است؛ چرا که مدل‌سازی موضوعی بدون نظارت انجام می‌گیرد. پژوهش‌های پیشین از تکنیک‌های مختلفی برای ارزیابی مدل موضوعی استفاده می‌کنند. بعضی از آن‌ها از قضاوت‌های انسانی برای ارزیابی مدل موضوعی استفاده می‌کنند (Chang et al. 2009). بعضی دیگر، از معیارهای عددی استفاده می‌کنند که نمونه آن‌ها معیار سرگشتگی[9] است (Blei et al. 2003). این معیار بدین‌صورت تعریف می‌شود:

$$P(D') = \sum_{w_d \in D'} P(w_d)\log(P(w_d)) \qquad (۲$$

1. delimitor characters
2. token
3. temming
4. stop words
5. Pandas
6. Numpy
7. Gensim
8. renormalization theory
9. perplexity

در این فرمول، D' مدل موضوعی است که قرار است کارایی آن سنجیده شود و w_d کلمات موجود در آن مجموعه است. یکی از اشکالاتی که به معیار سرگشتگی وارد است، این است که این معیار به قضاوت‌های انسان نزدیک نیست (Chang et al. 2009). به‌عبارت دیگر، احتمال وقوع کلمات و قضاوت‌های انسان در اغلب موارد مرتبط نیست و حتی در بعضی موارد ممکن است متضاد هم باشد.

انسجام معنایی[1]، معیار دیگری برای ارزیابی مدل موضوعی است که بهتر از سرگشتگی می‌تواند مدل‌های مختلف را با یکدیگر مقایسه نماید و شبیه‌ترین معیار به قضاوت انسانی است (Röder, Both & Hinneburg 2015). این معیار به‌صورت زیر تعریف می‌شود (ibid):

$$C_{UCI} = \frac{2}{W \times (W-1)} \sum_{i=2}^{W} \sum_{j=1}^{i-1} \log\left(\frac{P(w_i, w_j) + \varepsilon}{P(w_i).P(w_j)}\right) \quad (۳$$

$$P(w_i) = \frac{Number\ of\ documents\ that\ contain\ w_i}{total\ number\ of\ documents}$$

$$P(w_i, w_j) = \frac{Number\ of\ documents\ that\ contain\ w_i\ and\ w_j}{total\ number\ of\ documents}$$

در این فرمول، W تعداد کلمات و w_i و w_j به‌عنوان نمونه دو کلمه موجود در مجموعه را نشان می‌دهند. مقدار ε یک عدد بسیار کوچک در نظر گرفته می‌شود (ibid).

روش دیگری که در این پژوهش برای تخمین تعداد موضوعات موجود در نشریات مورد بررسی قرار گرفت، رویکردی مبتنی بر نظریه بازبهنجاری بود که در واقع، رویه ترکیب دو موضوع در یک موضوع است. به‌عنوان نتیجه ترکیب رویه ترکیب، ما یک موضوع جدید t با توزیع موضوع-کلمه به‌دست آوردیم. از آنجا که محاسبهٔ ماتریس Φ (همان ماتریس خروجی مدل موضوعی از کلمات و موضوعات) بستگی به مدل موضوعی مشخص دارد، فرمول‌بندی ریاضی رویّه بازبهنجاری نیز در واقع، وابسته به مدل است. افزون بر این، نتایج ترکیب بستگی دارد به اینکه موضوعات برای ترکیب چگونه انتخاب شوند.

در این پژوهش، از آنتروپی «رونو»[2] (Koltcov & Ignatenko 2020) برای رویّه انتخاب موضوع استفاده کردیم؛ بدین‌صورت که آنتروپی «رونو» برای هر موضوع به‌صورت

1. semantic coherence 2. Rényi

جداگانه بر اساس معادلهٔ زیر محاسبه می‌گردد.

$$H_{Renyi} = \ln(\tilde{P}) - T \times \ln(\rho) \quad (۴$$
$$\tilde{P} = E_{high} - T \times \ln(\rho)$$

در این فرمول، T تعداد موضوعات است و E_{high} کلمات با احتمال بالا هستند. در این پژوهش، مشابه با پژوهش‌های Koltcov & Ignatenko (2019, 2020) محدودهٔ احتمالی [0,1] را به دو بازه تقسیم کردیم. به‌عبارت دیگر، یک سیستم دو-سطحی در نظر گرفته شد که اولین سطح مربوط به کلمات با احتمال‌های بالاست و دومین سطح به کلمات با احتمالات کم نزدیک به صفر. در نتیجه، می‌توان تابع چگالی-حالات[1] را برای کلمات با احتمالات بالا تحت یک تعداد ثابت از موضوعات و پارامترها، طبق فرمول زیر معرفی نمود.

$$\rho = N/(WT) \quad (۵$$

در این معادله، NN تعداد کلمات با احتمال بالاست. منظور از احتمال بالا، احتمالاتی است که شرط $\rho > 1/W\rho > 1/W$ را ارضا کند. انتخاب چنین سطحی بر این اساس است که مقادیر $1/W1/W$ مقادیر اولیه ماتریس ϕ برای مدل‌های موضوعی در تخصیص پنهان دیریکله است. مقدار $W.TW.T$ تعداد کل حالات یک مدل موضوعی (سایز ماتریس ϕ) را مشخص کرده که برای نرمال‌سازی استفاده می‌شود. در طی فرایند مدل‌سازی موضوعی، احتمالات کلمات از توزیع $1/w1/w$ دور می‌شود. بخش کوچکی از کلمات، احتمالات بزرگ‌تر از مقدار آستانه می‌گیرند؛ در حالی که بخش بزرگ‌تر از کلمات، احتمالات کمتر از آن را می‌گیرند. بنابراین، E_{high} در معادلهٔ (۵) به‌صورت زیر تعریف می‌شود:

$$E_{high} = -\ln(\frac{1}{T}\sum_{w,t}(\phi_{wt}.\Omega(\phi_{wt} - {}^{1}/_{W}))) \quad (۶$$

در این معادله، ϕ_{wt} احتمال کلمهٔ w در موضوع t است و $\Omega(.)$ تابع پله[2] است که به‌صورت زیر تعریف می‌شود.

$$\Omega(\phi_{wt} - {}^{1}/_{W}) = 1 \quad if\ \phi_{wt} \geq {}^{1}/_{w} \quad (۷$$
$$\Omega(\phi_{wt} - {}^{1}/_{W}) = 0 \quad if\ \phi_{wt} < {}^{1}/_{w}$$

1. density-of-states function
2. step function

با استفاده از تابع پله، می‌توان کنترل نمود که در معادلهٔ (۶) فقط احتمالاتی که بزرگ‌تر از $1/W$ هستند، جمع می‌شوند.

بنابراین، این پژوهش برای به‌دست آوردن تعداد موضوعات از یک رویّهٔ تکرارشونده به‌شرح زیر استفاده کرد:

جفتی از موضوعات (t_1 و t_2) با کمترین آنتروپی «رونو» برای ترکیب انتخاب شدند. مقادیر بزرگ آنتروپی «رونو» نشان‌دهندهٔ این است که موضوعات با یکدیگر ارتباط کمی دارند؛ در حالی که مقادیر کمتر آنتروپی «رونو» نشان‌دهندهٔ بیشترین ارتباطات اطلاعاتی میان موضوعات است. بنابراین، موضوعاتی که ارتباط بیشتری با هم دارند، انتخاب شدند.

در این مرحله دو موضوع با کمترین آنتروپی با یکدیگر ترکیب شدند. فرض کنید t_1 و t_2 نشان‌دهنده دو موضوع با کمترین آنتروپی باشند؛ در این صورت ستون‌های مربوط به t_1 و t_2 در ماتریس Φ حذف شده و یک ستون جدید از ترکیب این دو موضوع با نام $\tilde{t}$ تولید گردید. احتمال کلمه در این موضوع جدید بر اساس معادلهٔ (۸) به‌دست می‌آید:

$$\phi_{w\tilde{t}} = \frac{c_{wt_1} + c_{wt_2} + \beta}{\left(\sum_{w \in \tilde{W}} c_{wt_1} + c_{wt_2}\right) + \beta W} \quad (۸$$

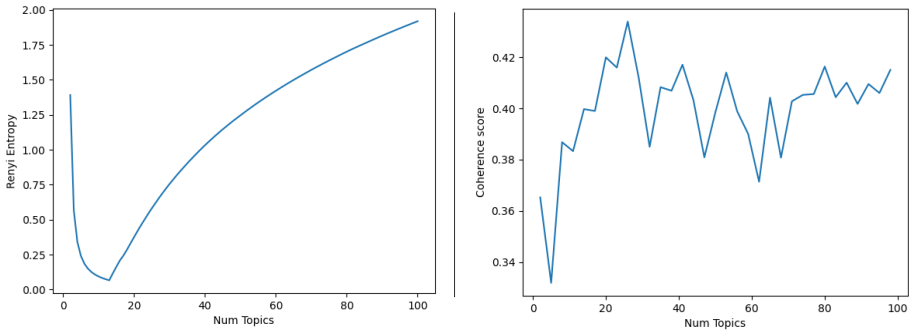
در این فرمول، c_{wt} تعداد دفعاتی است که کلمهٔ w به موضوع t نسبت داده می‌شود. مقدار β نیز ۰/۱ در نظر گرفته می‌شود (Koltcov & Ignatenko 2020).

ستون ϕ_{wt_1} به‌وسیلهٔ $\phi_{w\tilde{t}}$ و ستون ϕ_{wt_2} از ماتریس Φ حذف می‌شود. این مرحله باعث می‌شود که تعداد موضوعات در هر تکرار یکی کم شود. مرحله ۱ و ۲ به‌طور مرتب، تکرار می‌شود تا در نهایت، فقط دو موضوع باقی بماند. در هر بار پایان مرحله ۲، آنتروپی «رونو» برای ماتریس Φ بر اساس معادله (۴) محاسبه می‌گردد. سپس، نمودار آنتروپی «رونو» به‌عنوان یک تابع از تعداد موضوعات رسم شد و کمترین مقدار به‌منظور تشخیص تخمین تعداد بهینه از موضوعات محاسبه گردید.

٤. تجزیه‌وتحلیل یافته‌ها

در این بخش به ارائهٔ یافته‌ها و همچنین پاسخ به سؤالات پژوهشی می‌پردازیم. در این پژوهش دو سؤال پژوهشی مطرح شده است. یکی در مورد تعداد موضوعات موجود در نشریات مورد بررسی در این پژوهش، و دیگری مقایسهٔ زمانی دو روش ارائه‌شده در بخش قبلی است.

اولین نشریه‌ای که آن را بررسی می‌کنیم، نشریه «مطالعات مدیریت» در حوزه علوم انسانی است و هدف آن انتشار یافته‌های نوین پژوهشی در حیطه علوم ورزشی است. مقالات چاپ‌شده در سال‌های اخیر در این نشریه مورد بررسی قرار گرفت و روی آن، معیار انسجام به ازای تعداد موضوعات مختلف اعمال گردید که نتیجه در شکل ۲ (چپ) مشخص شده است. محور افقی تعداد موضوعات مختلف که از ۲ تا ۱۰۰ تغییر می‌کند و محور عمودی، معیار انسجام را نشان می‌دهد. با افزایش تعداد موضوعات، معیار انسجام از ۲ تا حدود ۲۲ افزایش پیدا کرده و به ازای مقادیر بزرگ‌تر، شروع به کاهش می‌کند. شکل ۲ (راست) معیار آنتروپی «رونو» را روی نشریهٔ «مطالعات مدیریت» نشان می‌دهد.

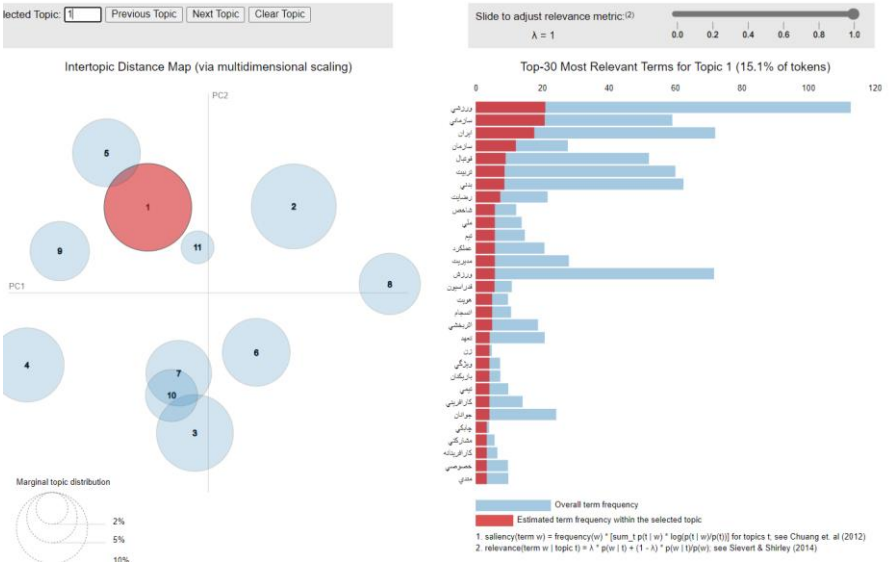


شکل ۲. راست: روش مبتنی بر نظریه بازبهنجاری روی نشریه «مطالعات مدیریت»؛ چپ: روش جست‌وجوی گریدی

به‌منظور ایجاد یک نمایه گرافیکی از نتایج مدل‌سازی موضوعی روی نشریات از روش LDAvis استفاده شد (Sievert & Shirley 2014). این نمایه گرافیکی از دو پنل تشکیل شده است که پنل سمت چپ، موضوعات و همچنین میزان اهمیت هر موضوع را نشان می‌دهد. هر موضوع با یک دایره مشخص شده است که اندازه هر دایره متناسب با میزان اهمیت آن موضوع است؛ به‌عبارت دیگر، هر چقدر شعاع یک دایره بیشتر باشد، موضوع متناسب با آن دایره از اهمیت بیشتری برخوردار است. با انتخاب هر موضوع در پنل سمت چپ، می‌توان کلمات موجود در آن موضوع را در پنل سمت راست مشاهده نمود. در کنار هر کلمه یک میله افقی وجود دارد که دو رنگ آبی و قرمز در آن میله وجود دارد. رنگ آبی نشان‌دهندهٔ تکرار کلمه در کل مقالات انتخاب‌شده در نشریه، و رنگ قرمز تکرار آن کلمه در موضوع انتخاب‌شده را مشخص می‌کند.

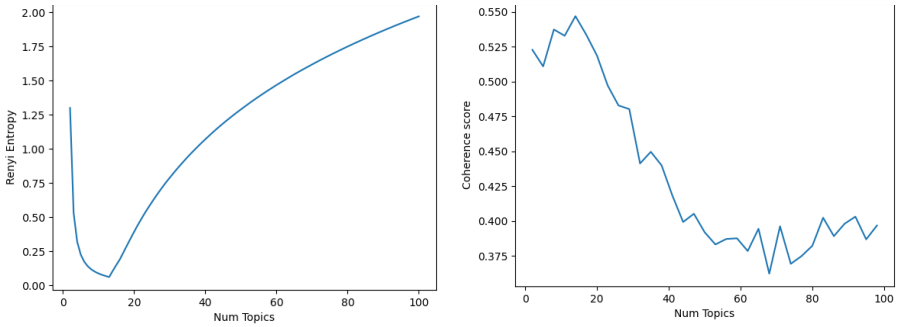
نمای گرافیکی از نمایش نتیجه مدل‌سازی موضوعی روی نشریه «مطالعات مدیریت»

در شکل ۳، مشخص شده است. پنل سمت چپ، موضوعات مختلف در این نشریه را نشان می‌دهد که اندازه هر دایره متناسب با میزان اهمیت هر موضوع است. همان‌طور که این شکل نشان می‌دهد، موضوع ۱ از اهمیت بالاتری برخوردار است. موضوعات ۳ و ۷ و ۱۰ با یکدیگر همپوشانی دارند. به‌عبارت دیگر، کلمات مشترکی میان این سه موضوع وجود دارد. موضوع ۱ و ۱۱ به‌ترتیب، پراهمیت‌ترین و کم‌اهمیت‌ترین موضوعات موجود در این نشریه هستند.



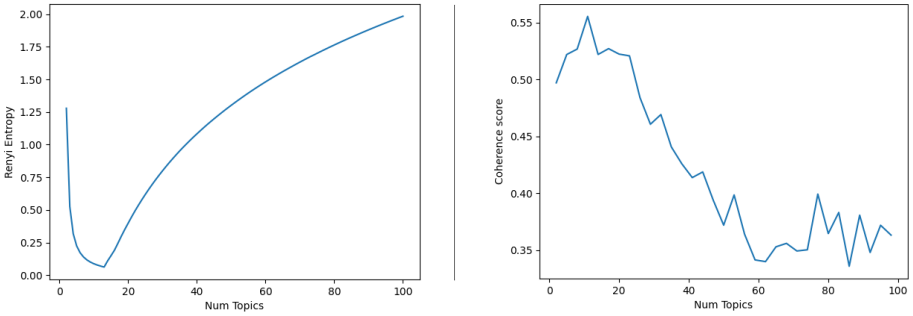
شکل ۳. نمای گرافیکی از موضوعات مختلف موجود در نشریه «مطالعات مدیریت»

«مهندسی برق و مهندسی کامپیوتر ایران»، نشریه دیگری در حوزه فنی و مهندسی است و به‌منظور به‌دست آوردن تعداد موضوعات در این نشریه، نتایج حاصل از «گریدی» با توجه به معیار انسجام و نظریه بازبهنجاری در شکل ۴، مشخص شده است.



شکل ٤. راست: روش مبتنی بر نظریه بازبهنجاری روی نشریه «مهندسی برق و کامپیوتر»؛ چپ: روش جست‌وجوی گریدی

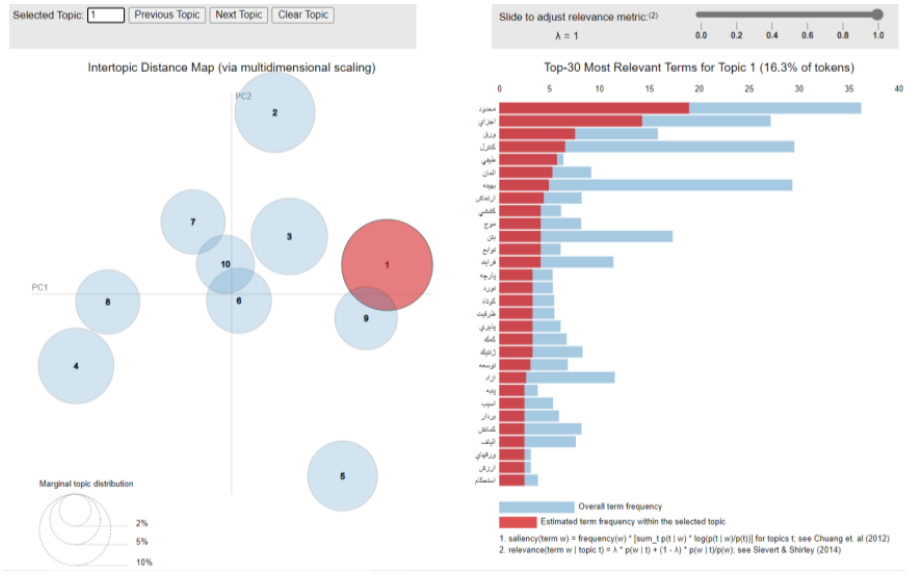
نشریه «روش‌های عددی در مهندسی» در حوزه فنی و مهندسی است که دستاوردهای پژوهشی محققان فارسی‌زبان را در زمینه‌های مختلف مهندسی به چاپ می‌رساند. به‌منظور پاسخ به پرسش اول پژوهش، از روش گریدی مبتنی بر انسجام (شکل ۵ (چپ)) و روش بازبهنجاری (شکل ۵ (راست)) استفاده شده است. نمودار در شکل ۵ (چپ) تا حدود تعداد ۱۲ موضوع به‌صورت صعودی و بعد از آن به‌صورت نزولی است. همین روند در شکل ۵ (راست) برعکس است؛ یعنی تا حدود ۱۲ موضوع نمودار نزولی و بعد از آن صعودی است. بالاترین معیار انسجام و کمترین معیار آنتروپی به معنای بهترین تعداد موضوع است.



شکل ۵. راست: روش مبتنی بر نظریه بازبهنجاری روی نشریه «روش‌های عددی در مهندسی»؛ چپ: روش جست‌وجوی گریدی

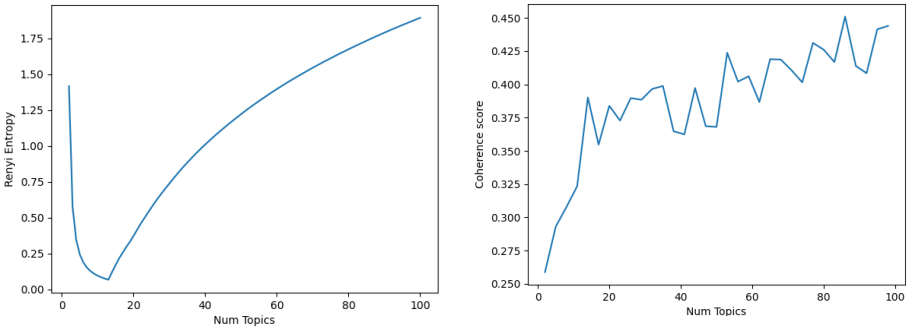
نمای گرافیکی از نتیجه مدل‌سازی موضوعی روی نشریه «روش‌های عددی در مهندسی» در شکل ۶، مشخص شده است. پنل سمت چپ، موضوعات و همچنین میزان اهمیت هر موضوع را نشان می‌دهد. به‌عنوان مثال، موضوع ۱ و ۲ بهترین موضوعات در

این نشریه هستند؛ چرا که نسبت به بقیه، ابعاد بزرگ‌تری دارند. با انتخاب هر موضوع در پنل سمت چپ، می‌توان کلمات موجود در آن موضوع را در پنل سمت راست مشاهده نمود. با توجه به این شکل، کلمات محدود، اجزای، ورق، کنترل، طیفی، المان و بهینه از مهم‌ترین کلمات در موضوع مشخص‌شده در پنل سمت چپ است.



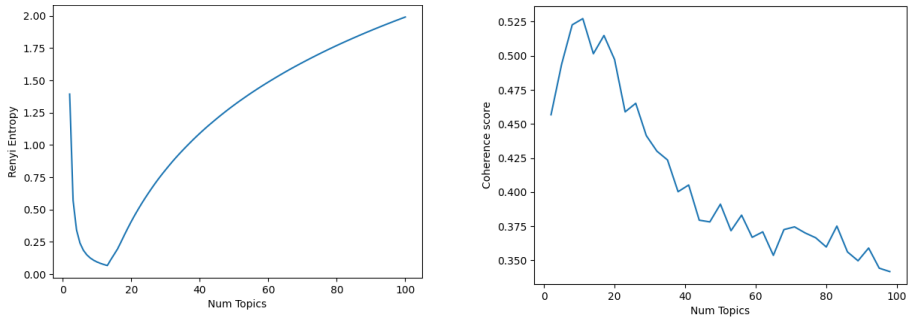
شکل ٦. نمای گرافیکی از موضوعات مختلف موجود در نشریه «روش‌های عددی در مهندسی»

یکی دیگر از نشریات معتبر در زمینهٔ علوم انسانی، نشریه «رهیافتی نو در مدیریت آموزشی» در حوزه مطالعات علوم تربیتی است که به چاپ مقالاتی در حیطه مسائل علوم تربیتی و آموزشی می‌پردازد. برای پاسخ به پرسش اول پژوهش برای این نشریه نیز از روش گریدی استفاده کردیم و نتایج آن در شکل ۷ (چپ) مشخص شده است؛ بدین‌صورت که به ازای تعداد موضوعات مختلف که می‌تواند از ۲ تا ۱۰۰ تغییر کند، معیار انسجام را به‌دست آورده و در نهایت، تعداد موضوع با بالاترین معیار انسجام را به‌عنوان یک تخمین مناسب از تعداد موضوعات و یا همان ابعاد مسئله در نظر می‌گیریم. در این شکل، معیار انسجام به‌صورت صعودی است و با افزایش تعداد موضوعات، معیار انسجام نیز افزایش پیدا می‌کند، به‌صورتی که در حدود ۸۵ موضوع بالاترین معیار انسجام به‌دست آمده است. به‌منظور بررسی بیشتر و به‌دست آوردن تعداد موضوعات، الگوریتم بازبهنجاری را نیز روی داده‌های این نشریه اعمال کردیم که نتایج آن در شکل ۷ (راست) نشان داده شده است.



شکل ۷. الف: روش جست‌وجوی گریدی؛
ب: روش مبتنی بر نظریه بازبهنجاری روی نشریه «رهیافتی نو در مدیریت آموزشی»

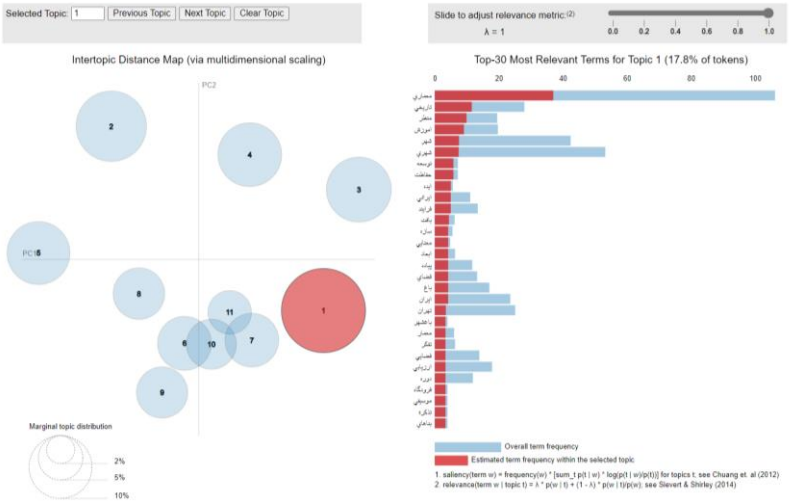
نشریه دیگری که در این پژوهش مورد بررسی قرار می‌گیرد، نشریه «صفه» در حوزه هنر و معماری است. برای این نشریه نیز پرسش‌های پژوهشی مورد بررسی قرار گرفتند. به‌منظور پاسخ به پرسش اول و یا همان به‌دست آوردن تعداد موضوعات یا ابعاد مسئله، از دو روش، یکی مبتنی بر گریدی و دیگری مبتنی بر الگوریتم بازبهنجاری استفاده شده است. نتایج آن در شکل ۸، نمایش داده شده است. در هر دو شکل، حدود ۱۱ موضوع بهترین نتیجه را به همراه دارد. شکل ۹، نمای گرافیکی از مدل‌سازی موضوعی روی این نشریه را نشان می‌دهد.



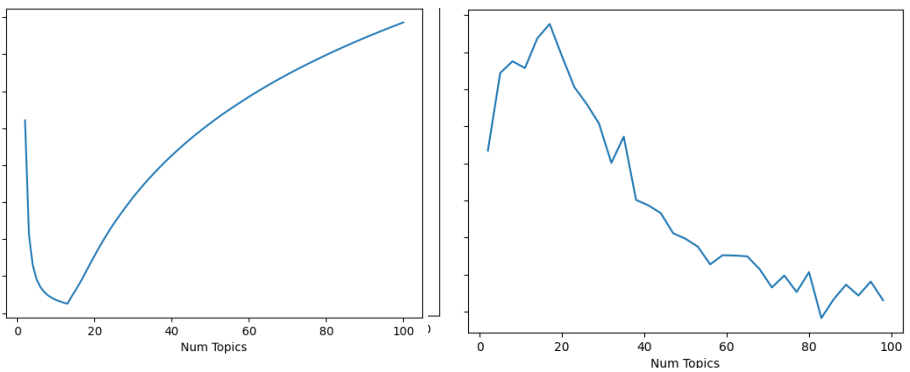
شکل ۸. راست: روش مبتنی بر نظریه بازبهنجاری روی نشریه «صفه»؛ چپ: روش جست‌وجوی گریدی

نشریه «مکانیک هوافضا» در حوزه فنی و مهندسی و به‌طور مشخص، مکانیک است. به‌منظور پاسخ به پرسش اول در این نشریه، از دو روش گریدی و بازبهنجاری استفاده کردیم. شکل ۱۰ (چپ) معیار انسجام را به ازای پارامترهای مختلف نشان می‌دهد و همان‌طور که این شکل نشان می‌دهد، این نشریه با تعداد موضوع حدود ۱۴، بهترین معیار انسجام را

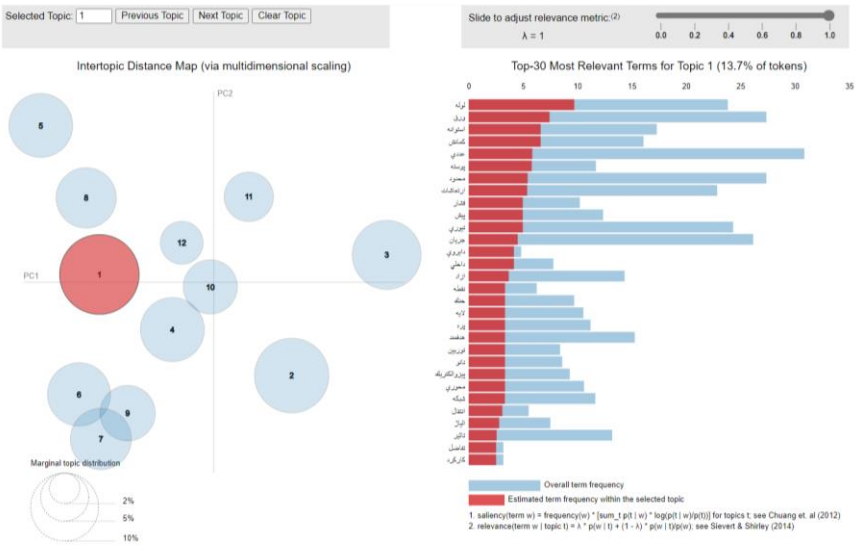
دربردارد. شکل ۱۰ (راست)، روند به‌دست آوردن تعداد موضوعات برای نشریه «مکانیک هوافضا» را با استفاده از الگوریتم بازبهنجاری نشان می‌دهد. همان‌طور که نمای گرافیکی از نشریه «مکانیک هوافضا» در شکل ۱۱، نشان می‌دهد، ۱۲ موضوع در این نشریه وجود دارد که این پارامتر نزدیکی بیشتری به روش تخمین پارامتر با روش بازبهنجاری دارد.



شکل ۹. نمای گرافیکی از موضوعات مختلف موجود در نشریه «صفه»



شکل ۱۰. راست: روش مبتنی بر نظریه بازبهنجاری روی نشریه «مکانیک هوافضا»؛ چپ: روش جست‌وجوی گریدی



شکل ۱۱. نمای گرافیکی از موضوعات مختلف موجود در نشریه «مکانیک هوافضا»

به‌منظور مقایسه زمانی میان دو روش گریدی مبتنی بر معیار انسجام و روش بازبهنجاری، مقایسه‌ای از مدت‌زمان اجرای دو روش به‌دست آوردن تعداد موضوعات و یا همان تعداد ابعاد مسئله در این پژوهش داشتیم (جدول ۲). همان‌طور که این جدول نشان می‌دهد، الگوریتم بازبهنجاری به‌صورت قابل ملاحظه‌ای نسبت به روش گریدی توانسته است تعداد موضوعات را با سرعت بالاتری بیابد.

جدول ۲. مقایسه زمان اجرا (ثانیه)؛ دو روش مختلف برای به‌دست آوردن تعداد موضوعات

عنوان نشریه	مدت‌زمان اجرا با استفاده از روش گریدی	مدت‌زمان اجرا با استفاده از بازبهنجاری و آنتروپی رونو
مطالعات مدیریت	۷۹۱,۴۹	۴۲,۸۶
مهندسی برق و مهندسی کامپیوتر ایران	۲۵۳۴,۹۴	۴۶,۶۴
روش‌های عددی در مهندسی	۸۶۰,۴۴	۴۴,۲۳
رهیافتی نو در مدیریت آموزشی	۱۵۴۶,۶۳	۴۲,۵۵
صفه	۶۹۴,۵۲	۴۶,۵۹
مکانیک هوافضا	۶۱۹,۰۱۷	۲۱,۲۵

۵. بحث و نتیجه‌گیری

در این پژوهش دو روش به‌دست آوردن تعداد موضوعات که یکی مبتنی بر «گریدی» و دیگری مبتنی بر نظریه بازبهنجاری است، بررسی گردید. بعضی از نشریات بررسی‌شده در این پژوهش، عملکرد یکسانی از هر دو روش «گریدی» و رویکرد مبتنی بر نظریه بازبهنجاری به همراه داشتند که از آن جمله می‌توان به نشریه «مهندسی برق و کامپیوتر ایران»، «روش‌های عددی در مهندسی»، و نشریه «صفه» نام برد. به‌عبارت دیگر، در این نشریات هر دو روش به‌کاررفته در این پژوهش به‌منظور تخمین تعداد موضوعات موجود در نشریات، نتیجه یکسانی را به همراه داشت. اگرچه، مدت‌زمان تخمین روش مبتنی بر نظریه بازبهنجاری در این نشریات نسبت به روش «گریدی» به طرز چشمگیری کمتر است. به‌عنوان مثال، مدت‌زمان اجرای الگوریتم بازبهنجاری در نشریه «مهندسی برق و مهندسی کامپیوتر ایران» ۴۶/۶۴ ثانیه بود؛ در حالی که اجرای روش «گریدی» مبتنی بر معیار انسجام روی همین نشریه ۲۵۳۴/۹۴ ثانیه زمان برد.

نتایج حاصل از دو روش تخمین تعداد موضوع در نشریه «مطالعات مدیریت» اختلاف چندانی نداشت و قابل چشم‌پوشی بود، ولی مدت‌زمان اجرای این دو روش در این نشریه اختلاف بسیار زیادی داشت. به‌عنوان مثال، مدت‌زمان اجرای الگوریتم با روش «گریدی» ۷۹۱/۴۹ و مدت‌زمان اجرا با استفاده از بازبهنجاری و آنتروپی «رونو» ۴۲/۸۶ بود؛ به‌عبارت دیگر، روش بازبهنجاری در حدود ۱۸ برابر سریع‌تر از روش مبتنی بر «گریدی» در این نشریه عمل کرد.

در نشریه «رهیافتی نو در مدیریت آموزشی»، عملکرد این دو روش در تخمین تعداد موضوعات موجود در نشریه، تفاوت بسیار زیادی داشت؛ به‌صورتی که روش «گریدی» در این نشریه ۸۵ موضوع را تشخیص داد و روش مبتنی بر نظریه بازبهنجاری ۱۲ موضوع را در این نشریه تخمین زد. تحلیل نتایج گرافیکی از این نشریه، مؤید درستی روش مبتنی بر نظریه بازبهنجاری بود.

همان‌طور که در قسمت‌های قبل توضیح داده شد، به‌دست آوردن تعداد موضوعات موجود در یک متن، یکی از بزرگ‌ترین چالش‌های موجود در مدل‌سازی موضوعی است. این پژوهش نشان داد که روش مبتنی بر نظریه بازبهنجاری می‌تواند با سرعت و دقت بسیار خوبی تعداد موضوعات موجود در یک متن را تشخیص دهد. بنابراین، روش مبتنی بر نظریه بازبهنجاری به‌عنوان یک روش دقیق و سریع به‌عنوان روش مؤثر برای یافتن تعداد

موضوعـات در مقـالات علمـی فارسـی پیشـنهاد می‌شـود. سـپس، بـا اسـتفاده از الگوریتم‌هـای تشـخیص موضوعـات ماننـد تخصیـص پنهـان دیریکلـه می‌تـوان موضوعـات موجـود در نشـریات را به‌صـورت توزیعـی از کلمـات موجـود در آن به‌دسـت آورد.

از کاربردهـای اجرایـی ایـن پژوهـش می‌تـوان بـه سـنجش میـزان تطابـق مقـالات چاپ‌شـده در نشـریات بـا اهـداف آن اشـاره کـرد. هـر نشـریه بـا توجـه بـه اهـداف و مأموریتـی کـه به‌طـور عمـده در قسـمت‌های «دربـاره نشـریه» یـا «اهـداف و چشـم‌انداز نشـریه» مشـخص می‌شـود، مقـالات مرتبـط را پذیـرش، بررسـی، و در نهایـت، چـاپ می‌کنـد. بـا اسـتفاده از مدل‌سـازی موضوعـی می‌تـوان بـه سـردبیران نشـریات کمـک کـرد تـا میـزان تطابـق مقـالات دریافتـی بـا اهـداف نشـریه در زمـان اولیـهٔ ارسـالِ مقالـه توسـط نویسـندگان را مشـخص نماینـد کـه بـرای یافتـن تعـداد موضوعـات در مقـالات نشـریات فارسـی، طبـق یافته‌هـای ایـن پژوهـش روش مبتنـی بـر نظریـه بازبهنجـاری پیشـنهاد می‌شـود.

از پژوهش‌هـای آتـی ایـن پژوهـش می‌تـوان بـه بررسـی اسـتفاده از روش‌هـای برنامه‌نویسـی پویـا به‌جـای نظریـه بازبهنجـاری اشـاره کـرد. همچنیـن، ارائـه یـک معیـار، به‌منظـور بررسـی میـزان تطابـق مقـالات چاپ‌شـده در هـر نشـریه بـا اهـداف آن به‌صـورت خـودکار از دیگـر تحقیقاتـی اسـت کـه در ادامـهٔ ایـن پژوهـش می‌توانـد انجـام شـود.

فهرست منابع

اسـدی قادیکلایـی، ام‌البنیـن، نجـلا حریـری، مریـم خادمـی، و فهیمـه باب‌الحوائجـی. ۱۴۰۰. مدل‌سـازی موضوعـی مقـالات پژوهشـگران ایـران در حـوزه غـدد درون‌ریـز و متابولیسـم در پایـگاه اسـتنادی وب علـوم. *پژوهشـنامه علم‌سـنجی*. صفحـه ۸، (۱۵) شـماره پیاپـی ۱۵: ۸-۴۹. DOI:10.22070/RSCI.2020.5813.1432.

دامـی، سـینا، و محمدرضـا الیکایـی. ۱۳۹۶. مدل‌سـازی موضوعـی رویدادهـای اخبـار مبتنـی بـر یادگیـری عمیـق افزایشـی. چهارمیـن کنفرانـس بین‌المللـی مطالعـات نویـن در علـوم کامپیوتـر و فنـاوری اطلاعـات. مشـهد.

دامـی، سـینا، و سـید احمـد طاهـرزاده. ۱۳۹۶. شناسـایی تهدیدهـای امنیتـی بـا اسـتفاده از مدل‌سـازی موضوعـی LDA و ماشـین بـردار پشـتیبان. کنفرانـس ملـی فناوری‌هـای نویـن در مهندسـی بـرق و کامپیوتـر. اصفهـان.

رحیمی، مرضیه، مرتضی زاهدی، و هدی مشایخی. ۱۳۹۷. یک مدل موضوعی احتمالاتی مبتنی بر روابط محلی واژگان در پنجره‌های همپوشان. *پردازش علائم و داده‌ها* ۴، پیاپی ۳۸: ۵۷-۷۰. DOI:10.29252/jsdp.15.4.57

زرمهر، فاطمه، علی منصوری، و حسین کارشناس. ۱۴۰۰. مدل‌سازی موضوعی و کاربرد آن در پژوهش‌ها؛ مروری بر ادبیات تخصصی. *پژوهشنامه کتابداری و اطلاع‌رسانی* ۱۱ (۱): ۲۳-۳۹. DOI:10.22067/infosci.2021.24128.0

زمانی، محسن، روح‌الله دیانت، و مهدی صادق‌زاده. ۱۳۹۳. دسته‌بندی متون فارسی با استفاده از روش آنالیز معنایی پنهان احتمالاتی. اولین همایش ملی کاربرد سیستم‌های هوشمند (محاسبات نرم) در علوم و صنایع. قوچان.

شکری، سعید، و بهروز معصومی. ۱۳۹۵. خوشه‌بندی معنایی متن با استفاده از تخصیص پنهان دیریکله و الگوریتم ژنتیک. چهارمین کنفرانس بین‌المللی پژوهش در علوم و تکنولوژی. ترکیه.

گیلوری، عباس. ۱۳۷۹. نمایه‌سازی خودکار (گذشته، حال، آینده). *تحقیقات اطلاع‌رسانی و کتابخانه‌های عمومی (پیام کتابخانه سابق)* ۳۹: ۱۷-۲۵.

هاشم‌زاده، محمدجواد، زینب نخعی، و حسین مرادی مقدم. ۱۳۹۲. کاربرد و تعدیل قانون زیف و الگوی بازو در بازشناسی واژه‌های بازدارنده زبان فارسی با استفاده از خوشه زبانی مقالات علمی-پژوهشی رشته کتابداری و اطلاع‌رسانی. *پژوهشنامه کتابداری و اطلاع‌رسانی* ۳ (۲): ۱۹۱-۲۰۸.

References

Barbieri, N., G. Manco, F. Ritacco, M. Carnuccio, & A. Bevacqua. 2013. Probabilistic topic models for sequence data. *Machine learning* 93 (1): 5-29. DOI:10.1007/s10994-013-5391-2.

Blei, D. M. 2012. Probabilistic topic models. *Communications of the ACM* 55 (4): 77-84. DOI:10.1145/2133806.2133826

Blei, D. M., A. Y. Ng, & M. J. Jordan. 200). Latent dirichlet allocation. *The Journal of machine Learning research* 3: 993-1022.

Chang, J., S. Gerrish, C. Wang, J. L. Boyd-Graber, & D. M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In Advances in neural information processing systems (pp. 288-296). Vancouver, British Columbia, Canada.

Cheng, X., Q. Cao, & S. Liao. 2022. An overview of literature on COVID-19, MERS and SARS: Using text mining and latent Dirichlet allocation. *Journal of Information Science* 48 (3): 304-320. DOI:10.1177/0165551520954674

Davarpanah, M. R., M. Sanji, & M. Aramideh. 2009. Farsi lexical analysis and stop word list. *Library Hi Tech*. DOI:10.1108/07378830910988559.

De Finetti, B. 2017. Theory of probability: A critical introductory treatment (Vol. 6). United Kingdom: John Wiley & Sons.

: John Wiley & Sons.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, & R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41 (6): 391-407. DOI:10.1002/(SICI)1097-4571

Dudoit, S., J. Fridlyand, & T. P. Speed. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association* 97 (457): 77-87. DOI:10.1198/016214502753479248.

Griffiths, T. L., M. Steyvers, & J. B. Tenenbaum. 2007. Topics in semantic representation. *Psychological review* 114 (2): 211. DOI:10.1037/0033-295X.114.2.211

Hofmann, T. 2013. Probabilistic latent semantic analysis. arXiv preprint arXiv:1301.6705. DOI:10.48550/arXiv.1301.6705.

Jameel, S., W. Lam, & L. Bing. 2015. Supervised topic models with word order structure for document classification and retrieval learning. *Information Retrieval Journal* 18 (4): 283-330. DOI:10.1007/s10791-015-9254-2.

Kadanoff, L. P. 2000. Statistical physics: statics, dynamics and renormalization. *World Scientific Publishing Company* DOI:10.1142/4016.

Kherwa, P., & P. Bansal. 2017. Latent Semantic Analysis: An Approach to Understand Semantic of Text. In 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC) (pp. 870-874). IEEE. DOI:10.1109/CTCEEC.2017.8455018.

Kherwa, P., & P. Bansal. 2020. Topic Modeling: A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems* 7 (24). DOI:10.4108/eai.13-7-2018.159623.

Koltcov, S. N. 2017. A thermodynamic approach to selecting a number of clusters based on topic modeling. *Technical Physics Letters* 43 (6): 584-586. DOI:10.1134/S1063785017060207.

_____, & V. Ignatenko. 2020. Renormalization approach to the task of determining the number of topics in topic modeling. In Science and Information Conference (pp. 234-247). Springer, Cham. DOI:10.1007/978-3-030-52249-0_16.

_____, & O. Koltsova. 2019. Estimating Topic Modeling Performance with Sharma–Mittal Entropy. *Entropy* 21 (7): 660. DOI:10.3390/e21070660.

Lee, D. D., & H. S. Seung. 2001. Algorithms for non-negative matrix factorization. In Advances in neural information processing systems (pp. 556-562).

Noji, H., D. Mochihashi, & Y. Miyao. 2013. Improvements to the Bayesian topic n-gram models. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1180-1190). Washington, USA.

Röder, M., A. Both, & A. Hinneburg. 2015. Exploring the space of topic coherence measures. In Proceedings of the eighth ACM international conference on Web search and data mining (pp. 399-408). Shanghai, China.

Sadeghi, M., & J. Vegas. 2014. Automatic identification of light stop words for Persian information retrieval systems. *Journal of information science* 40 (4): 476-487.

Sato, I., & H. Nakagawa. 2010. Topic models with power-law using Pitman-Yor process. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 673-682). Washington, DC, USA.

Sievert, C., & K. Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70). Baltimore, Maryland, USA.

Wang, X., A. McCallum, & X. Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In Seventh IEEE international conference on data mining (ICDM 2007) (pp. 697-702). IEEE. Omaha, Nebraska, USA.

Wang, C., & D. M. Blei. 2011. Collaborative topic modeling for recommending scientific articles. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 448-456). San Diego, California USA.

Yang, G., D. Wen, N. S. Chen, & E. Sutinen. 2015. A novel contextual topic model for multi-document summarization. *Expert Systems with Applications* 42 (3): 1340-1352.

نیلوفر مظفری

متولد ۱۳۶۴ دارای مدرک دکتری در رشته هوش مصنوعی از دانشگاه شیراز است.
ایشان هم‌اکنون استادیار مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری است.
داده‌کاوی، یادگیری ماشین، تحلیل شبکه‌های اجتماعی و پردازش زبان‌های طبیعی از جمله علایق پژوهشی وی است.

پژوهش‌نامه

پردازش و مدیریت اطلاعات