Open Access

Digital and Smart Libraries Researches

Summer (2024) 11(3): 47-60

DOI: https://doi.org/10.30473/mrs.2024.11582

Received: 21/Nov/2024 Accepted: 02/Feb/2025

ORIGINAL ARTICLE

Recognizing Semantic Patterns and Extracting Top Topics for the Thesaurus of Information Science and Epistemology by Relying on Advanced Text Processing and Content Analysis Techniques

Mohsen HajiZeinolabedini[1*], Hamid Keshavarz[2], Mahnam Zamani Kalajahi[3]

1. Assistant Professor, Department of Knowledge and Information Science, Shahid Beheshti University, Tehran, Iran.
2. Assistant Professor, Department of Knowledge and Information Science, Shahid Beheshti University, Tehran, Iran.
3. Msc Student, Department of Knowledge and information Science, Shahid Beheshti University, Tehran, Iran.

Correspondence
Mohsen HajiZeinolabedini
Email: zabedini@gmail.com

How to cite
HajiZeinolabedini, M., Keshavarz, H., & Zamani Kalajahi, M. (2024). Recognizing Semantic Patterns and Extracting Top Topics for the Thesaurus of Information Science and Epistemology … . *Digital and Smart Libraries Researches*, 11(3), 47-60.

ABSTRACT

In the age of information explosion, the field of information science and knowledge seeks to simplify and improve the thesaurus production process. This goal is realized by using text mining techniques and machine learning algorithms. The proposed approach includes automatically extracting topics from unstructured text data and identifying key concepts in the field of information science and knowledge. The main goal of this research is to improve and develop the thesaurus by focusing on text mining techniques. This approach effectively facilitates information retrieval and simplifies the thesaurus generation process. This study includes several main steps. First, abstracts of articles related to the field of information science and knowledge were collected from the Web of Science citation database in the period of 1968-2022. Data were preprocessed in Python to remove unnecessary characters and symbols. Then, TextRank algorithm was applied using Pandas and NLTK libraries to discover hidden topics in texts. This iterative process led to the identification of top topics in the subject area. Finally, by analyzing and comparing the existing manual thesaurus and examining the criteria of subject coherence and thematic coverage, the effectiveness of the proposed approach was evaluated and the top terms were selected. This method effectively used big data to extract key topics in the field of information science and knowledge. This study has extracted key topics and selected top topics using text mining techniques and TextRank algorithm. The results indicate the identification of 17 main issues in the field of information science and knowledge. These topics include important areas such as archives and information centers, artificial intelligence, bibliography, classification, collection development, controlled vocabulary, digital libraries, information organization, information retrieval and data extraction, information science and librarianship, information systems and resources, knowledge management, Libraries and community services are metadata, reference services, subject headings, and scientology. This list of top topics effectively represents key concepts in the field of information science and knowledge and can be used as a basis for developing a thesaurus and improving the information retrieval process .Using text mining methods and advanced algorithms, this research extracted and proposed key topics for the term Ras through detailed analysis of textual sources.

KEYWORDS

Thesaurus, Information Science and Knowledge, Text Mining, Top Topics.

نشریه تخصصی
پژوهش‌های کتابخانه‌های دیجیتالی و هوشمند

سال یازدهم، شماره سوم، پیاپی چهل و دوم، پاییز ۱۴۰۳ (۶۰-۴۷)

تاریخ دریافت: ۱۴۰۳/۰۹/۰۱ تاریخ پذیرش: ۱۴۰۳/۱۱/۱۰

DOI: https://doi.org/10.30473/mrs.2024.11582

«مقاله - پژوهشی»

تشخیص الگوهای معنایی و استخراج موضوعات رأس برای اصطلاحنامه علم اطلاعات و دانش‌شناسی با تکیه‌بر تکنیک‌های پیشرفته پردازش متن و تحلیل محتوا

محسن حاجی‌زین‌العابدینی[۱*]، حمید کشاورز[۲]، مهنام زمانی کلجاهی[۳]

[۱]. استایار، گروه علم اطلاعات و دانش‌شناسی، دانشگاه شهید بهشتی، تهران، ایران.
[۲]. استادیار، گروه علم اطلاعات و دانش‌شناسی دانشگاه شهید بهشتی، تهران، ایران.
[۳]. دانشجوی کارشناسی ارشد، گروه علم اطلاعات و دانش‌شناسی دانشگاه شهید بهشتی، تهران، ایران.

نویسنده مسئول: محسن حاجی‌زین‌العابدینی
رایانامه: zabedini@gmail.om

چکیده

در عصر انفجار اطلاعات، حوزه علم اطلاعات و دانش‌شناسی به دنبال ساده‌سازی و ارتقای فرآیند تولید اصطلاحنامه است. این هدف با استفاده از تکنیک‌های متن‌کاوی و الگوریتم‌های یادگیری ماشین تحقق می‌یابد. رویکرد پیشنهادی شامل استخراج خودکار موضوعات از داده‌های متنی بدون ساختار و شناسایی مفاهیم کلیدی در حوزه علم اطلاعات و دانش‌شناسی است. هدف اصلی این پژوهش، بهبود و توسعه اصطلاحنامه با تمرکز بر تکنیک‌های متن‌کاوی است. این رویکرد به‌طور مؤثری بازیابی اطلاعات را تسهیل می‌کند و فرآیند تولید اصطلاحنامه را ساده‌سازی می‌کند. روش‌شناسی پژوهش شامل چند مرحله اصلی است. ابتدا، چکیده‌های مقالات مرتبط با حوزه علم اطلاعات و دانش‌شناسی از پایگاه استنادی Web of Science در بازه زمانی ۲۰۲۲-۱۹۶۸ جمع‌آوری شدند. داده‌ها در پایتون پیش‌پردازش شدند تا از نویسه‌ها و نمادهای غیرضروری پاک‌سازی شوند. سپس، الگوریتم TextRank با استفاده از کتابخانه‌های Pandas و NLTK برای کشف موضوعات پنهان در متن‌ها اعمال شد. این فرآیند تکراری به شناسایی موضوعات رأس در حوزه موضوعی منجر شد. در نهایت، با تحلیل و مقایسه اصطلاحنامه دستی موجود و بررسی معیارهای انسجام موضوع و پوشش موضوعی، اثربخشی رویکرد پیشنهادی ارزیابی و اصطلاحات رأس انتخاب شدند. این روش به‌طور مؤثری از داده‌های بزرگ برای استخراج موضوعات کلیدی در حوزه علم اطلاعات و دانش‌شناسی استفاده کرد. یافته‌های پژوهش بیان می‌کند که این مطالعه با استفاده از تکنیک‌های متن‌کاوی و الگوریتم TextRank، به استخراج موضوعات کلیدی و انتخاب موضوعات رأس پرداخته است. نتایج نشان‌دهنده شناسایی ۱۷ موضوع اصلی در حوزه علم اطلاعات و دانش‌شناسی است. این موضوعات شامل حوزه‌های مهمی مانند آرشیوها و مراکز اطلاعاتی، هوش مصنوعی، کتابشناختی، رده‌بندی، توسعه مجموعه، واژگان کنترل شده، کتابخانه‌های دیجیتال، سازمان‌دهی اطلاعات، بازیابی اطلاعات و استخراج داده‌ها، علم اطلاعات و کتابداری، نظام‌های اطلاعات و منابع، مدیریت دانش، کتابخانه‌ها و خدمات اجتماعی، فراداده، خدمات مرجع، سرعنوان‌های موضوعی و علم‌سنجی هستند. این فهرست موضوعات رأس به‌طور مؤثری نماینده مفاهیم کلیدی در حوزه علم اطلاعات و دانش‌شناسی است و می‌تواند به‌عنوان پایه‌ای برای توسعه اصطلاحنامه و بهبود فرآیند بازیابی اطلاعات استفاده شود. این پژوهش با بهره‌گیری از روش‌های متن‌کاوی و الگوریتم‌های پیشرفته، به استخراج و پیشنهاد موضوعات کلیدی برای اصطلاح رأس از طریق تجزیه و تحلیل دقیق منابع متنی، پرداخت.

واژه‌های کلیدی

اصطلاحنامه، علم اطلاعات و دانش‌شناسی، متن‌کاوی، موضوعات رأس.

استناد به این مقاله:

حاجی‌زین‌العابدینی، محسن؛ کشاورز، حمید و زمانی کلجاهی، مهنام (۱۴۰۳). تشخیص الگوهای معنایی و استخراج موضوعات رأس برای اصطلاحنامه علم اطلاعات و دانش‌شناسی با تکیه‌بر تکنیک‌های پیشرفته پردازش متن و تحلیل محتوا. *پژوهش‌های کتابخانه‌های دیجیتالی و هوشمند*، ۱۱(۳)، ۶۰-۴۷.

مقدمه

اصطلاحنامه یکی از بهترین ابزارهای نمایه‌سازی و اطلاع‌رسانی است. این ابزار مجموعه‌ای از واژگان است که روابط مترادف، وابستگی و سلسله مراتبی بین آن‌ها برقرار کرده است. اصطلاحنامه‌ها کمک می‌کنند تا موضوع یک‌رشته را با تمام جنبه‌های اصلی و فرعی وابسته به نظم و ترتیب خاصی ارائه دهند. آن‌ها بین استفاده‌کنندگان و سازمان‌دهی‌کنندگان اطلاعات هماهنگی ایجاد می‌کنند و با ارائه اصطلاحات مشخص، روابط مفهومی آشکار، یادداشت‌های دامنه و یادداشت‌های تاریخی، یکدستی در فرایند ذخیره و بازیابی اطلاعات را فراهم می‌کنند.

در عصر دیجیتال، مدیریت و تحلیل اطلاعات علمی چالش‌برانگیز است. علم اطلاعات و دانش‌شناسی نقش کلیدی در سازمان‌دهی و بازیابی اطلاعات دارد. اصطلاح‌نامه‌ها به‌عنوان چارچوب مفهومی برای تعریف و سازمان‌دهی اصطلاحات عمل می‌کنند، اما استخراج موضوعات رأس برای آن‌ها چالش‌برانگیز است. در حالی که توجه به اصطلاحنامه‌نویسی خودکار کم‌تر است، اما برخی سازمان‌ها به تدوین اصطلاحنامه‌های سنتی می‌پردازند. پیاده‌سازی فناوری‌های تولید خودکار اصطلاحنامه‌ها هنوز فراگیر نیست و فرآیند تولید خودکار، روابط از طریق روش‌های آماری شناسایی می‌شوند (حسینی بهشتی، ۱۳۸۲)، در حالی که تمرکز بر تحلیل معناشناختی وجود ندارد. برای تولید اصطلاحنامه از الگوهای فاصله‌برداری و الگوریتم‌های دسته‌بندی استفاده می‌شود. این وضعیت نشان‌دهنده نیاز به تحول در روش‌های تولید اصطلاحنامه‌ها و تقویت نقش اصطلاحنامه‌نویسی خودکار در مدیریت اطلاعات علمی است.

فرآیند تولید اصطلاحنامه با استفاده از متن‌کاوی، یک روش پیشرفته برای استخراج و مدیریت اطلاعات از متون علمی است. این روش با استفاده از تکنیک‌های متنوعی مانند طبقه‌بندی، خوشه‌بندی و خلاصه‌سازی خودکار متون، قادر به شناسایی، استخراج و مدیریت اطلاعات نهفته در متون علمی است. پس از بررسی متون با استفاده از استخراج واژه‌های کلیدی، می‌توان به اطلاعات موجود در متن پی برد. تکنیک‌های مختلفی مانند ساختار مفهومی، کاوش قوانین انجمنی، درخت تصمیم‌گیری و روش‌های استنتاج قوانین برای استخراج اطلاعات از متون استفاده می‌شوند. متن‌کاوی به‌عنوان یک حوزه در حال رشد، هدف اصلی آن استخراج اطلاعات معنادار از متون طبیعی است. این فرایند هوشمندانه، با تبدیل داده‌های متنی غیرساخت‌یافته به اطلاعات معنادار، امکان دسترسی به دانش و اطلاعات مفیدی از حجم عظیمی از داده‌های متنی را فراهم می‌کند. متن‌کاوی ارتباط نزدیک با پردازش زبان طبیعی، آمار، یادگیری ماشینی، استدلال استخراج اطلاعات، مدیریت دانش و سایر رشته‌های مرتبط دارد. همچنین متن‌کاوی به‌عنوان ابزاری قدرتمند در علم اطلاعات و دانش‌شناسی، امکان کشف و استخراج دانش پنهان در متون علمی را فراهم می‌کند (پونز پوراتا و همکاران[1]، ۲۰۰۷) این روش به بهبود فرآیندهای تصمیم‌گیری، تحقیقات علمی و توسعه دانش در حوزه‌های مختلف منجر می‌شود.

با توجه به اهمیت خودکارسازی تولید موضوعات اصطلاحنامه، پژوهشگران به استخراج و پیشنهاد اصطلاحات رأس با استفاده از متن‌کاوی پرداخته‌اند. این مطالعه با مرور مقالات حوزه علم اطلاعات و دانش‌شناسی، به مفاهیم اصلی و اصطلاحات رأس دست یافته و اطلاعات لازم را برای نمایه‌سازان و دانشجویان در نظر گرفته است. تمرکز اصلی این پژوهش بر خودکارسازی کشف موضوعات اصطلاحنامه است، که به‌عنوان یک جزء کلیدی در فرآیند ذخیره و بازیابی اطلاعات تلقی می‌شود. این مطالعه به دلیل وجود اصطلاحنامه‌های تک‌زبانه و چندزبانه، به‌ویژه اصطلاحنامه تک‌زبانه انگلیسی، اهمیت دارد.

اصطلاح‌نامه‌ها ساختارهای پیچیده‌ای هستند که می‌توانند براساس نیازها و اهداف خاص طراحی شوند. اصطلاحات رأس، نقاط مرکزی در ساختار اصطلاحنامه، اما همیشه در همه آن‌ها وجود ندارند. در حوزه علم اطلاعات و دانش‌شناسی، استفاده از متن‌کاوی برای انتخاب موضوعات رأس اصطلاحنامه چالش‌برانگیز است. این حوزه نیاز به الگوریتم‌های مناسب برای شناسایی موضوعات کلیدی در داده‌های متنی بزرگ دارد. پژوهش حاضر با هدف توسعه رویکردی جدید برای پیشنهاد موضوعات رأس اصطلاحنامه با استفاده از متن‌کاوی و الگوریتم TextRank انجام شده است. این رویکرد نه‌تنها به بهبود دقت و کارایی فرآیند پیشنهاد موضوعات می‌انجامد، بلکه به ارتقاء کیفیت پژوهش‌ها در علم اطلاعات و دانش‌شناسی کمک می‌کند. برای انجام این پژوهش، از نرم‌افزار پایتون برای شناسایی واژگان کلیدی در حوزه علم اطلاعات و دانش‌شناسی استفاده شد. سپس الگوریتم موردنظر بر روی مقالات اعمال شد. این رویکرد می‌تواند سرعت و دقت در پردازش و تحلیل داده‌های متنی را در محیط‌های علمی و پژوهشی افزایش دهد.

1. Pons-Porrata et al

اهداف پژوهش

هدف کلی

از مهم‌ترین اهداف این پژوهش بهبود ذخیره و بازیابی اطلاعات و ارائه الگوهای مناسب برای استخراج و پیشنهاد موضوعات رأس برای اصطلاحنامه علم اطلاعات و دانش‌شناسی از طریق متن‌کاوی مقالات این حوزه است. این فرایند از دو بخش مهم نمایه‌سازی و کاوش تشکیل می‌شود.

هدف اصلی این پژوهش متن‌کاوی مقالات حوزه علم اطلاعات دانش‌شناسی به‌منظور استخراج و پیشنهاد موضوعات رأس برای اصطلاحنامه علم اطلاعات و دانش‌شناسی و ارائه الگوریتم‌های مناسب جهت استخراج مفاهیم و موضوعات اصلی جهت نمایه‌سازی است.

اهداف فرعی

از اهداف فرعی این پژوهش نیز می‌توان موارد زیر را نام برد:

۱- تحلیل و بررسی موضوعات موجود در اصطلاحنامه علم اطلاعات و دانش‌شناسی.

۲- استخراج اطلاعات مرتبط با موضوعات رأس از متون علمی این حوزه.

۳- تحلیل و بررسی تکنیک‌های متن‌کاوی برای استخراج اطلاعات.

پرسش‌های پژوهش

پرسش اصلی

با استفاده از الگوهای به دست آمده چه واژه‌های رأسی برای اصطلاحنامه علم اطلاعات و دانش‌شناسی می‌توان پیشنهاد داد؟

پرسش‌های فرعی

۱ .چه تکنیک‌های متن‌کاوی برای استخراج اطلاعات موضوعات رأس از متن‌های علمی مورد استفاده قرار می‌گیرد؟

۲ .چه چالش‌هایی در استخراج اطلاعات موضوعات رأس از متن‌های علمی وجود دارد؟

مبانی نظری

تعریف اصطلاحنامه

اصطلاحنامه، یک ساختار پیچیده و پویای اطلاعاتی است که بر اساس راهنمای یونسکو در سال ۱۹۷۰ تعریف شده است. این ساختار، مجموعه‌ای از اصطلاحات کنترل شده است که به شیوه معنایی به یکدیگر پیوسته‌اند و یک زمینه خاص از دانش را پوشش می‌دهند. (جسوس هولاندا و همکاران[1]، ۲۰۰۴) اصطلاحنامه می‌تواند ساختار الفبایی یا نظام‌مند داشته باشد، شامل توصیفگرها و غیرتوصیفگرها، و روابط بین آن‌ها را تعیین کند. از دیدگاه کنت، اصطلاحنامه باید به‌عنوان یک مجموعه از اصطلاحات یک نظام ذخیره و بازیابی در نظر گرفته شود که در یک شکل منجمن و با معنا در کنار هم قرار گرفته‌اند.

اصطلاح‌نامه‌ها در نظام درونداد و نظام برونداد، هر دو، موقعیت‌هایی مؤثر دارند. این مفهوم، معادل فارسی واژه لاتین "Thesaurus"است که ریشه یونانی دارد و به معنای مخزن و گنجینه است. اصطلاح‌نامه‌ها با توجه به نوع کارکرد و هدف، تعاریف مختلفی دارند. از دیدگاه یونسکو، اصطلاحنامه مجموعه‌ای از واژگان کنترل شده و پویاست که برای برگرداندن زبان طبیعی مدارک به زبان مشترک بین نمایه‌ساز و کاربر است.

از دیدگاه استاندارد ملی آمریکا، اصطلاحنامه واژگان کنترل شده‌ای است که براساس الگویی مشخص سازمان‌دهی شده تا روابط بین اصطلاحات را به صورت واضح و روشن به نمایش بگذارد. از دیدگاه استاندارد بین‌المللی ISO 2788، اصطلاحنامه مجموعه‌ای از واژگان کنترل شده زبان نمایه‌ای است که روابط از پیش تعیین شده بین مفاهیم را نشان می‌دهد. اصطلاحنامه نه‌تنها روابط سلسله‌مراتبی، بلکه روابط وابستگی اصطلاحات، یادداشت‌های لازم برای هر اصطلاح و ارجاعات از اصطلاحات نامرتبط را به نمایش می‌گذارد. (ایچسن و دکستر کلارک[2]، ۲۰۰۴) این ساختار، نوعی طبقه‌شناسی است که از طریق زبان کنترل شده به‌عنوان یک ابزار شناخته و تثبیت شده برای بازیابی اطلاعات به نمایه‌سازی و بازیابی اطلاعات کمک می‌کند. همچنین، ابزاری برای مدیریت دانش و تبادل اطلاعات است که می‌توان برای طبقه‌بندی خودکار اسناد، نمایه‌سازی خودکار و گسترش جستجو استفاده کرد. در مجموع، اصطلاحنامه، یک ابزار طراحی شده برای کنترل واژگان نظام‌های ذخیره و بازیابی اطلاعات است که به افزایش حداکثری انسجام در توصیف مفاهیم یک مجموعه از اصطلاحات منتخب با روابط مناسب بین آن‌ها کمک می‌کند.

پس از تعیین حوزه اصطلاحنامه، انتخاب اصطلاحات برای ایجاد این ساختار یک مرحله کلیدی است. این انتخاب بر اساس معیارهای استاندارد تدوین اصطلاحنامه یا استراتژی‌های گروه تدوین‌کننده انجام می‌شود. انتخاب اصطلاحات و ایجاد روابط سلسله‌مراتبی بین آن‌ها می‌تواند به‌عنوان اولین مرحله

1. Jesus Holanda et al.
2. Aitchison and Dextre Clarke

تدوین اصطلاحنامه در نظر گرفته شود. در این فرآیند، اصطلاح رأس، که بالاترین سطح در سلسله‌مراتب واژه‌های کنترل شده است، نقشی اساسی ایفا می‌کند. این اصطلاح، جامع‌ترین اصطلاح ممکن در حوزه موضوعی است و معمولاً در اصطلاحنامه‌های چندرشته‌ای کاربرد دارد. اصطلاح رأس، حوزه موضوعی را تعیین و مشخص می‌کند، و در برخی سیستم‌ها مانند یونسکو، از MT (Micro Term) به‌جای TT استفاده می‌شود. در اصطلاحنامه‌های فارسی، اصطلاح رأس معمولاً با علامت اختصاری «ار» نشان داده می‌شود. این مرحله، پایه و اساس ساختار سلسله‌مراتبی اصطلاحنامه را شکل می‌دهد و برای ایجاد یک ساختار منطقی و قابل‌اعتماد در اصطلاحنامه ضروری است.

متن‌کاوی

متن‌کاوی، یک فناوری نوظهور است که هدف آن استخراج اطلاعات معنی‌دار از داده‌های متنی ساختارنیافته است. این روش، بر اساس پردازش پیچیده زبان طبیعی، به شناسایی الگوها، روندها و مدل‌های مفید در داده‌های متنی می‌پردازد. متن‌کاوی برای استخراج، مدیریت و بهره‌برداری از دانش در متون مختلف مانند فایل‌های متنی، مقالات و پیام‌های الکترونیکی استفاده می‌شود. در مقایسه با داده‌کاوی، متن‌کاوی حوزه‌ای میان‌رشته‌ای است که از بازیابی اطلاعات، داده‌کاوی، یادگیری ماشینی، آمار و زبان‌شناسی محاسباتی نشانه گرفته است. این روش، ارزش اقتصادی بالایی دارد زیرا بسیاری از اطلاعات در شکل متنی ذخیره شده‌اند.

متن‌کاوی با استفاده از پردازش زبان طبیعی، امکان استخراج دانش از متون ساختارنیافته را فراهم می‌کند و نیاز به محدودیت در منابع اطلاعاتی را برمی‌انگیزد. با تکیه‌بر پردازش زبان طبیعی، متن‌کاوی می‌تواند اطلاعات ارزشمند را از متون زبان طبیعی استخراج کند و به‌طور خودکار دانش را از این منابع استخراج کند. این روش، به‌طور قابل‌توجهی کاربردپذیری کشف دانش از داده‌ها را افزایش داده است و فرصت‌های جدیدی برای استخراج و پردازش اطلاعات از متون ساختارنیافته فراهم کرده است.

پیشینه پژوهش

تیمورپور (۱۳۸۸)، در رساله دکتری خود با عنوان «کشف روندها نوظهور در حوزه‌های علمی بر پایه خوشه‌بندی پویا با رویکرد متن‌کاوی و تحلیل پیوند»، به مطالعه و تحلیل خوشه‌بندی خلاصه مقالات و حق امتیاز فناوری نانو پرداخته است. این پژوهش، با تمرکز بر تکنیک‌های پیشرفته تحلیل داده و پردازش متن، به دنبال شناسایی و دسته‌بندی زیر حوزه‌های فناوری نانو است. برای این purpose، تیمورپور از رویکردی ترکیبی استفاده کرده است که شامل خوشه‌بندی پویا، متن‌کاوی و تحلیل پیوند است. خوشه‌بندی پویا، یک روش تحلیل داده است که به شناسایی گروه‌های مرتبط از داده‌ها کمک می‌کند. متن‌کاوی، یک تکنیک پردازش زبان طبیعی است که به استخراج مفاهیم و الگوهای معنایی از متن‌ها می‌پردازد. تحلیل پیوند، یک روش است که روابط بین مفاهیم و داده‌ها را بررسی می‌کند. در این مطالعه، تیمورپور از معیار ضریب سیلوئت برای مقایسه شباهت بین خوشه‌ها استفاده کرده است. ضریب سیلوئت، یک شاخص است که به‌اندازه شباهت یا تفاوت بین دو خوشه کمک می‌کند. با استفاده از این روش، تیمورپور به شناسایی و دسته‌بندی زیر حوزه‌های فناوری نانو پرداخته است. این روش ترکیبی از خوشه‌بندی پویا، متن‌کاوی و تحلیل پیوند، امکان شناسایی روندها و الگوهای نوظهوری در حوزه فناوری نانو را فراهم می‌کند و به تحلیل داده‌های پیچیده‌تر در این زمینه کمک می‌کند.

باباآغائی (۱۳۹۱)، در یک پژوهش با عنوان «کشف ساختار درونی مطالعات روان‌شناسی مثبت به روش متن‌کاوی»، به تحلیل و دسته‌بندی پیچیده‌ای از داده‌های مربوط به این حوزه پرداخت. این مطالعه، از ترکیبی از روش‌های پیشرفته تحلیل داده و پردازش متن استفاده کرد. ازجمله این روش‌ها، خوشه‌بندی بود که برای کشف ساختار درونی و روابط موضوعی بین مقالات استفاده شد. خوشه‌بندی، یک تکنیک تحلیل داده است که به گروه‌بندی داده‌ها براساس ویژگی‌های مشترک آن‌ها کمک می‌کند. در این پژوهش، باباآغائی از خوشه‌بندی برای شناسایی گروه‌های مرتبط از مقالات روان‌شناسی مثبت استفاده کرد. این روش، امکان تحلیل داده‌های پیچیده و شناسایی الگوهای نا آشکاری در داده‌های متنی را فراهم می‌کند. علاوه‌بر خوشه‌بندی، باباآغائی از متن‌کاوی نیز استفاده کرد. متن‌کاوی، یک رویکرد پردازش زبان طبیعی است که به استخراج مفاهیم و روابط معنایی از متن‌ها می‌پردازد. این روش، امکان تحلیل دقیق‌تر از مفاهیم و نظریه‌های روان‌شناسی مثبت را فراهم کرد. در ادامه، باباآغائی به تحلیل ارتباط موضوعی بین مقالات حوزه روان‌شناسی مثبت که از پایگاه اسکوپوس جمع‌آوری شده بودند، پرداخت. این تحلیل، به شناسایی مقالات با بیشترین و کمترین درصد شباهت در موضوعات مختلف کمک کرد. نتیجه این تحلیل، یک ساختار پیچیده‌تر و دقیق‌تر از روابط بین مقالات و نظریه‌های روان‌شناسی مثبت را ارائه داد. این روش ترکیبی از خوشه‌بندی و متن‌کاوی، امکان کشف ساختار درونی و روابط موضوعی پیچیده‌تر در حوزه روان‌شناسی

مثبت را فراهم کرد و به تحلیل داده‌های پیچیده‌تر در این زمینه کمک کرد.

قنادی نژاد و همکاران (۱۴۰۲)، در یک پژوهش با عنوان «تحلیل موضوعی تولیدات علمی پژوهشگران ایرانی در حوزه علم اطلاعات و دانش‌شناسی با رویکرد متن‌کاوی»، به بررسی و تحلیل پیچیده‌ای از داده‌های مربوط به این حوزه پرداخته‌اند. این مطالعه، نقش مهمی در نشان دادن نقاط ضعف، کاستی‌ها و مسیر پیشرفت و توسعه رشته علم اطلاعات و دانش‌شناسی ایفا می‌کند. در این تحقیق، پژوهشگران از روش‌های متن‌کاوی برای تجزیه و تحلیل مقالات استفاده کرده‌اند. متن‌کاوی، یک رویکرد پیشرفته پردازش زبان طبیعی است که به استخراج مفاهیم و روابط معنایی از متن‌ها می‌پردازد. این روش، امکان تحلیل دقیق‌تر از مفاهیم و نظریه‌های علم اطلاعات و دانش‌شناسی را فراهم می‌کند. پژوهشگران در این مطالعه، از دو منبع اصلی داده استفاده کرده‌اند: مقالات مستخرج از وبگاه نشریات ایرانی نمایه شده در پایگاه ISI در فاصله سال‌های ۱۳۹۸-۱۳۵۱، و مقالات انجام‌شده توسط پژوهشگران ایرانی که از سال ۲۰۱۹-۱۹۴۵ در پایگاه Web of Science نمایه شده‌اند. این روش ترکیبی از متن‌کاوی و تحلیل داده‌های ساختاریافته، امکان شناسایی الگوهای نا آشکاری و روابط پیچیده‌تر در داده‌های متنی را فراهم می‌کند. یافته‌های این مطالعه نشان می‌دهند که اکثر پژوهش‌های این رشته به حوزه‌های مختلفی از جمله کتابخانه‌ها و مراکز اطلاعاتی و آرشیو، پژوهش، مطالعه و نشر، علم‌سنجی و اطلاع‌سنجی و اینترنت و مطالعات وب اختصاص یافته‌اند. این تحلیل، به شناسایی مناطق قوت و ضعف در حوزه علم اطلاعات و دانش‌شناسی کمک کرده و مسیرهای احتمالی برای توسعه این رشته را پیشنهاد می‌کند. نتیجه این تحلیل، یک تصویر دقیق‌تر و جامع‌تر از وضعیت فعلی و مسیرهای پیشرفت مستقبل این حوزه را ارائه می‌دهد.

پوراتا پونز و همکاران[1] (۲۰۱۱)، در یک پژوهش با عنوان «کشف موضوع بر اساس تکنیک‌های متن‌کاوی»، به توسعه و کاربردیابی پیشرفته‌ای از روش‌های متن‌کاوی در زمینه‌های مختلف پرداخته‌اند. این مطالعه، هدفمندانه به بررسی و تحلیل تکنیک‌های متن‌کاوی برای خلاصه‌بندی، طبقه‌بندی و دسته‌بندی موضوعات می‌پردازد. متن‌کاوی، یک رویکرد پیشرفته پردازش زبان طبیعی است که به استخراج مفاهیم، روابط معنایی و الگوهای نا آشکاری از متن‌ها می‌پردازد. در این تحقیق، پژوهشگران از تکنیک‌های متن‌کاوی برای تحلیل و پردازش داده‌های متنی استفاده کرده‌اند. این روش‌ها، امکان شناسایی ساختارهای معنایی پیچیده‌تر در متن‌ها را فراهم می‌کنند و به تحلیل داده‌های متنی ساختارنیافته کمک می‌کند. یافته‌های این مطالعه نشان می‌دهند که متن‌کاوی می‌تواند به‌طور مؤثری در زمینه‌های مختلفی از جمله پردازش زبان طبیعی، بازیابی اطلاعات، یادگیری ماشینی و تحلیل داده‌های متنی کاربرد داشته باشد. این روش، به شناسایی و استخراج مفاهیم کلیدی، شناسایی روابط بین مفاهیم و حتی کشف الگوهای نا آشکاری در متن‌ها کمک می‌کند. نتیجه این تحلیل، یک تصویر دقیق‌تر و جامع‌تر از نحوه عملکرد تکنیک‌های متن‌کاوی در پردازش و تحلیل داده‌های متنی ارائه می‌دهد. این مطالعه، اهمیت و کاربردیابی متن‌کاوی در زمینه‌های مختلف را نشان می‌دهد و به توسعه و بهبود این روش در پردازش داده‌های متنی کمک می‌کند.

یان و همکاران[2] (۲۰۱۵)، در یک پژوهش با عنوان «تجزیه و تحلیل مقالات پژوهشی در تجارت الکترونیکی بر اساس رویکرد متن‌کاوی»، به تحلیل و پردازش پیچیده‌ای از داده‌های مربوط به این حوزه پرداخته‌اند. این مطالعه، در سال ۲۰۱۵ انجام شد و از روش متن‌کاوی برای استخراج و تحلیل مفاهیم کلیدی از مقالات در زمینه تجارت الکترونیکی استفاده کرد. متن‌کاوی، یک رویکرد پیشرفته پردازش زبان طبیعی است که به استخراج مفاهیم، روابط معنایی و الگوهای نا آشکاری از متن‌ها می‌پردازد. در این تحقیق، پژوهشگران ۶۸ کلیدواژه کلیدی را از مقالات بین سال‌های ۲۰۱۳-۲۰۰۲ استخراج کردند. این روش، امکان شناسایی ساختارهای معنایی پیچیده‌تر در متن‌ها را فراهم می‌کند و به تحلیل داده‌های متنی ساختارنیافته کمک می‌کند. یافته‌های این مطالعه نشان می‌دهند که تجارت الکترونیکی به سه حوزه اصلی تکنولوژی، مدیریت و مشتریان تقسیم می‌شود. این تحلیل، به شناسایی مناطق اصلی تمرکز و توسعه در حوزه تجارت الکترونیکی کمک می‌کند. علاوه‌بر این، پژوهشگران دریافتند که تمام مقالات در زمینه تجارت الکترونیکی در هفت حوزه مهمی همپوشانی دارند: اینترنت، رفتار مصرف‌کننده، رضایت مشتری، خرید آنلاین، شهرت و مدیریت دانش. این یافته‌ها، به شناسایی زمینه‌های مشترک و ارتباطات بین مفاهیم در حوزه تجارت الکترونیکی کمک می‌کند و به تحلیل داده‌های پیچیده‌تر در این زمینه کمک می‌کند. نتیجه این تحلیل، یک تصویر دقیق‌تر و جامع‌تر از ساختار و حوزه‌های اصلی تجارت الکترونیکی ارائه

1. Pons-Porrata et al

2. Yan et al

می‌دهد و به توسعه و بهبود این حوزه در زمینه‌های مختلف کمک می‌کند.

لیو و همکاران[1] (۲۰۲۴)، در یک پژوهش با عنوان «روش خلاصه‌سازی خودکار متن بر اساس الگوریتم Text Rank بهبودیافته و خوشه‌بندی K-Means"، به توسعه و کاربردیابی پیشرفته‌ای از روش‌های خلاصه‌سازی و استخراج عبارات کلیدی از متن‌ها پرداخته‌اند. این مطالعه، هدفمندانه به ترکیب و بهبود دو رویکرد پیشرفته پردازش زبان طبیعی استفاده می‌کند: در این تحقیق، پژوهشگران این دو روش را ترکیب کرده‌اند تا یک روش خلاصه‌سازی خودکار متن را توسعه دهند. این روش ترکیبی، شامل دو فرایند اصلی است: خوشه‌بندی و مرتب‌سازی. خوشه‌بندی، متن را به جملات تقسیم می‌کند و آن‌ها را به گروه‌های مرتبط دسته‌بندی می‌کند. مرتب‌سازی، جمله‌های خوشه‌بندی شده را براساس نمره آن‌ها مرتب می‌کند. در نهایت، این روش، یک خلاصه از متن را انتخاب می‌کند که شامل عبارات کلیدی و مفاهیم اصلی است. هدف این ترکیب، دستیابی به نوآوری منحصربه‌فردی در انتخاب مراکز اولیه است که می‌تواند مشکلات بهینه‌سازی خوشه‌بندی و افزونگی را بهتر حل کند. این روش، امکان تحلیل دقیق‌تر و فهم عمیق از مفاهیم و نظریه‌های متن‌کاوی را فراهم می‌کند. نتیجه این ترکیب، یک روش خلاصه‌سازی خودکار متن است که می‌تواند به‌طور مؤثری از داده‌های متنی ساختارنیافته استخراج کند و مفاهیم کلیدی را شناسایی و دسته‌بندی کند.

این روش، به تحلیل داده‌های پیچیده‌تر در حوزه‌های مختلف پردازش زبان طبیعی کمک می‌کند و به توسعه و بهبود این حوزه در زمینه‌های مختلف کمک می‌کند.

در چارچوب ادبیات تحقیقاتی موجود، تحقیقی که به استخراج و ارائه موضوعات رأس اصطلاحنامه علم اطلاعات و دانش‌شناسی مرتبط بپردازید، در حوزه متن‌کاوی، هنوز مورد مطالعه قرار نگرفته است. این خلأ تحقیقاتی، ضرورت انجام یک مطالعه جامع در این زمینه را آشکار می‌سازد. پژوهش‌های اخیر در حوزه متن‌کاوی، نشان‌دهنده تنوع قابل‌ملاحظه‌ای در موضوعات مورد مطالعه هستند. از جمله این موضوعات می‌توان به:

- بررسی تغییرات قیمت سهام بلافاصله پس از انتشار مقالات خبری
- کشف موضوعات جدید با استفاده از تکنیک‌های متن‌کاوی
- توسعه الگوریتم‌های خوشه‌بندی در داده‌کاوی و متن‌کاوی

این تحقیقات، به دست آوردن نتایج قابل‌توجهی در زمینه‌های مختلف صنعتی و تخصصی داشته‌اند. همچنین، استفاده از تکنیک‌های متن‌کاوی، به بهبود فرآیندهای تحلیل و استخراج اطلاعات از داده‌های متنی کمک کرده است. با توجه به این نکته، این پژوهش به دنبال پاسخ به سؤال اصلی خود در مورد عدم وجود تحقیقی که به استخراج و ارائه موضوعات رأس اصطلاحنامه علم اطلاعات و دانش‌شناسی مرتبط بپردازید، خواهد بود. هدف اصلی این تحقیق، بررسی این موضوع و ارائه راه‌حل‌هایی برای پر کردن این شکاف تحقیقاتی در ادبیات علمی است.

روش انجام پژوهش

این پژوهش از نوع کاربردی است که از روش متن‌کاوی برای حل مسائل استفاده می‌کند. مراحل این روش شامل پیش‌پردازش داده، استخراج، انتخاب ویژگی و ایجاد الگو و مدل است. ابتدا یک مجموعه داده مناسب از چکیده مقالات بین‌المللی در رشته علم اطلاعات و دانش‌شناسی (۱۹۶۸–۲۰۲۲) از پایگاه استنادی وب آو ساینس جمع‌آوری می‌شود. سپس داده‌ها به‌طور دقیق پردازش می‌شوند تا به داده‌های قابل استفاده تبدیل شوند. در ادامه، الگوریتمی بر روی این داده‌ها اجرا می‌شود تا یک الگو مناسب برای استخراج و پیشنهاد اصطلاحات رأس برای اصطلاحنامه علم اطلاعات و دانش‌شناسی ارائه دهد. این روش نه‌تنها بهینه‌سازی فرایندهای تحلیل و استخراج اطلاعات از داده‌های متنی را ممکن می‌سازد، بلکه می‌تواند نتایج مفیدی را برای صنایع و حوزه‌های مختلف فراهم کند.

برای پیاده‌سازی الگوریتم TextRank، مراحل پیش‌پردازش متن به شرح زیر انجام می‌شود:

۱. استانداردسازی نمایش کلمات:

– نشانه‌سازی: به جای نوشتن کلمات با حروف بزرگ و کوچک، همه کلمات را به حروف کوچک تبدیل می‌کنند.

– حذف توقف: علائم توقف مانند نقطه، ویرگول، کاما و غیره را حذف می‌کنند.

– ریشه‌یابی/لماتیشن‌سازی: کلمات را به اشکال ساده‌تر تبدیل می‌کنند (مثلاً "running" به "run").

۲. نمایش نمودار TextRank:

– متن را به‌عنوان یک نمودار نمایش می‌دهند.

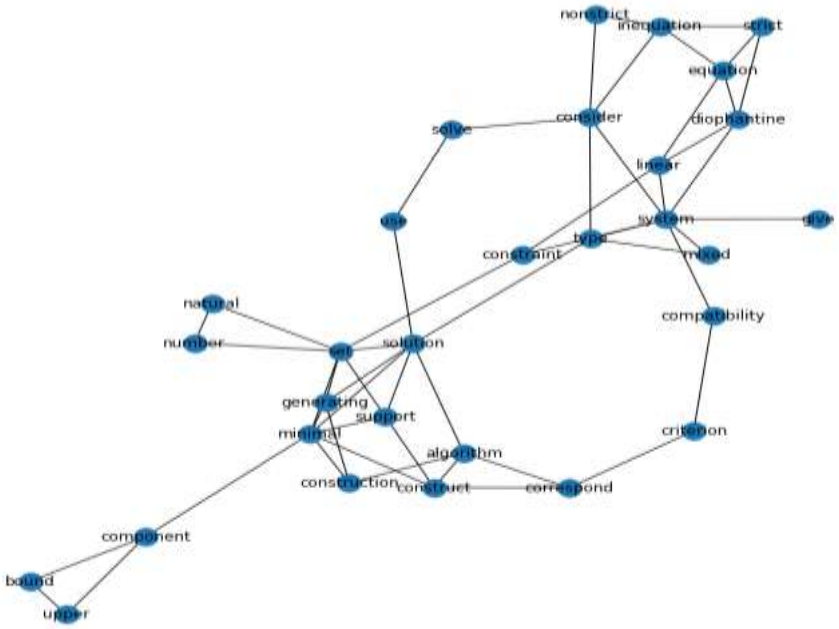
– گره‌ها کلمات یا عبارات را نشان می‌دهند.

– یال‌ها روابط بین گره‌ها را نمایش می‌دهند.

– هر گره به یک کلمه یا عبارت خاص در متن نسبت داده می‌شود.

1. Liu et al

این نمودار به صورت یک شبکه پیچیده از کلمات و روابط بین آن‌ها ظاهر می‌گیرد. برای مثال، در شکل زیر یک نمونه از نمودار TextRank نشان داده شده است که در آن گره‌ها کلمات را و یال‌ها روابط بین گره‌ها را نمایش می‌دهند. این نمودار برای اعمال الگوریتم TextRank و استخراج اصطلاحات کلیدی آماده می‌شود.



نمودار ۱. نمودار روابط بین کلمات در الگوریتم text rank

برای تحلیل روابط هم‌زمانی در متن، یک ماتریس هم‌زمانی طراحی می‌شود که نشان‌دهنده حضور فیزیکی کلمات و عبارات در مجاورت یکدیگر در یک پنجره زمانی خاص است. این ماتریس معمولاً پراکنده است، زیرا اکثر جفت کلمات یا عبارات در تمام متن‌ها وجود ندارند.

پس از ساخت ماتریس هم‌زمانی، مرحله بعدی محاسبه وزن یال‌های متصل‌کننده گره‌ها در نمودار TextRank است. وزن هر یال نمایانگر شدت یا قدرت رابطه بین دو کلمه یا عبارت است. در الگوریتم TextRank، این وزن معمولاً براساس معیارهای شباهت یا ارتباط بین کلمات یا عبارات هم‌زمان محاسبه می‌شود.

برای محاسبه وزن یال‌ها، چندین معیار شباهت وجود دارند که می‌توان برای این هدف استفاده کنند:

۱. شباهت لفظی: مقایسه توالی حروف کلمات؛

۲. شباهت معنایی: اندازه‌گیری شباهت مفهوم بین کلمات؛

۳. شباهت ساختاری: ارزیابی ساختار نحوی کلمات؛

۴. شباهت زمینه‌ای: اندازه‌گیری حضور مشترک کلمات در زمینه‌های خاص.

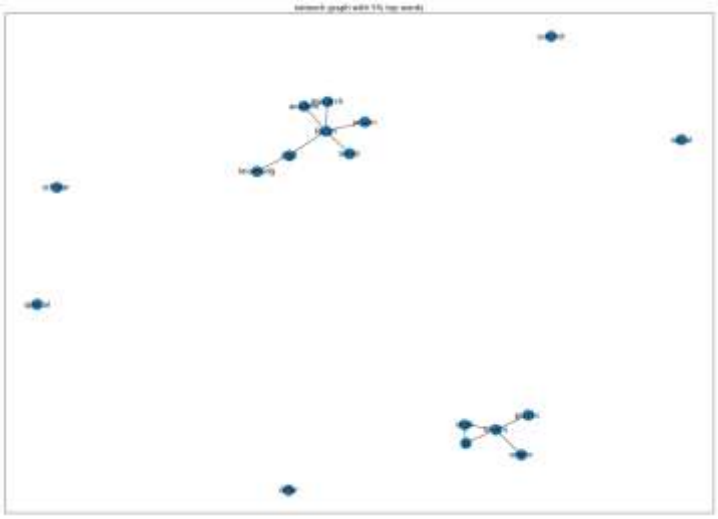
براساس انتخاب معیار مناسب، وزن یال‌ها محاسبه می‌شوند که این وزن‌ها نقش کلیدی در شناسایی اصطلاحات کلیدی و استخراج موضوعات از متن دارند.

یافته‌ها

در این پژوهش، با توجه به جامعه پژوهش که شامل تمامی مقالات نمایه شده بین‌المللی در حوزه علم اطلاعات و دانش‌شناسی در پایگاه استنادی وب آو ساینس است، تعداد ۴۲۹۷ رکورد بازیابی شد. پس از حذف اطلاعاتی که نیازی به آن‌ها نبود، تعداد رکوردها کاهش یافت و در نهایت منجر به تولید ۴۲۹۷ رکورد شد.

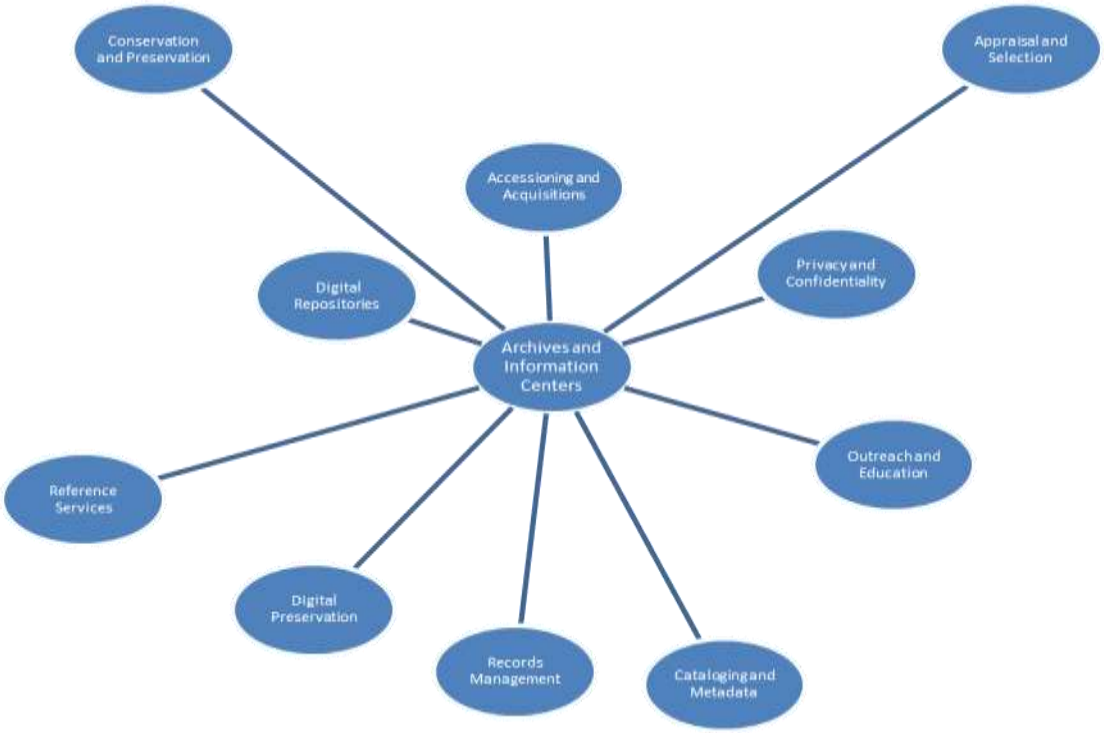
پس از رسم نمودارهای مربوطه، به سمت توسعه یک مدل مفهومی حرکت کردیم که در آن هر گراف به‌عنوان یک ورودی یا مدخل برای اصطلاحنامه تعریف می‌شود. در این چارچوب، گره مرکزی که در قسمت اولیه پروژه شناسایی شده است، به‌عنوان اصطلاح کلیدی یا «رأس» محسوب می‌گردد .علاوه‌بر این، عبارات و مفاهیمی که از طریق یال‌ها (یا اتصالات) به این گره مرکزی متصل هستند، به‌عنوان عبارات وابسته و مرتبط به اصطلاح کلیدی در نظر گرفته می‌شوند. این عبارات براساس امتیازات و اولویت‌های خود مرتب‌سازی

می‌شوند تا اهمیت نسبی آن‌ها در رابطه با اصطلاح کلیدی نشان داده شود. این رویکرد نه‌تنها به ساختاردهی بهتر اطلاعات کمک می‌کند بلکه امکان تحلیل عمیق‌تر و فهم بهتری از روابط میان مفاهیم مختلف را فراهم می‌آورد. نمودارهای ارائه شده در ادامه، بر اساس فرآیند مرتب‌سازی و سازمان‌دهی شکل‌های حاصل از مراحل قبلی، در چارچوب یک مدل مفهومی طراحی شده‌اند که به زبان ساده و قابل‌فهمی بیانگر کلیت پروژه است. این مدل مفهومی با هدف ارائه تصویری روشن و جامع از ساختار و روابط موجود در پژوهش طراحی شده است، به‌گونه‌ای که مخاطبان بتوانند به‌سادگی درک کاملی از کلیت کار داشته باشند.



شکل ۱ روابط بین اصطلاحات

توسعه همکاری و صلح، فدراسیون فوتبال ایران، کنفدراسیون فوتبال آسیا.



شکل ۲. نمودار ترسیم شده از نتایج به دست آمده

یک نمونه از نمودارهای بازیابی شده در شکل ۲ نمایش داده شده است. با توجه به نمودار، عبارت " Archives and Information Centers" به‌عنوان اصطلاح کلیدی یا «رأس» شناسایی شده است. اصطلاحات مرتبط با این اصطلاح رأس براساس وزن و اهمیت آن‌ها مرتب‌سازی شده‌اند؛ به‌گونه‌ای که هرچه وزن یک کلمه در این گروه بیشتر باشد، فاصله گره مربوطه تا گره اصلی (یعنی اصطلاح رأس) کمتر خواهد بود. این امر نشان‌دهندهٔ ارتباط مفهومی قوی‌تر بین اصطلاحات وابسته و اصطلاح رأس است.

وزن هریک از عبارات در گروه اول به صورت زیر محاسبه شده است:

Accessioning and Acquisitions: 3.06001
Digital Repositorie: 2.02549
Privacy and Confidentiality: 2.02158
Outreach and Education: 1.89652
Digital Preservation :1.79625
Records Management: 1.56985
Cataloging and Metadata: 1.53874
Reference Services: 0.962548
Conservation and Preservation: 0.52145
Appraisal and Selection: 0.015987

سپس در پایان با توجه به نتایج به دست آمده از تجزیه و تحلیل‌ها یک نتیجه کلی در قالب ۱۷ موضوع رأس ارائه داده شد. بدین ترتیب با اندازه‌گیری امتیاز هریک از عبارات، عباراتی که بیشترین امتیاز را داشتند به‌عنوان موضوعات رأس برگزیده شدند که به ترتیب زیر است:

جدول ۱. اصطلاح های رأس برگزیده

اصطلاح رأس برگزیده شده
Archives and information centers
Artificial Intelligence
Bibliographic Databases
Classification
Collection Development
Controlled Vocabulary
Digital libraries
Information Organization
Information Retrieval and data Extraction
Information Science and library
Information Systems and Resources
Knowledge Management
Libraries and Community Services
Metadata
Reference Services
Subject Headings
Scientometrics

پاسخ به پرسش‌های پژوهش

سؤال اول: با استفاده از الگوهای به دست آمده چه واژه‌های رأسی برای اصطلاحنامه علم اطلاعات و دانش‌شناسی می‌توان پیشنهاد داد؟

با توجه به الگوهای به دست آمده از تجزیه و تحلیل متون حوزه علم اطلاعات و دانش‌شناسی با استفاده از تکنیک‌های متن‌کاوی می‌توان واژه‌ها یا عبارات زیر را به‌عنوان موضوع رأس پیشنهاد داد:

Information centers and Archives
Artificial Intelligence
Bibliographic Databases
Classification
Collection Development
Vocabulary
Digital libraries

Information Organization
Information Retrieval and data Extraction
Information Science and Library
Information Systems and Resource
Knowledge Management
Libraries and Community Services
Metadata
Reference Service
Subject Headings
Scientometrics

سؤال دوم: چه تکنیک‌های متن‌کاوی برای استخراج اطلاعات موضوعات رأس از متن موردنظر مورد استفاده قرار گرفت؟

برای استخراج اطلاعات موضوعات رأس از متون موردنظر با استفاده از تکنیک‌های متن‌کاوی، از تکنیک‌های زیر استفاده شد.

تحلیل متن[1]: تحلیل متن فرآیندی است که در آن، متن‌ها برای استخراج، تحلیل و درک اطلاعات موردنظر بررسی شد. این فرآیند شامل بررسی ساختار متن، شناسایی کلمات کلیدی، اصطلاحات، عبارات مرتبط و موضوعات مورد بحث است (زادگانکار و آگراوال، ۲۰۲۴).

تحلیل ساختاری متن[2]: از این تحلیل برای بررسی ساختار پاراگراف‌ها، بندها و عبارات مرتبط با موضوعات استفاده شد.

تحلیل معنایی متن[3]: از این تحلیل برای تحلیل ساختار پاراگراف‌ها، بندها و عبارات مرتبط با موضوعات مهم متن استفاده شد.

تحلیل کلمات کلیدی[4]: از این تحلیل برای تحلیل و شناسایی کلمات کلیدی و اصطلاحات استفاده شد.

تحلیل متن با استفاده از ترکیب کلمات[5]: از این تکنیک برای شناسایی کلمات و عبارات که با هم در متن به‌طور مکرر به ظاهر می‌شوند استفاده شد.

تحلیل متن با استفاده از ترکیب کلمات و عبارات با توجه به معنا[6]: از این تکنیک برای شناسایی عبارات مرتبط باهم توجه به معنای آن‌ها استفاده شد.

پردازش زبان طبیعی[7]: پردازش زبان طبیعی (NLP) یک شاخه از هوش مصنوعی است که به رایانه‌ها اجازه می‌دهد تا زبان انسانی را به‌طور معنادار درک کنند، تحلیل کنند و با آن کار کنند. این فرآیند شامل ترجمه ماشینی، تشخیص گفتار، تشخیص احساسات و تحلیل متن است. در زمینه استخراج اطلاعات موضوعات رأس از متن‌های علمی، NLP بسیار کاربردی و مفید است (پوراتا پونز و همکاران، ۲۰۱۱)

سؤال سوم: چه چالش‌هایی در استخراج اطلاعات موضوعات رأس از متن‌های علمی وجود داشت؟

پیچیدگی و حجم زیاد متون

متنی که به‌عنوان جامعه پژوهش از وب آو ساینس استخراج شده بود تا مورد تجزیه و تحلیل قرار گیرد دارای حجم زیاد و پیچیدگی خاص که شامل استفاده از زبان واژگانی و ترکیبات کلمات متعدد بود. این پیچیدگی تا حدودی باعث شد استخراج اطلاعات دقیق و معنادار از متن‌ها بسیار دشوار باشد و زمان طولانی صرف پیش‌پردازش و پردازش داده‌ها شود (بارونی و ستیاسیلان[8]، ۲۰۲۰).

عدم توانایی در فهم زبان طبیعی

الگوریتم TextRank بر اساس هم‌رخدادی کلمات کار می‌کند و به‌طور مستقیم از معنا یا ساختار جملات استفاده نمی‌کند. این موضوع باعث شد که الگوریتم مورد استفاده در فهم پیچیدگی‌های زبان طبیعی، مانند استعاره‌ها، بازی با کلمات، یا جملات چندمعنایی، محدودیت داشته باشد (بارونی و ستیاسیلان، ۲۰۲۰).

نیاز به تنظیم دقیق پارامترها

عملکرد TextRank بسیار وابسته به تنظیم دقیق پارامترهایی مانند اندازه پنجره هم‌رخدادی، تعداد تکرار و ضریب کاهش بود. انتخاب نادرست این پارامترها می‌توانست منجر به نتایج غیرقابل اعتمادی شود (لیو و همکاران، ۲۰۲۴).

تغییر معنا و استفاده متغیر از کلمات

کلمات و عبارات در متن به صورت‌ها و شکل‌های متعددی استفاده شده بود که با توجه به زمینه و موضوع متن معنای متفاوتی داشت، ازاین‌رو تشخیص معنی دقیق موضوعات رأس چالش‌برانگیز شد.

1. Text Analysis
2. Text Structure Analysis
3. Text Semantic Analysis
4. Keyword Analysis
5. Co-occurrence Analysis
6. Semantic Phrase Analysis
7. Natural Language Processing - NLP
8. Baruni & Sathiaseelan

نوع و سبک نوشتار

مقالاتی که از وب آو ساینس استخراج شد نوع و سبک نوشتار متفاوتی داشتند، از جمله مقالات تحقیقاتی، مقالات مروری و غیره. این تنوع در نوع و سبک نوشتار باعث شد تا استخراج اطلاعات موضوعات رأس چالش‌برانگیز باشد.

نیاز به دانش متخصص

برای استخراج دقیق و معنادار اطلاعات موضوعات رأس از متن علمی موردنظر، نیاز به دانش متخصص در حوزه‌های زبان‌شناسی، علم اطلاعات و دانش‌شناسی و کامپیوتر بود تا با راهنمایی‌های آن‌ها مشکلات و چالش‌هایی که وجود داشت حل شود.

نیاز به تکنیک‌های پیشرفته پردازش زبان طبیعی

برای استخراج دقیق و معنادار اطلاعات موضوعات رأس از متن علمی موردنظر، نیاز به استفاده از تکنیک‌های پیشرفته پردازش زبان طبیعی و یادگیری ماشین بود. امتحان و تست کردن برخی الگوریتم‌ها در جهت یافتن الگوریتم موردنظر، صرف‌نظر از وقت و زمان چالش‌برانگیز بود.

این چالش‌ها نشان داد استخراج اطلاعات موضوعات رأس از متن‌های علمی یک فرآیند پیچیده و نیازمند توجه و دانش است.

بحث و نتیجه‌گیری

نتایج حاصل از پژوهش حاضر به شناسایی موضوعات رأس اصطلاحنامه علم اطلاعات و دانش‌شناسی، بر اساس تکنیک‌های متن‌کاوی منجر شده است. نتایجی که در جدول ۵-۱ نشان داده شده است به‌عنوان موضوعات رأس اصطلاحنامه علم اطلاعات و دانش‌شناسی شناسایی شده است.

همگام با رشد پژوهش‌ها و پیشرفت فناوری‌های جدید در ۱۰ سال گذشته، به دلیل رویکرد پژوهشگران به استفاده از ابزار جدید، موضوعاتی بیشتر مورد توجه پژوهشگران در دنیا قرار گرفته است که مستلزم استفاده از این ابزار نوظهور بوده و این موضوعات با فناوری‌ها و ابزارهای مبتنی بر فناوری‌ها و زندگی امروزی ارتباط یافته است .در دنیای علم اطلاعات و دانش‌شناسی، تولید اصطلاحنامه به صورت سنتی به دلیل چندین عامل رخ داده است.

یکی از دلایل اصلی این امر، نگرانی از کیفیت و دقت اطلاعات است. اصطلاحنامه‌های سنتی توسط متخصصان انسانی ویرایش و بررسی دقیق شده‌اند تا اطمینان حاصل شود که اطلاعاتی که ارائه می‌شود دقیق و قابل اعتماد است. این رویکرد سنتی به‌ویژه در مواردی که نیاز به تفسیر دقیق و فهم عمیق از مفاهیم وجود دارد، اهمیت زیادی پیدا می‌کند. با این حال، استفاده از فناوری‌ها مانند متن‌کاوی برای تولید اصطلاح‌نامه‌ها می‌تواند مزایای بسیاری داشته باشد. متن‌کاوی، که یکی از شاخه‌های پردازش زبان طبیعی[1] است، به ماشین‌ها اجازه می‌دهد تا زبان انسانی را درک کنند و آن را به صورت خودکار پردازش کنند. این فناوری می‌تواند حجم زیادی از داده‌های غیرمعمول (مانند پست‌های رسانه‌های اجتماعی، نظرات محصول، بازخوردهای مشتری، و غیره) را تجزیه و تحلیل کند که این کار بدون استفاده از فناوری بسیار زمان‌بر و گاهی اوقات ناممکن است. استفاده از متن‌کاوی برای تولید اصطلاح‌نامه‌ها می‌تواند به سازمان‌ها کمک کند تا به صورت کارآمد و کم هزینه اطلاعات مربوطه را از داده‌های خام استخراج کنند. همچنین این فرآیند به سازمان‌ها امکان می‌دهد تا به‌سرعت و با دقت بالا، اطلاعات مرتبط را شناسایی کنند که این امر برای تصمیم‌گیری‌های مبتنی بر داده و پاسخگویی به نیازهای کاربران حیاتی است. بنابراین، استفاده از فناوری‌های مدرن مانند متن‌کاوی نه‌تنها می‌تواند به بهبود کیفیت و سرعت تولید اصطلاح‌نامه‌ها کمک کند، بلکه همچنین می‌تواند به سازمان‌ها کمک کند تا به‌طور مؤثرتر از داده‌های بزرگ و پیچیده استفاده کنند، که در دنیای امروز بیش از همیشه ضروری است.

مقایسه نتایج حاصل از پژوهش‌های صورت گرفته در خصوص تکنیک‌های متن‌کاوی برای کشف موضوعات پنهان داخل منابع که در داخل کشور انجام شده است بیانگر این است که استفاده از تکنیک‌های متن‌کاوی برای تولید اصطلاحنامه در اولویت پژوهشگران قرار نگرفته یا به نسبت، کمتر مورد توجه قرارگرفته است. با این حال در سایر حوزه‌ها مانند فناوری نانو به‌طور مثال برای کشف موضوعات و روندهای نوظهور در این حوزه به خوشه‌بندی خلاصه مقالات پرداخته‌اند سپس با استفاده از تکنیک‌های متن‌کاوی بر اساس شباهت مقایسه‌های انجام داده شده موضوعات مهم و اساسی را در این حوزه شناسایی کرده‌اند (تیمورپور، ۱۳۸۸). یا در حوزه مهندسی صنایع با استفاده از تکنیک‌های متن‌کاوی موضوعات پرکاربرد در پنج دهه اخیر این موضوع را شناسایی کرده‌اند (ابوالصدق، ۱۳۹۰).

1. NLP

همچنین از این رویکرد برای نگاشت ساختار مفهومی علم اطلاعات و دانش‌شناسی نیز استفاده کرده‌اند که ۱۵۰ مفهوم را براساس الگوریتم TF-IDF انتخاب کردند (حسن‌زاده و همکاران، ۱۳۹۷).

بنابراین به‌طور کلی، از مقایسه نتایج این پژوهش و نتایج پژوهش‌های انجام شده در داخل کشور، می‌توان گفت که هنوز روند جهانی تولید اصطلاحنامه با استفاده از تکنیک‌های متن‌کاوی در پژوهش‌های داخلی به‌درستی پیگیری نشده است. این موضوع بر عهده تصمیم‌گیران و برنامه‌ریزان علمی است تا با ایجاد ابزارها و شرایط مناسب، تولید اصطلاحنامه‌ها را به سمت استفاده از رویکردهای مدرن و جهانی هدایت کنند. دسته دیگری از پژوهش‌های صورت گرفته در داخل کشور در این خصوص به منظور ارائه مدل برای استخراج اطلاعات از مستندات متنی مبتنی بر متن‌کاوی (آقاکاردان و کیهانی‌نژاد، ۱۳۹۱) و تحلیل محتوایی مقالات علمی با استفاده از متن‌کاوی انجام شده است که نتایج قابل‌ملاحظه‌ای از این پژوهش‌ها حاصل شده است. در واقع این دسته از پژوهشگران با توجه به نیازهای جامعه علمی و ضرورت استفاده از فناوری‌های نوین توانسته‌اند قدم مهمی را در جهت تحلیل و استخراج موضوعات مهم از متون علمی بردارند. نتایج به دست آمده از پیشینه این پژوهش که توسط پژوهشگران خارج از کشور انجام شده است نتایج جالب و قابل‌توجهی دارد. بررسی نتایج پژوهش‌های انجام شده در خارج از کشور را در این خصوص می‌توان در چهار دسته کلی تقسیم‌بندی کرد که شامل: استفاده از رویکردهای متن‌کاوی برای پیش‌بینی روندها (آس[1]، ۲۰۱۱)، خلاصه‌سازی متون (پوراتا پونز و همکاران، ۲۰۱۱؛ لیو و همکاران، ۲۰۲۴)،

مدل‌سازی موضوعی (سیلواتانانوسارن و کولکانجاناپیبان[2]، ۲۰۲۲) و تجزیه و تحلیل متن جهت استخراج موضوعات اساسی و کلیدی حوزه موردنظر (بارونی و ستیاسیلان[3]، ۲۰۲۰) است. در دسته دیگری از پژوهش‌ها که به‌عنوان قدم مهمی در فرایند خودکارسازی اصطلاح‌نامه‌ها به شمار می‌رود پژوهشی است که در قالب تحقیق و بررسی الگوریتم ساخت هوشمند اصطلاحنامه دانش موضوعی بر اساس منابع ادبی است که با هدف بررسی و توسعه یک الگوریتم برای ساخت اصطلاحنامه براساس داده‌های منابع گسترده ادبیات است (وانگ و همکاران[4]، ۲۰۲۱) که همسو با این پژوهش است.

تحلیل دقیق‌تر نتایج به دست آمده، تأکید می‌کند که تلفیقی از رویکردهای مختلف موجود در حوزه متن‌کاوی و تلاش‌های بیشتری برای افزایش توجه پژوهشگران به استفاده از این فناوری، می‌تواند گامی مهم و تعیین‌کننده در تولید اصطلاحنامه‌های علمی باشد. این ترکیب و تمرکز، نه‌تنها به بهبود کیفیت و دقت اطلاعات منجر می‌شود، بلکه همچنین می‌تواند به توسعه ابزارها و روش‌های جدید برای مدیریت و تحلیل داده‌های علمی کمک کند.

پیشنهادهای پژوهش

به صورت کلی برای پژوهش‌های آینده موارد زیر برای پژوهشگران پیشنهاد می‌گردد:

- استفاده از محتوای منابع سایر پایگاه‌های اطلاعاتی.

در این پژوهش داده‌های موردنظر از پایگاه اطلاعاتی وب آو ساینس جمع‌آوری شد. پیشنهاد می‌شود در پژوهش‌های آتی از منابع پایگاه اطلاعاتی اسکوپوس نیز استفاده شود.

- شناسایی منابع موجود دیگر در حوزه علم اطلاعات و دانش‌شناسی.

از آنجایی که پژوهشگران در این پژوهش به متون کتاب‌ها، گزارش‌های علمی، منابع صوتی و غیره دسترسی نداشتند، فقط مقالاتی که در پایگاه استنادی وب آو ساینس نمایه شده‌اند مورد تجزیه و تحلیل قرار گرفتند. پیشنهاد می‌گردد در صورت دسترسی به متون منابع مختلف، متن‌کاوی روی آن‌ها نیز صورت گیرد.

- استفاده از روش یادگیری عمیق و سایر الگوریتم‌های موجود.

با توجه به اینکه روش به کار رفته در این پژوهش، فرایند متن‌کاوی بدون نظارت است، پیشنهاد می‌شود پژوهشگران از یادگیری ماشین نیز در راستای، کشف روند موضوعات حوزه‌های علمی مختلف بهره ببرند و نتایج را با روش‌های بدون نظارت مقایسه نمایند. همچنین از سایر الگوریتم‌های موجود برای استخراج موضوعات رأس استفاده کنند.

1. Aase
2. Silwattananusarn and Kulkanjanapiban
3. Baruni & Sathiaseelan
4. Wang et al

References

Aase, K. G. (2011). *Text mining of news articles for stock price predictions* (Master's thesis, Institutt for datateknikk informasjonsvitenskap).

Abol-sadegh, S. (2011). *Application of Text Mining in Reviewing Industrial Engineering Literature*. Master's Thesis, Industrial Engineering, Faculty of Engineering, Yazd University, Yazd, Iran. (In Persian)(

Aitchison, J., & Clarke, S. D. (2004). The thesaurus: a historical viewpoint, with a look to the future. *Cataloging & classification quarterly*, *37*(3-4), 5-21. Doi: 10.1300/J104v37n03_02

Baba-Aghaei, S. (2013). *Discovery of the Internal Structure of Positive Psychology Studies Using Text Mining*. Master's Thesis, Information Science and Knowledge Management, Faculty of Educational Sciences and Psychology, Allameh Tabataba'i University, Tehran, Iran. (In Persian)

Baruni, J. S., & Sathiaseelan, J. G. R. (2020). Keyphrase extraction from document using RAKE and TextRank algorithms. Int. J. Comput. Sci. Mob. Comput, 9(9), 83-93. Doi:10.47760/IJCSMC.2020.v09i09.009

De Jesus Holanda, A., Pisa, I. T., Kinouchi, O., Martinez, A. S., & Ruiz, E. E. S. (2004). Thesaurus as a complex network. *Physica A: Statistical Mechanics and its Applications*, *344*(3-4), 530-536. Doi: 10.1016/j.physa.2004.06.025

Ghanadinezhad, F., Osareh, F., & Ghane, M.R. (2023). Thematic analysis of scientific productions of Iranian researchers in the field of knowledge and information science with text mining approach. *Library and Information Sciences*, 26(2), 223-249. (In Persian) Doi:10.30481/lis.2021.298842.1862

Hassanzadeh, M., Zandian, F., Ahmadi Meinagh, S.S. (2018). Mapping the cognitive structure and its evolution in "Knowledge and Information Science": text mining approach (2004-2013). *Scientometric research journal*, 4(8), 123-142. (In Persian) Doi: 10.22070/rsci.2018.616

Kardan, A.A., & Kaihaninejad, M. (2012). Proposing a Model for Extracting Information from Textual Documents, Based on Text Mining in E-learning, *Journal of Information and Communication Technology*, 4(11), 47-54. (In Persian)

Kit, C., & Nie, J. Y. (2023). Information retrieval and text mining. In *Routledge Encyclopedia of Translation Technology* (pp. 601-642). Routledge.

Liu, W., Sun, Y., Yu, B., Wang, H., Peng, Q., Hou, M., & Liu, C. (2024). Automatic Text Summarization Method Based on Improved TextRank Algorithm and K-Means Clustering. Knowledge-Based Systems, 287(1), 111447. Doi: 10.1016/j.knosys.2024.111447

Pons-Porrata, A., Berlanga-Llavori, R., & Ruiz-Shulcloper, J. (2007). Topic discovery based on text mining techniques. *Information Processing & Management*, 43(3), 752-768. https://doi.org/10.1016/j.ipm.2006.06.001

Silwattananusarn, T., & Kulkanjanapiban, P. (2022). A text mining and topic modeling based bibliometric exploration of information science research. *IAES International Journal of Artificial Intelligence*, 11(3), 1057. DOI: http://doi.org/10.11591/ijai.v11.i3.pp1057-1065

Teimourpour, B. (2009). *Discovery of Emerging Trends in Scientific Fields Based on Dynamic Clustering Using Text Mining and Link Analysis*. Doctoral Dissertation, Information Technology in Industrial Engineering, Faculty of Engineering, Tarbiat Modares University, Tehran, Iran. (In Persian)

Wang, X., Xu, X., Zhang, J., Zhu, Y., Fan, Y., & Xu, P. (2021). Research on intelligent construction algorithm of subject knowledge thesaurus based on literature resources. In *Journal of Physics: Conference Series*. 1955(012038). IOP Publishing.

Yan, B. N., Lee, T. S., & Lee, T. P. (2015). Analysis of research papers on E-commerce: (2000–2013) based on a text mining approach. *Scientometrics*, 105(1), 403-417. Doi: 10.1007/s11192-015-1675-6