Overlap Analysis of Interface Elements in the Representation of Textual Documents: A Study Using the "Rapid Automatic Keyphrase Extraction Algorithm"

Yaghoub Norouzi[1], Elham Yalveh[2], Ashkan Khatir[3]

Abstract

Purpose: The present study investigated the degree of overlap of keywords extracted from interface elements in the representation of text documents using the "Rapid Automatic Keyphrase Extraction Algorithm."

Method: In this research, the "Rapid Automatic Keyphrase Extraction Algorithm" was used.

Keywords were extracted from a dataset including 500 scientific articles in five different subject groups. Then, the overlap between the keywords of the title, abstract, and keywords of the authors was examined.

Findings: The results showed that the overlap between title keywords and authors' keywords was about 45%, and the overlap between abstract keywords and authors' keywords was about 18%. Further, it was observed that the keywords of the title covered 22% of the keywords of the abstract. The results also showed that the overlap and dispersion between the keywords of the abstract and the keywords of the authors and between the keywords of the abstract and the keywords of the title were balanced and almost the same. However, it was observed that the keywords of the title and the keywords of the authors were more scattered, which indicates the possibility of more overlap between the keywords of the title and the keywords of the author of an article compared to the keywords of the abstract and the keywords of the author, as well as the keywords of the abstract and the keywords of the title. Is. In addition, there was a good understanding of the concepts and topics of the research field in the fields of psychology and public administration, while the need to improve and strengthen the knowledge of concepts was observed in the fields of information technology and public law. The amount of overlap between abstract keywords and authors' keywords in five subject groups was about 20%.

Conclusion: Appropriate use of keywords, writing abstracts with content in harmony with the topic and choosing suitable titles can help to improve the process of extracting concepts, storing and retrieving scientific articles, including that keywords, abstracts and titles can be used as input for algorithms for extracting concepts, as well As parts of the information storage structure, they can contribute significantly to the speed of users' access to the information they need and as input for information retrieval algorithms for quick access to related articles.

Article Type: Research Article
Article history:
Received: 2 Apr. 2024
Accepted: 26 May 2024

1. Professor, Knowledge and Information Science Group, University of Qom, Qom, Iran (Corresponding Author)
ynorouzi@gmail.com

2. PhD Candidate, Knowledge & Information Science, University of Qom, Qom, Iran
elham.yalveh2018@gmail.com

3. Assistant Professor, Information Technology, Iranian Research Institute for Information Science and Technology (IranDoc), Tehran, Iran
khatir@students.irandoc.ac.ir

Keywords

Text Documents, Keyword Extraction, Keyword Overlap, Document Representation, Data Dispersion

Citation: Norouzi, Y., Yalveh, E, & Khatir, A. (2025). Overlap Analysis of Interface Elements in the Representation of Textual Documents: A Study Using the "Rapid Automatic Keyphrase Extraction Algorithm". *Librarianship and Information Organization Studies*, 35(4): 95-122.
Doi: 10.30484/nastinfo.2024.3594.2276

Publisher: National Library and Archives of I.R. of Iran

تحلیل همپوشانی عناصر واسط در بازنمایی اسناد متنی: مطالعه‌ای به روش الگوریتم «RAKE»

یعقوب نوروزی[1] | الهام یلوه[2] | اشکان خطیر[3]

چکیده

هدف: پژوهش حاضر با هدف بررسی میزان همپوشانی کلیدواژه‌های استخراج‌شده از عناصر واسط در بازنمایی اسناد متنی با استفاده از الگوریتم «Rake» انجام شد.

روش: در این پژوهش، با استفاده از الگوریتم «Rake» کلیدواژه‌های مجموعه داده‌ای شامل ۵۰۰ مقاله علمی در پنج گروه موضوعی مختلف استخراج شد. سپس همپوشانی بین کلیدواژه‌های عنوان، چکیده و کلیدواژه‌های نویسندگان موردبررسی قرار گرفت.

یافته‌ها: نتایج نشان داد که همپوشانی بین کلیدواژه‌های عنوان و کلیدواژه‌های نویسندگان حدود ۴۵ درصد و همپوشانی بین کلیدواژه‌های چکیده و کلیدواژه‌های نویسندگان حدود ۱۸ درصد بود. در ادامه مشاهده شد که کلیدواژه‌های عنوان دارای پوشش ۲۲ درصدی کلیدواژه‌های چکیده بودند. نتایج همچنین نشان داد که همپوشانی و پراکندگی بین کلیدواژه‌های چکیده و کلیدواژه‌های نویسندگان و نیز بین کلیدواژه‌های چکیده و کلیدواژه‌های عنوان به‌طور متوازن و تقریباً یکسان بود؛ اما مشاهده شد که کلیدواژه‌های عنوان و کلیدواژه‌های نویسندگان دارای پراکندگی بیشتری بودند؛ که نشان‌دهندهٔ احتمال همپوشانی بیشتر بین کلیدواژه‌های عنوان و کلیدواژه‌های نویسندهٔ یک مقاله در مقایسه با کلیدواژه‌های چکیده و کلیدواژه‌های نویسنده و همچنین کلیدواژه‌های چکیده و کلیدواژه‌های عنوان است. بعلاوه درک خوبی از مفاهیم و مباحث حوزهٔ پژوهشی در رشته‌های روانشناسی و مدیریت دولتی وجود داشت، درحالی‌که در رشته‌های فناوری اطلاعات و حقوق عمومی نیاز به بهبود و تقویت درک مفاهیم مشاهده شد. میزان همپوشانی بین کلیدواژه‌های چکیده و کلیدواژه‌های نویسندگان در پنج گروه موضوعی حدود ۲۰ درصد بود.

نتیجه‌گیری: استفاده مناسب از کلیدواژه‌ها، نوشتن چکیده‌هایی با محتوای هماهنگ با موضوع موردنظر و انتخاب عناوین متناسب می‌تواند به بهبود فرایند استخراج مفاهیم، ذخیره‌سازی و بازیابی مقالات علمی کمک کند، ازجمله اینکه کلیدواژه‌ها، چکیده‌ها و عناوین می‌توانند به‌عنوان ورودی برای الگوریتم‌های استخراج مفاهیم، همچنین به‌عنوان بخش‌هایی از ساختار ذخیره‌سازی اطلاعات در سرعت دسترسی کاربران به اطلاعات موردنیازشان و به‌عنوان ورودی برای الگوریتم‌های بازیابی اطلاعات برای دسترسی سریع به مقالات مرتبط کمک بسزایی داشته باشند.

کلیدواژه‌ها

اسناد متنی، استخراج کلیدواژه، همپوشانی کلیدواژه، بازنمایی اسناد، پراکندگی داده‌ها

۱. استاد، گروه علم اطلاعات و دانش‌شناسی، دانشگاه قم، قم، ایران (نویسندهٔ مسئول)
ynorouzi@gmail.com

۲. دانشجوی دکتری، علم اطلاعات و دانش شناسی، دانشگاه قم، قم، ایران
elham.yalveh2018@gmail.com

۳. استادیار، فناوری اطلاعات، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، تهران، ایران
khatir@students.irandoc.ac.ir

فصلنامهٔ مطالعات کتابداری و سازماندهی اطلاعات، ۳۵ (٤)، زمستان ۱٤۰۳

نوع مقاله: پژوهشی
تاریخ دریافت: ۱۴۰۳/۰۱/۱۴
تاریخ پذیرش: ۱۴۰۳/۰۳/۰۶

ناشر: سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

استناد: نوروزی، یعقوب، یلوه، الهام و خطیر، اشکان (۱۴۰۳). تحلیل همپوشانی عناصر واسط در بازنمایی اسناد متنی: مطالعه‌ای به روش الگوریتم «RAKE». *مطالعات کتابداری و سازماندهی اطلاعات*، ۳۵(۴): ۹۵-۱۲۲.

Doi: 10.30484/nastinfo.2024.3594.2276

مقدمه

در عصر حاضر شاهد حجم زیادی از اطلاعات با رشد روزافزون و درعین‌حال پویا هستیم. بخش قابل‌توجهی از این اطلاعات در قالب داده‌های متنی در بستر وب و پایگاه داده‌های متنی قرارگرفته است. این‌گونه محتواها قالب‌هایی همچون نوشتارهای الکترونیکی، مقالات پژوهشی و خبری، پایان‌نامه‌ها، کتاب‌ها، وب‌نگاشت‌ها، گزارش‌ها و نظایر آن را در بر می‌گیرد. درعین‌حال، سرعت رشد این دسته از اطلاعات نسبت به سایر قالب‌های اطلاعاتی همچنان بیش‌ازپیش در حال افزایش است. در دنیای واقعی، مجموعه داده‌ها که نیاز به پردازش دارند، همیشه استاندارد نیستند (یلوه و همکاران، ۱۴۰۰؛ Song et al., 2014) به همین دلیل در صورت نیاز به یک منبع اطلاعاتی، با در نظر گرفتن این احتمال که حداکثر در ۵۰ نتیجه اول مطالب به‌طور کامل مرتبط باشند، فرایند زمان‌بر و وقت‌گیر خواهد بود (کریمی منش، ۱۳۹۲) از طرفی در بعضی موارد نتایج جستجو ازنظر معنا و مفهوم ارتباط چندانی با کلمات مورد هدف برای جستجو ندارند. به همین دلیل ارائه روش‌هایی برای غلبه برای مشکلات می‌تواند راهگشا باشد. این امر به کاربر کمک بسزایی خواهد کرد که با صرف هزینه زمانی کمتر، مرتبط‌ترین و مناسب‌ترین مطلب را به دست آورد.

در فرایند جستجو اولین قدم انتخاب کلیدواژه مناسب و تا حد ممکن مرتبط است که نماینده محتوای سند مورد جستجو باشد. کلیدواژه‌ها زیرمجموعه‌ای از کلمات یا عبارات یک سند هستند که می‌توانند به‌طور خلاصه معرف محتوای یک سند باشند و همچنین توصیفی از محتوای سند را ارائه کنند و نقش بسزایی در درک دقیق و سریعی از محتوا دارند (; Rose et al., 2010; Zhang, 2008;Subramanian & Karthik, 2017). استخراج

خودکار کلیدواژه[1] (AKE) با شناسایی و استخراج مجموعه کوچکی از کلمات، عبارات کلیدی، کلمات کلیدی، یا بخش‌های کلیدی از سند می‌تواند در توصیف محتوای سند تأثیرگذار باشد (Hulth, 2003). ازآنجاکه کلیدواژه کوچک‌ترین واحد اطلاعاتی است که می‌تواند محتوای سند را بازگو کند، بسیاری از داده‌کاوی‌های متنی می‌توانند از آن بهره ببرند، مانند نمایه‌سازی خودکار[2]، خلاصه‌سازی خودکار[3]، طبقه‌بندی خودکار[4]، خوشه‌بندی خودکار[5]، فیلترینگ خودکار[6]، تشخیص و ردیابی موضوع[7]، تصویرسازی اطلاعات[8] و غیره. در این راستا استخراج کلیدواژه‌ها نقش تعیین‌کننده‌ای در زمینه‌های خلاصه‌سازی متون، برچسب‌گذاری اسناد، بازیابی اطلاعات و استخراج موضوع از متن دارد (Liu, 2020; Zhang, 2008)؛ بنابراین، می‌توان با تجمیع کلمات، غنی‌سازی کلیدواژه‌ها، انتخاب کلیدواژه‌ها و توصیفگرهای مناسب سیستمی، بازیابی مناسب‌تری در اختیار داشت (خطیر و گنجه‌فر، ۱۳۹۷).

ازجمله راه‌های دسترس‌پذیری اطلاعات ذخیره و سازماندهی آن است. شیوهٔ ارائه محتوای موضوعی مدارک از مهم‌ترین عواملی است که بر دسترسی موضوعی مؤثر است. محتوای موضوعی مدارک ممکن است توسط کلیدواژه‌هایی از عنوان، متن، چکیده و یا با استفاده از واژگان کنترل‌شده نمایش داده شود (انصاری، ۱۳۷۹). مخاطب در صورت مطالعه یک چکیده، مناسب و یا نامناسب بودن آن را با هدف مورد جستجوی خود، درخواهد یافت. علاوه بر این، عنوان، چکیده و کلیدواژه‌های هر اثر نقش مؤثری در دستیابی به آن در زنجیره اطلاعات دارند و هر یک به‌نوعی در شناساندن آن نقش تأثیرگذاری دارند. از سوی دیگر ازجمله مؤلفه‌های موردنیاز در نظام ذخیره و بازیابی خودکار اطلاعات، امکان دستیابی سریع و مستمر به یک سند توسط کاربر است؛ بنابراین، باید شیوه‌ها و راهکارهایی برگزیده و یا مورد کنکاش قرار گیرد که بتواند این امکان را فراهم سازد. شناسایی کلیدواژه‌ها از متن با روش‌های

1. Automatic Keyword Extraction (AKE)
2. Automatic Indexing
3. Automatic Summarization
4. Automatic Classification
5. Automatic Clustering
6. Automatic Filtering
7. Topic Detection and Tracking
8. Information Visualization

معمول کاری زمان‌بر و پرهزینه است. استخراج خودکار کلیدواژه‌ها باعث صرفه‌جویی در زمان می‌شود و همچنین می‌تواند با موردتوجه قرار دادن دقت در روش‌های استخراج خودکار در کاهش خطاهای انسانی نیز مؤثر واقع شود (Baruni & Sathiaseelan, 2020). لذا در این پژوهش آنچه به‌عنوان مسئله پژوهشی مورد هدف قرارگرفته شده است تا حد امکان غلبه بر دو چالش زمان‌بر بودن جستجو در داده‌های متنی و همچنین نامرتبط بودن نتایج بازیابی آن‌هاست. برای رسیدن به این مهم استخراج ماشینی کلیدواژه‌ها با استفاده از الگوریتم مناسب نقش بسزایی دارد. الگوریتم «Rake[1]» ازجمله الگوریتم‌های مناسب جهت استخراج خودکار کلیدواژه‌ها است. ازآنجایی‌که در این پژوهش داده‌های مورداستفاده به زبان فارسی هستند، باید به این نکته نیز توجه داشت که ساختار و قواعد زبان فارسی نسبت به زبان انگلیسی پیچیده‌تر است. این الگوریتم می‌تواند مستقل از زبان در رسیدن به نتایج مطلوب تأثیرگذار باشد. در این رابطه ازجمله راهکارها، مقایسه و تطابق بین کلیدواژه‌هایی که توسط نویسنده یک اثر به‌صورت دستی به آن داده‌شده با کلیدواژه‌هایی است که به‌صورت خودکار از ابزارهای واسط در متن ازجمله چکیده و عنوان استخراج می‌شود؛ بنابراین توجه به این مقوله و اهمیت آن موجب شد تا پژوهشی در این رابطه موردتوجه قرار گیرد تا بتواند گامی را در این جهت برای بهبود عملکرد ذخیره و بازیابی اطلاعات داشته باشد. این پژوهش با هدف استخراج ماشینی کلیدواژه‌ها با استفاده از الگوریتم استخراج کلیدواژه Rake از عناوین مقالات و چکیده‌های مقالات و بررسی هم‌پوشانی کلیدواژه‌های استخراج‌شده با یکدیگر و با کلیدواژه‌های نویسندگان انجام شده است.

پیشینهٔ پژوهش

برای یافتن آثار علمی منتشرشده خارجی درزمینهٔ موضوعی این پژوهش جستجوی منابع با کلیدواژه‌های Keyword، Indexing، Descriptor، overlapping keywords، Distribution، Extracting keyword، matching keyword در پایگاه‌های IEEE، Springer، Science Direct، ACM Digital Library، Scopus، ProQuest، و google

[1]. Rapid Automatic Keyphrase Extraction

scholar انجام شد. همچنین برای جستجو منابع داخلی نیز از کلیدواژه‌های همپوشانی[1] کلیدواژه‌ها، توصیفگرها، نمایه‌سازی، توزیع کلیدواژه، استخراج کلیدواژه، انطباق کلیدواژه در پایگاه‌های اطلاعاتی الکترونیکی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، پرتال جامع علوم انسانی و مرجع دانش (سیویلیکا) جستجو صورت گرفت. همچنین از موتور جستجوی گوگل و علم نت نیز استفاده شد. برخی از نتایج حاصل به شرح زیر است. البته با توجه به نوع مسئله، تعداد این‌گونه پژوهش‌ها اندک است و همین مسئله اهمیت پرداخت به آن را دوچندان می‌کند.

در پژوهشی که توسط دانش و رحیمی (۱۴۰۲) انجام شد، با استفاده روش متن‌کاوی و الگوریتم‌ها و فنون مربوط به آن و همچنین طبقه‌بندی متون با رویکرد تحلیلی-تطبیقی بر روی چکیده و عنوان انتشارات COVID-19 نمایه شده در پایگاه PubMed Central® (PMC) موردبررسی قرار گرفت. تحلیل داده‌ها توسط آن‌ها نشان داد که covid، infect و cell از مهم‌ترین واژگان بکار رفته در انتشارات بین‌المللی COVID-۱۹ و patient، SARS-Cov و covid مهم‌ترین واژگان انتشارات ملی هستند. آن‌ها در پژوهش خود به این نتیجه رسیدند که در خصوص روند تغییرات واژگان مورداستفاده در انتشارات COVID-۱۹ تفاوت اساسی بین مهم‌ترین واژه‌های انتشارات بین‌المللی با ملی و تأکید پژوهش‌های بین‌الملل بر کرونا و عفونت ناشی از آن و در سطح ملی بر بیماران و کرونا است. نتیجه مهم دیگر تغییرات سالانه به وجود آمده در واژه‌ها در سطح انتشارات ملی و بین‌المللی بود. محرابی و همکاران (۱۴۰۰) در پژوهشی روشی برای بهبود استخراج کلمات کلیدی از متن فارسی بر پایه الگوریتم RAKE ارائه دادند که یک الگوریتم استخراج خودکار کلیدواژه محسوب می‌شود. آن‌ها برای بررسی نقاط ضعف الگوریتم و ارائه راهکار پیشنهادی از مجموعه‌ای از فرادادههای پایان‌نامه و رساله‌های فارسی استفاده کردند. راهکار پیشنهادی در آزمایش و ارزیابی صورت گرفته بر روی این داده‌ها افزایش دقت، بازخوانی و معیار F را نشان داد. داورپناه (۱۳۷۵) در پژوهشی میزان همخوانی عناوین مقالات فارسی با محتوای آن‌ها را در زمینه‌های مختلف علمی موردبررسی قرار داد. نتایج حاصل از پژوهش وی نشان داد که عناوین مقالات در حوزه

1. Overlap

علوم انسانی در مقایسه با علوم پایه، مهندسی و علوم پزشکی سازگاری کمتری با محتوا دارند. بنی اقبال و همکاران (۱۳۹۰) به مقایسه واژه‌های عنوان و چکیده پایان‌نامه‌ها با توصیفگرهای تعیین‌شده در نمایه سازمان اسناد و کتابخانه ملی ایران پرداختند. آن‌ها در پژوهش خود به این نتیجه رسیدند که میزان مطابقت واژه‌های عنوان با توصیفگرهای موجود در نمایه سازمان، ۴۷ درصد و میزان مطابقت واژه‌های چکیده با توصیفگرها ۵۳/۵ درصد است. خطیر و گنجه‌فر (۱۳۹۷) در پژوهشی به تحلیل توزیع و تمرکز کلیدواژه‌های ۵۲۷ پایان‌نامه و رساله در سه رشته مهندسی عمران، مکانیک، برق و میزان تطابق با توصیفگرها، عنوان و چکیده پرداختند. آن‌ها علاوه بر بررسی معایب موجود در انتخاب نمایه و نگارش چکیده، اطلاعاتی در رابطه با نحوه انتخاب کلیدواژه‌ها در جهت بهره‌گیری در طراحی سیستمی برای استخراج خودکار کلیدواژه‌ها ارائه کردند. نتایج حاصل از پژوهش آن‌ها نشان داد که عموماً نمایه‌های انتخاب‌شده بیش از ۶۰ درصد توسط نویسنده و نمایه‌ساز حرفه‌ای از ۴۰ درصد ابتدایی چکیده انتخاب شده‌اند. دیگر تحلیل‌های آماری این پژوهش نشان داد که میزان انطباق بین توصیفگرها و کلیدواژه‌ها ۸ درصد اختلاف دارد که نشان‌دهنده میزان تفاوت نظر زیاد بین نویسندگان پارساها و نمایه‌سازان است. درزی خلردی و رضوی (۱۳۹۷) میزان همخوانی واژگان کلیدی مقاله‌های مجلات منتشرشده با درجه علمی-پژوهشی در دانشگاه علوم کشاورزی و منابع طبیعی ساری با اصطلاح‌نامه کب[1] موردبررسی قرار دادند. نتایج حاصل از پژوهش آن‌ها نشان داد کمتر از نیمی از کلیدواژه‌ها دارای همخوانی کامل هستند. این مسئله نشانگر عدم هماهنگی میان اصطلاحات نویسندگان، اصطلاحات موجود در اصطلاح‌نامه‌ها و جست‌وجوگران اطلاعات است. همچنین کلیدواژه‌های مقالات مجله‌ها ازلحاظ قواعد نگارشی و مفرد و جمع بودن نیاز به یکدست شدن دارند تا همخوانی بیشتری با اصطلاح‌نامه داشته باشند. قاضی میر سعید و مسعودی (۱۳۹۸) در پژوهشی به‌منظور شناسایی سطح دانش مؤلفان دندانپزشکی در استفاده از ابزار MeSH[2] در پژوهش‌های خود به مقایسه کلیدواژه‌های مقالات منتشرشده در مجلات دندانپزشکی ایرانی نمایه شده در PubMed پرداختند. آن‌ها کلیدواژه‌ها را بر اساس میزان انطباقشان با توصیفگرهای پذیرفته‌شده MeSH در سه گروه انطباق دقیق،

1. CAB Thesaurus
2. Medical Subject Headings

انطباق نسبی و نامنطبق طبقه‌بندی کردند که ۴۴/۳ درصد انطباق دقیق، ۱۴ درصد انطباق نسبی و ۴۱/۷ درصد نامنطبق بود. همچنین آن‌ها در سال ۲۰۱۶ نیز پژوهشی بر روی کلیدواژه‌های مجله علوم پیراپزشکی با MeSH انجام دادند که از مجموع ۱۱۴۳ کلیدواژه مندرج در ۲۶۹ مقاله منتشرشده این مجله، ۲۴/۲ درصد انطباق دقیق، ۳۵/۸ درصد انطباق نسبی و ۴۰ درصد نامنطبق بودند. انصاری و همکاران (۱۴۰۰) پژوهشی را با هدف تعیین میزان همخوانی کلیدواژه‌های انگلیسی پایان‌نامه‌های دانشکده‌های پرستاری و مامایی دانشگاه‌های علوم پزشکی تهران، ایران و شهید بهشتی با سر عنوان‌های موضوعی پزشکی مش در بازه زمانی ۵ ساله (۱۳۹۹-۱۳۹۵) انجام دادند. نتایج بررسی آن‌ها نشان داد که در طی سال‌های موردبررسی ۳۸/۵ درصد واژه‌ها با سر عنوان‌های موضوعی مش همخوانی کامل، ۴۲/۱ درصد عدم همخوانی و ۱۹/۴ درصد با همخوانی نسبی دارند. آن‌ها در بررسی یافته‌های خود به این نتیجه رسیدند که نیاز است دانشجویان و پژوهشگران دقت بیشتری را برای انتخاب کلیدواژه‌های خود داشته باشند که این امر را منوط بر نیاز به آموزش و حساس‌سازی آنان در انتخاب کلیدواژه مناسب دانستند. گیل لیوا و آلونسو آرویو[1] (۲۰۰۷) کلیدواژه‌های داده‌شده توسط نویسندگان مقالات علمی در پایگاه داده توصیفی را بررسی کردند. نتایج حاصل از بررسی آن‌ها نشان داد بیش از ۴۶ درصد کلیدواژه‌های مورداستفاده نویسندگان در حوزه‌های فیزیکی و مهندسی، کشاورزی و علوم و فناوری اطلاعات با توصیفگرهای موجود در پایگاه‌های اطلاعاتی مربوط تطابق دارد. نئول و همکاران[2] (۲۰۱۰) در پژوهشی کلیدواژه‌های نویسندگان در مقالات مجلات پزشکی را موردبررسی قرار دادند. آن‌ها با بررسی حدود ۳۰۰ کلیدواژه به این نتیجه رسیدند که بیش از ۶۰ درصد کلیدواژه‌های اختصاص داده‌شده توسط مؤلفان با اصطلاحات به‌کاررفته در نمایه مربوطه دارای همخوانی است. در پژوهشی که کیپ[3] (۲۰۱۱) انجام داد، نمایه‌سازی آنلاین از دیدگاه سه گروه خوانندگان، نویسندگان و نمایه‌سازان حرفه‌ای موردبررسی قرار گرفت. نتایج وی نشان داد میزان انطباق بین کلیدواژه نویسنده و برچسب‌های خوانندگان مقاله و کلیدواژه‌ها ۳۳ درصد، میزان انطباق بین توصیفگرها و برچسب‌ها ۱۶ درصد و میزان انطباق بین

1. Gil- Leiva & Alonso- Arroyo
2. Névéol et al.
3. Kipp

کلیدواژه‌ها و توصیفگرها ۱۹ درصد است. در پژوهشی که توسط کیم و همکاران[1] (۲۰۱۶) بر روی کلیدواژه‌های مقالات حوزه چشم‌پزشکی انجام شد، از مجموع ۱۹۵۲ کلیدواژه مندرج در ۴۰۹ مقاله مجله انجمن بینایی و چشم‌پزشکی کره که با MeSH مقایسه شدند، ۲۲/۴ درصد انطباق دقیق، ۴۱/۸ درصد انطباق نسبی، ۳۵/۵ درصد نامنطبق و ۰/۳ درصد هم با کلیدواژه‌های اشتباه نگارش یافته به دست آمد. پارسایی محمدی و همکاران[2] (۲۰۱۷) در پژوهشی با هدف بررسی کلیدواژه‌های مقالات نمایه شده در IranMedex از نظر منشأ، ساختار و نمایه‌سازی و انطباق آن‌ها با فرهنگ اصطلاحات پزشکی فارسی[3] و سر عنوان‌های موضوعی پزشکی[4] به این یافته رسیدند. اگرچه ازنظر منشأ نمایه‌سازی تفاوت معنی‌داری بین نسبت انواع مختلف واژگان کلیدی فارسی و انگلیسی نمایه شده در IranMedex وجود ندارد، اما میزان انطباق واژگان فارسی و انگلیسی با اصطلاح‌نامه پزشکی فارسی و سر عنوان‌های موضوعی پزشکی در سال‌های مختلف متفاوت است.

مرور پیشینه‌های داخلی و خارجی مرتبط با پژوهش حاضر نشان می‌دهد که انتخاب یک کلیدواژه مناسب درزمینهٔ ذخیره و بازیابی اطلاعات در میان منابع مکتوب در رشته‌های متفاوتی ازجمله پزشکی، مهندسی، کشاورزی و سایر علوم به‌عنوان یک دغدغه توسط پژوهشگران موردبررسی قرارگرفته است؛ و این نشان از تلاش برای به حداقل رساندن گزینش‌های ناصحیح در به‌کارگیری کلیدواژه‌های نامناسب و غیر مرتبط به یک سند متنی است. همچنین با توجه به متغیر بودن نتایج پژوهش‌های آن‌ها می‌توان گفت که با ادامه پژوهش‌هایی ازاین‌دست همچنان می‌توان امیدوار بود که با پژوهش‌های دقیق‌تر به نتایج مناسبی در جهت ذخیره و بازیابی این دسته از منابع رسید. بررسی‌های صورت گرفته بر روی پژوهش‌های مشابه در داخل و خارج از کشور که به مواردی از آن اشاره شد حاکی از آن است که روش غالب برای استخراج کلیدواژه‌ها و بررسی آن‌ها در این پژوهش‌ها روش‌های آماری است؛ اما پژوهش حاضر با استفاده از الگوریتم «Rake» جهت استخراج کلیدواژه‌ها در جهت

1. Kim et al.
2. Parsaei-Mohammadi et al.
3. Persian medical thesaurus
4. Medical Subject Headings

بررسی میزان همپوشانی کلیدواژه‌های استخراج‌شده با کلیدواژه‌های مقالات موردبررسی، سعی بر رسیدن به پاسخ مناسب پرسش‌های زیر برای هدف موردپژوهش خود است:

۱) همپوشانی بین کلیدواژه‌های عنوان با کلیدواژه‌های نویسندگان به چه میزان است؟

۲) همپوشانی بین کلیدواژه‌های چکیده با کلیدواژه‌های نویسندگان به چه میزان است؟

۳) همپوشانی بین کلیدواژه‌های چکیده با کلیدواژه‌های عنوان به چه میزان است؟

۴) در کدام‌یک از رشته‌های موردبررسی در پژوهش حاضر انتخاب کلیدواژه توسط نویسنده با همپوشانی بیشتری انجام شده است؟

روش پژوهش

در پژوهش حاضر برای استخراج خودکار کلیدواژه‌ها، از الگوریتم «Rake» بهبودیافته استفاده شده است. همان‌گونه که محرابی و همکاران (۱۴۰۰) نیز بیان کرده‌اند، الگوریتم اصلی «Rake» که توسط رز و همکاران[1] (۲۰۱۰)، جهت استخراج خودکار کلیدواژه در زبان انگلیسی معرفی شده است مستقل از منبع اسناد و زبان متون است؛ بنابراین، با توجه به این مزیت و با توجه به اینکه پژوهش حاضر بر روی مجموعه داده‌های فارسی انجام شده است، از الگوریتم بهبودیافته آن جهت استخراج خودکار کلیدواژه‌ها بهره گرفته شد. مراحل پیش‌پردازش با توجه به الگوریتم «Rake» بهبودیافته شامل دو مرحله اصلی پیش‌پردازش متن و پردازش متن است که مراحل پیش‌پردازش آن شامل نرمال‌سازی[2]، تک‌واژسازی[3]، برچسب‌گذاری نحوی[4]، قطعه‌بندی[5]، حذف کلمات توقف[6] و علائم نگارشی و مراحل پردازش آن شامل دسته‌بندی کلمات و تشکیل عبارات کاندید، محاسبهٔ امتیاز، درجه و نمره عبارات کاندید، مرتب کردن عبارات بر اساس امتیاز آن‌ها، انتخاب عبارات کلیدی، استخراج عبارات کلیدی منتخب و تبدیل عبارات کلیدی منتخب به کلمات کلیدی است.

---

1. Rose et al.
2. Normalization
3. Tokenization
4. Part-of-speech tagging
5. Segmentation
6. Stop Words

در این پژوهش برای استخراج داده‌ها از عنوان، چکیده و کلیدواژه‌های نویسندگان[1] استفاده شد که در پژوهش‌های قبلی ازجمله خطیر و گنجه‌فر (۱۳۹۷) نیز موردتوجه قرار گرفته است. همچنین نتایج حاصل از مرحله پیش‌پردازش پژوهش یلوه و همکاران (۱۴۰۲) حاکی از آن است که این الگوریتم برای داده‌های متنی فارسی مناسبت بیشتری دارد. بر اساس الگوریتم «Rake» بهبودیافته پس از محاسبهٔ امتیازهای مربوط به هر عبارت که شامل اسم‌ها و صفات می‌شد، درنهایت پس از مرتب کردن امتیازها به‌صورت نزولی عباراتی به‌عنوان عبارات کلیدی انتخاب شد که به ازای ۲۰ درصد بیشترین امتیازها دارای بیشترین امتیاز محاسبه شده بود. با توجه به اینکه الگوریتم «Rake» عبارات کلیدی را استخراج می‌کند، در ادامه عبارات کلیدی منتخب به کلمات کلیدی منتخب تبدیل شد و میزان همپوشانی با توجه به هدف پژوهش بررسی شد.

همچنین جهت ساختارمندکردن داده‌ها، ابتدا عملیات پیش‌پردازش انجام شد که طی مراحلی ازجمله نرمال‌سازی، تک‌واژسازی، برچسب‌گذاری نحوی، قطعه‌بندی، حذف کلمات توقف و علائم نگارشی صورت گرفت. سپس ویژگی‌های (کلمات) زائد و نامرتبط حذف شد. برای پیش‌پردازش متون فارسی از کتابخانه متن‌باز هضم که یکی از بسته‌های زبان «پایتون» است استفاده شد. ابزار مورداستفاده برای تجزیه‌وتحلیل جهت مراحل پیش‌پردازش، پردازش و ارزیابی، زبان برنامه‌نویسی «پایتون» بود. مجموعه داده‌ای که برای این پژوهش در نظر گرفته شد مربوط به مقاله‌های علمی در حوزه‌های مختلف در قالب فایل «اکسل» بود. به‌منظور ایجاد یکدستی در داده‌ها و همچنین مقایسه بین رشته‌های علمی از روش گزینشی بر پایه یک حد آستانه که ۱۰۰ پژوهش در هر رشته را در برمی‌گرفت برای پنج رشته انتخابی استفاده شد. درنهایت ۵۰۰ مقاله به‌عنوان مجموع داده در پژوهش حاضر در پنج گروه روانشناسی، فناوری اطلاعات و ارتباطات، معماری، مدیریت دولتی و حقوق عمومی که هر گروه شامل ۱۰۰ عنوان مقاله علمی می‌شد، انتخاب شد. پردازش و تجزیه‌وتحلیل داده‌ها نیز بر روی سه فیلد چکیده، عنوان و کلیدواژه‌های نویسندگان انجام شد.

در مرحله پیش‌پردازش با توجه به الگوریتم «Rake» بهبودیافته جهت پیش‌پردازش

[1]. Author

داده‌ها فرایند نرمال‌سازی انجام شد که طی آن بدون اینکه شاهد تغییر معنایی در متن باشیم علائم نگارشی، حروف، فاصله‌های بین کلمات، اختصارات و غیره به شکل استاندارد تبدیل شدند. سپس تک‌واژسازی که تقسیم یک متن به کلمات، عبارات و یا دیگر قسمت‌های معنی‌دار است انجام شد. با استفاده از روش برچسب‌گذاری نحوی به کلمات، نقش کلمه در جمله مانند اسم، فعل، حرف‌اضافه، صفت، قید و غیره تعیین شد. در فرایند قطعه‌بندی کلمات دارای بار معنایی ازجمله اسامی و صفات از تجزیه ساختار گرامری به دست آمد. در انتها، حذف کلمات توقف و علائم نگارشی انجام شد. کلمات توقف کلماتی هستند که معمولاً بدون وابستگی به یک موضوع خاص در متن هستند و به‌تنهایی دارای بار معنایی نیستند (مانند حروف ربط، حروف اضافه و غیره) که برای انجام این مرحله در پژوهش از فهرست کلمات توقف دریکی از پروژه‌های «گیت هاب»[1] (Kharazi, ۲۰۱۵) استفاده شد؛ و در ادامه این فرایند ویژگی‌های (کلمات) زائد و نامرتبط حذف شدند. سپس پس از مرحله پردازش، بردار کلمات حاصل از عبارات منتخب استخراج‌شده توسط الگوریتم «Rake» جهت بررسی میزان همپوشانی تشکیل شد. در ادامه با ایجاد ماتریس سند-کلمه[2] نمایش اسناد در یک روش یکپارچه و آماده‌سازی آن‌ها برای تحلیل فراهم شد. در این ماتریس هر سلول محتوای مربوط به آن سند را با توجه به انتخاب روش تشکیل این ماتریس نشان می‌دهد. به‌این‌ترتیب، در این مرحله نیز ماتریس سند-کلمه تشکیل شد.

الگوریتم Rake بهبودیافته

این الگوریتم بر مبنای سه راهکار، نرمال‌سازی درجه عبارات نسبت به طول آن‌ها، وزن‌دهی به عبارات تکراری و استفاده از مفهوم انحراف استاندارد، جهت استخراج کلمات از اسناد فارسی بهینه‌شده است.

نرمال‌سازی درجه عبارات نسبت به طول آن‌ها: در این الگوریتم نرمال‌سازی درجه عبارات نسبت به طول آن‌ها بر اساس روابط محاسبه می‌شود.

1. GitHub
2. Document-term matrix (DTM)

$$d\acute{e}g_w(w_j) = \frac{deg_w(w_j)}{length(p_i \cdot w_j)}$$

که در این رابطه w_j نشان‌دهنده کلمه و p_i نشان‌دهنده عبارت است.

$$d\acute{e}g_w(p_i) = \sum_{\substack{j=1 \\ p_i \in p_w}}^{N} \frac{deg_w(w_j)}{length(p_i \cdot w_j)}$$

$$Sc\acute{o}re_w(w_j) = \frac{d\acute{e}g_w(w_j)}{freq(w_j)}$$

که در آن $d\acute{e}g_w(w_j)$ درجه جدید برای کلمه w_j، و $Sc\acute{o}re_w(w_j)$ امتیاز جدید برای کلمه w_j است. در انتها امتیاز هر عبارت بر پایه درجه جدید طبق رابطه زیر محاسبه می‌شود:

$$Sc\acute{o}re_{\ RAKE}(P_i) = \sum_{j=1}^{n} Sc\acute{o}re_w\ (w_j)$$

وزن‌دهی به عبارات تکراری: در این الگوریتم به‌جای محاسبه امتیاز عبارت بر پایه الگوریتم RAKE فراوانی عبارت در نظر گرفته شده است که بر طبق رابطه زیر محاسبه می‌شود:

$$Score_{new}(p_i) = Sc\acute{o}re_{\ RAKE}(p_i) \times freq(p_i)$$

استفاده از مفهوم انحراف استاندارد: در این الگوریتم به‌جای قرار دادن امتیاز کلمات که حاصل تقسیم تعداد کلمات بر درجه آن‌ها است، انحراف استاندارد کلمات از میانگین امتیازهایشان قرار داده می‌شود. برای یک نمونه N عضو به‌صورت $x_1 x_2 x_3 \cdots x_N$ با میانگین μ، انحراف استاندارد هر عضو از میانگین برابر است با رابطه:

$$SD_i(x) = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

امتیاز نهایی هر عبارت در این الگوریتم طبق رابطه زیر محاسبه می‌شود:

$$Score_{proposed}(p_i) = \sum_{j=1}^{N} SD_i\left(Scóre_{\ RAKE}(w_j)\right) \times freq(p_i)$$

که در آن عبارت p_i دارای کلمات $w_1 w_2 w_3 \cdots w_N$ است که در آن نرمال‌سازی طول عبارات با استفاده از امتیاز جدید برای کلمات با $\left(Scóre_{\ RAKE}(w_j)\right)$، فراوانی عبارت با $freq(p_i)$، و انحراف استاندارد امتیاز کلمات با $\sum_{j=1}^{N} SD_i\left(Scóre_{\ RAKE}(w_j)\right)$ در نظر گرفته می‌شود.

در پژوهش حاضر، با توجه به الگوریتم Rake بهبودیافته، فرایند صورت گرفته شامل دو مرحله اصلی پیش‌پردازش متن و پردازش متن است که مراحل پیش‌پردازش آن شامل:

۱) نرمال‌سازی متن؛
۲) تک‌واژسازی؛
۳) برچسب‌گذاری نحوی؛
۴) قطعه‌بندی؛
۵) حذف کلمات توقف و علائم نگارشی.

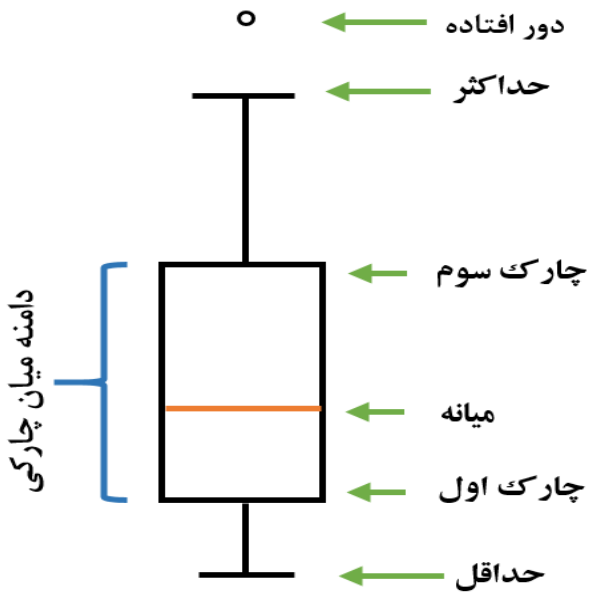
و مراحل پردازش آن شامل:

۱) دسته‌بندی کلمات و تشکیل عبارات کاندید؛
۲) محاسبه امتیاز، درجه و نمره عبارات کاندید؛
۳) مرتب کردن عبارات بر اساس امتیاز آن‌ها؛
۴) انتخاب عبارات کلیدی؛
۵) استخراج عبارات کلیدی منتخب؛
۶) تبدیل عبارات کلیدی منتخب به کلمات کلیدی؛
۷) پالایه کردن کلمات کم‌اهمیت جهت کاهش ابعاد بالای ویژگی و استخراج کلمات یکتا.

یافته‌ها

در این بخش با تمرکز بر یافته‌های پژوهش بر اساس پرسش‌های پژوهش نتایج به‌دست‌آمده

ارائه می‌شود. در این پژوهش برای بررسی و تحلیل داده‌ها از نمودار جعبه‌ای[1] که یک ابزار و روش استاندارد برای تفسیر داده‌ها و نمایش توزیع آن‌ها در مقالات علمی و پژوهشی است، استفاده شد. این نمودار در شکل ۱ قابل‌مشاهده است.



شکل ۱- نمودار جعبه‌ای

با استفاده از این نمودار می‌توان توزیع داده‌ها را بررسی کرد و مقایسه‌ای آسان بین چندین مجموعه داده داشت و همچنین داده‌های پرت را مشاهده کرد. این نمودار با استفاده از خطوط و جعبه‌هایی که نشان‌دهنده مقادیر آماری است، می‌تواند اطلاعات مهمی درباره توزیع و پراکندگی داده‌ها ارائه دهد. چندین شاخص مرکزی و پراکندگی آماری در این نمودار است که عبارت‌اند از:

میانه[2]: این شاخص در وسط داده‌ها قرار دارد. این قسمت در نمودار به‌وسیله خطی نارنجی‌رنگ دیده می‌شود.

1. Boxplot
2. Median

چارک اول[1] (Q1): مقداری را نشان می‌دهد که ۲۵ درصد داده‌ها از آن کوچک‌تر هستند. از سوی دیگر می‌توان این مقدار را میانه داده‌هایی دانست که بین کوچک‌ترین مقدار (با توجه به داده‌های پرت) و میانه قرار گرفته‌اند.

چارک سوم[2] (Q3): مقداری توسط این شاخص نشان داده می‌شود که ۷۵ درصد داده‌ها از آن کوچک‌تر هستند. به‌عبارت‌دیگر می‌توان این مقدار را میانه داده‌هایی دانست که بین بزرگ‌ترین مقدار (با توجه به داده‌های پرت) و میانه قرار گرفته‌اند.

دامنه میان چارکی[3] (IQR): توسط این شاخص فاصله بین چارک اول و سوم نشان داده می‌شود.

خطوط[4]: فاصله بین چارک اول تا کمترین مقدار و همچنین بیشترین مقدار توسط این خطوط پر می‌شود.

حداکثر[5] (Max): بزرگ‌ترین مقدار در این نمودار، بیشترین مقداری است که حداکثر ۱/۵ دامنه میان چارکی از چارک سوم فاصله دارد.

حداقل[6] (Min): کوچک‌ترین مقداری که این نمودار نشان می‌دهد، کمترین مقداری است که حداکثر ۱/۵ برابر دامنه میان چارکی از چارک اول فاصله دارد.

داده‌های دورافتاده[7] (پرت): در این نمودار داده‌هایی دورافتاده محسوب می‌شوند که از Minimumکوچک‌تر و یا ازMaximum بزرگ‌تر هستند، داده پرت محسوب می‌شوند.

پرسش اول: همپوشانی بین کلیدواژه‌های عنوان با کلیدواژه‌های نویسندگان به چه میزان است؟

بر اساس نتایج به‌دست‌آمده که در جدول ۱ قابل‌مشاهده است، با توجه به معیار تمرکز میانه،

1. First Quartile
2. Third Quartile
3. Interquartile Range
4. Whikers
5. Maximum
6. Minimum
7. Outlier

میزان همپوشانی بین کلیدواژه‌های عنوان با کلیدواژه‌هایی که نویسنده برای پژوهش خود برگزیده است ۴۵ درصد همپوشانی داشته است. همچنین بر اساس چارک اول ۲۵ درصد از مقاله‌ها کمتر از ۳۶ درصد اشتراک کلیدواژه بین کلیدواژه‌های عنوان و کلیدواژه‌های نویسندگان را دارند و درعین‌حال ۷۵ درصد از مقاله‌ها بیش از ۳۶ درصد اشتراک کلیدواژه بین کلیدواژه‌های عنوان و کلیدواژه‌های نویسندگان را دارد؛ و در ادامه همان‌گونه که در چارک سوم قابل‌مشاهده است، ۷۵ درصد مقالات دارای کمتر از ۶۴ درصد اشتراک کلیدواژه بین کلیدواژه‌های عنوان و کلیدواژه‌های نویسندگان است و همین‌طور ۲۵ درصد از مقاله‌ها بیشتر از ۶۴ درصد دارای اشتراک کلیدواژه بین کلیدواژه‌های عنوان و کلیدواژه‌های نویسندگان است. در شکل ۲ که نمودار جعبه‌ای نتایج است، این میزان همپوشانی قابل‌مشاهده است.

پرسش دوم: همپوشانی بین کلیدواژه‌های چکیده با کلیدواژه‌های نویسندگان به چه میزان است؟

در پاسخ به پرسش دوم پژوهش همان‌گونه که در جدول ۱ مشاهده می‌شود، میزان همپوشانی بین کلیدواژه‌های چکیده با کلیدواژه‌های نویسندگان بر اساس معیار تمرکز میانه ۱۸ درصد است که بر اساس چارک اول ۲۵ درصد از مقالات کمتر از ۱۲ درصد اشتراک کلیدواژه بین کلیدواژه‌های چکیده و کلیدواژه‌های نویسندگان را دارند و درعین‌حال ۷۵ درصد از مقاله‌ها بیشتر از ۱۲ درصد اشتراک کلیدواژه بین کلیدواژه‌های عنوان و کلیدواژه‌های نویسندگان را دارد؛ و همان‌طور که در چارک سوم مشاهده می‌شود، ۷۵ درصد مقالات دارای کمتر از ۲۵ درصد اشتراک کلیدواژه بین کلیدواژه‌های عنوان و کلیدواژه‌های نویسندگان است. همچنین ۲۵ درصد از مقاله‌ها بیشتر از ۲۵ درصد دارای اشتراک کلیدواژه بین کلیدواژه‌های عنوان و کلیدواژه‌های نویسندگان است. در نمودار جعبه‌ای که در شکل ۲ آمده است، این میزان همپوشانی قابل‌مشاهده است.

پرسش سوم: همپوشانی بین کلیدواژه‌های چکیده با کلیدواژه‌های عنوان به چه میزان است؟

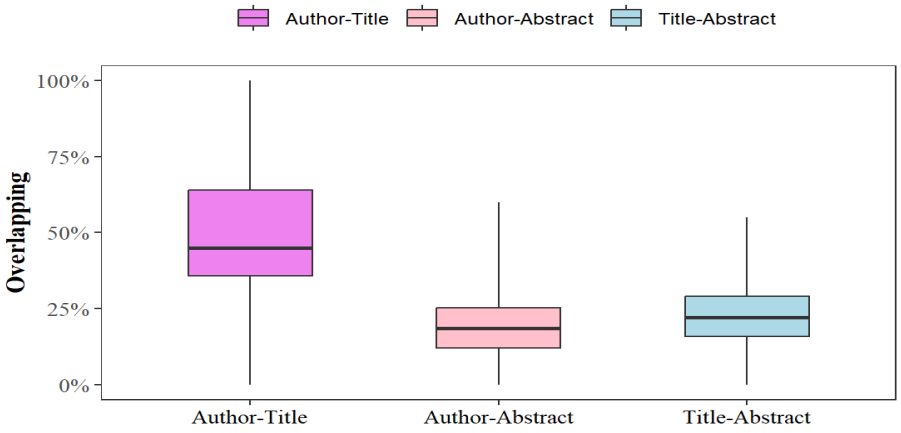
میزان پوشش کلیدواژه‌های چکیده توسط کلیدواژه‌های عنوان با پوشش ۲۲ درصد بر اساس معیار تمرکز میانه در جدول ۱ قابل‌مشاهده است. ازآنجایی‌که چارک اول نشان‌دهنده این است که ۲۵ درصد از مقالات کمتر از ۱۶ درصد اشتراک کلیدواژه بین کلیدواژه‌های چکیده و

کلیدواژه‌های عنوان را دارند، ۷۵ درصد از مقالات بیشتر از ۱۶ درصد اشتراک کلیدواژه بین کلیدواژه‌های چکیده و کلیدواژه‌های عنوان را دارد. همان‌گونه که در چارک سوم نیز دیده می‌شود، ۷۵ درصد مقالات دارای کمتر از ۲۹ درصد اشتراک کلیدواژه بین کلیدواژه‌های چکیده و کلیدواژه‌های عنوان است. همچنین ۲۵ درصد از مقالات بیشتر از ۲۹ درصد دارای اشتراک کلیدواژه بین کلیدواژه‌های چکیده و کلیدواژه‌های عنوان است. در نمودار جعبه‌ای که در شکل ۲ آمده است، این میزان همپوشانی قابل‌مشاهده است.

جدول ۱- میزان همپوشانی بین کلیدواژه‌های عنوان/نویسنده، چکیده/نویسنده و چکیده/عنوان

همپوشانی کلیدواژه‌ها شاخص (درصد)	عنوان/ نویسنده (Author/ Title)	چکیده/ نویسنده (Author/ Abstract)	چکیده/ عنوان (Title / Abstract)
میانه	۴۵	۱۸	۲۲
میانگین	۴۹	۲۰	۲۳
چارک اول (Q1) (۲۵)	۳۶	۱۲	۱۶
چارک سوم (Q3) (۷۵)	۶۴	۲۵	۲۹
حداقل (Min)	۰	۰	۰
حداکثر (Max)	۱۰۰	۶۰	۵۵

حداکثر میزان همپوشانی کلیدواژه‌های عنوان با کلیدواژه‌هایی نویسندگان در بین مجموعه ۱۰۰ درصد، همچنین حداکثر میزان همپوشانی کلیدواژه‌های چکیده با کلیدواژه‌هایی نویسندگان در بین مجموعه داده‌ها ۶۰ درصد و درنهایت حداکثر میزان همپوشانی بین کلیدواژه‌های عنوان با کلیدواژه‌هایی چکیده نیز ۵۵ درصد است. حداقل میزان همپوشانی بین کلیدواژه‌های عنوان با کلیدواژه‌های نویسندگان، کلیدواژه‌های چکیده با کلیدواژه‌های نویسندگان و کلیدواژه‌های عنوان با کلیدواژه‌های چکیده در بین مقاله‌ها نیز صفر درصد است. در نمودار جعبه‌ای که در شکل ۲ دیده می‌شود، این میزان همپوشانی را می‌توان مشاهده کرد.



شکل ۲- نمودار جعبه‌ای میزان همپوشانی بین کلیدواژه‌های عنوان/نویسنده، چکیده/نویسنده و چکیده/عنوان

پرسش چهارم: در کدام‌یک از رشته های موردبررسی در پژوهش حاضر انتخاب کلیدواژه توسط نویسنده با همپوشانی بیشتری انجام شده است؟

مجموعه داده شامل ۵۰۰ تحقیق علمی در پنج گروه روانشناسی (G۱)، فناوری اطلاعات و ارتباطات (G۲)، معماری (G۳)، مدیریت دولتی (G۴) و حقوق عمومی (G۵) و هر گروه شامل ۱۰۰ پژوهش گروه‌بندی شد. نتایج مربوط به این پرسش از پژوهش در جداول ۲ و ۳ و همچنین شکل ۳ قابل‌مشاهده است:

جدول ۲- میزان همپوشانی بین کلیدواژه‌های عنوان/نویسنده به تفکیک گروه‌های موضوعی در پژوهش حاضر

رشته شاخص (درصد)	روانشناسی (G۱)	فناوری اطلاعات و ارتباطات (G۲)	معماری (G۳)	مدیریت دولتی (G٤)	حقوق عمومی (G۵)
میانه	۵۶	۴۵	۳۸	۵۴	۵۰
میانگین	۵۴	۴۸	۳۹	۵۳	۵۱
چارک اول (Q1) (۲۵)	۴۲	۳۳	۲۸	۳۸	۳۸
چارک سوم (Q3) (۷۵)	۶۷	۶۴	۴۸	۶۸	۶۷
حداقل (Min)	۱۱	۰	۱۰	۰	۰
حداکثر (Max)	۱۰۰	۹۰	۸۰	۱۰۰	۱۰۰

همان‌گونه که در جدول ۲ قابل‌مشاهده است، بر اساس معیار تمرکز میانه به‌طور متوسط رشته روانشناسی با ۵۶ درصد بیشترین میزان همپوشانی بین کلیدواژه‌های عنوان با کلیدواژه‌هایی که نویسنده برای پژوهش خود برگزیده است را در میان رشته‌های دیگر داشته است. بازه بین چارک اول و چارک دوم نشان‌دهنده پراکندگی میانه داده‌هاست. بر همین اساس چارک اول نشان می‌دهد که ۲۵ درصد از مقاله‌ها در رشته روانشناسی کمتر از ۴۲ درصد اشتراک کلیدواژه بین کلیدواژه‌های عنوان و کلیدواژه‌های نویسندگان را دارند و درعین‌حال ۷۵ درصد از مقاله‌ها در این رشته بیشتر از ۴۲ درصد اشتراک کلیدواژه بین کلیدواژه‌های عنوان و کلیدواژه‌های نویسندگان را دارد؛ و در چارک سوم خلاف این حالت است، یعنی ۷۵ درصد مقالات در این رشته دارای کمتر از ۶۷ درصد اشتراک کلیدواژه بین کلیدواژه‌های عنوان و کلیدواژه‌های نویسندگان است و همین‌طور ۲۵ درصد از مقاله‌ها در این رشته بیشتر از ۶۷ درصد دارای اشتراک کلیدواژه بین کلیدواژه‌های عنوان و کلیدواژه‌های نویسندگان است. همچنین با توجه به معیار تمرکز میانه، در رشته معماری با ۳۸ درصد همپوشانی کمترین میزان همپوشانی بین کلیدواژه‌های عنوان و کلیدواژه‌های نویسندگان مشاهده شد. نتایج حاصل در شکل ۲ نمودار جعبه‌ای Author-Title برای پنج گروه از رشته‌های موردبررسی قابل‌مشاهده است. حداکثر میزان همپوشانی کلیدواژه‌های عنوان با کلیدواژه‌های نویسندگان در بین مقاله‌ها با ۱۰۰ درصد همپوشانی نیز مربوط به رشته‌های روانشناسی، مدیریت دولتی و حقوق عمومی است و حداقل میزان همپوشانی کلیدواژه‌های عنوان با کلیدواژه‌های نویسندگان که در بین مقاله‌ها وجود داشته است، در رشته‌های فناوری اطلاعات و ارتباطات، مدیریت دولتی و حقوق عمومی با همپوشانی صفر مشاهده شد.

طبق استاندارد سازمان بین‌المللی استاندارد[1] (ISO) چکیده عبارت است از نشان دادن «محتوای یک مدرک، به‌صورت خلاصه، دقیق و بی‌هیچ نقد و تفسیر». این بخش از هر مقاله به خواننده کمک می‌کند که در صورت تناسب محتوا با نیاز اطلاعاتی وی از آن در راستای پژوهش خود بهره‌مند شود؛ بنابراین میزان دقت و همپوشانی مناسب این بخش از پژوهش‌های علمی نقش بسزایی در نمایه‌سازی خودکار داده‌های متنی در این پژوهش‌ها دارد. بر اساس

[1]. International Organization for Standardization

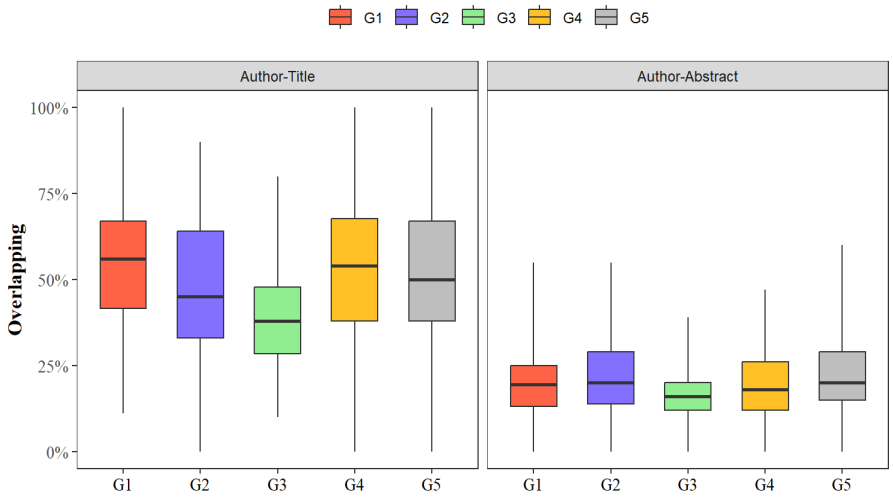
پرسش چهارم پژوهش، نتایج حاصل از میزان همپوشانی بین کلیدواژه‌های چکیده و کلیدواژه‌های نویسندگان در مجموعه داده در این پژوهش قابل‌مشاهده است:

جدول ۳- میزان همپوشانی بین کلیدواژه‌های چکیده/نویسنده به تفکیک گروه‌های موضوعی در پژوهش حاضر

رشته شاخص (درصد)	روانشناسی (G۱)	فناوری اطلاعات و ارتباطات (G۲)	معماری (G۳)	مدیریت دولتی (G٤)	حقوق عمومی (G٥)
میانه	۲۰	۲۰	۱۶	۱۸	۲۰
میانگین	۲۰	۲۲	۱۶	۱۹	۲۳
چارک اول (Q1) (۲۵)	۱۳	۱۴	۱۲	۱۲	۱۵
چارک سوم (Q3) (۷۵)	۲۵	۲۹	۲۰	۲۶	۲۹
حداقل (Min)	۰	۰	۰	۰	۰
حداکثر (Max)	۵۵	۵۵	۳۹	۴۷	۶۰

مطابق با آنچه در جدول ۳ آمده بر اساس معیار تمرکز میانه، میزان همپوشانی بین کلیدواژه‌های چکیده با کلیدواژه‌های نویسندگان به‌طور یکسان در سه رشته حقوق عمومی، فناوری اطلاعات و ارتباطات و روانشناسی با ۲۰ درصد همپوشانی نسبت به دو رشته دیگر یعنی رشته مدیریت دولتی با ۱۸ درصد و رشته معماری با ۱۶ درصد دارای بیشترین میزان همپوشانی است. این در حالی است که در سایر شاخص‌های دیگر نیز این میزان همپوشانی قابل‌مشاهده است. به‌عنوان‌مثال، برای رشته فناوری اطلاعات بر اساس چارک اول ۲۵ درصد از مقاله‌ها در این رشته کمتر از ۱۴ درصد اشتراک کلیدواژه بین کلیدواژه‌های چکیده و کلیدواژه‌های نویسندگان را دارند و همین‌طور ۷۵ درصد از مقاله‌ها در این رشته بیشتر از ۱۴ درصد اشتراک کلیدواژه بین کلیدواژه‌های چکیده و کلیدواژه‌های نویسندگان را دارد. بر اساس چارک سوم نیز ۷۵ درصد مقالات در این رشته دارای کمتر از ۲۹ درصد اشتراک کلیدواژه بین کلیدواژه‌های چکیده و کلیدواژه‌های نویسندگان است و همین‌طور ۲۵ درصد از مقاله‌ها در این رشته بیشتر از ۲۹ درصد دارای اشتراک کلیدواژه بین کلیدواژه‌های چکیده و کلیدواژه‌های نویسندگان است. حداکثر میزان همپوشانی کلیدواژه‌های چکیده با کلیدواژه‌های نویسندگان در بین مقاله‌ها با ۶۰ درصد همپوشانی در رشته حقوق عمومی و با اختلاف ۵ درصد با میزان همپوشانی ۵۵

درصد مربوط به رشته‌های روانشناسی و فناوری اطلاعات و ارتباطات است و حداقل میزان همپوشانی کلیدواژه‌های چکیده با کلیدواژه‌های نویسندگان که در بین مقاله‌ها وجود داشته است، در پنج رشته موردبررسی در پژوهش حاضر با همپوشانی صفر مشاهده شد. نتایج به‌دست‌آمده در شکل ۳ نمودار جعبه‌ای Author-Abstract پنج گروه از رشته‌های موردبررسی را نشان می‌دهد.



شکل ۳- نمودار جعبه‌ای میزان همپوشانی کلیدواژه‌های عنوان/نویسنده و چکیده/نویسنده به تفکیک گروه‌های موضوعی در پژوهش حاضر

نتیجه‌گیری

مقاله‌های علمی به‌عنوان یکی از اسناد متنی پراهمیت درزمینهٔ پژوهشی نیازمند سازمان‌دهی مناسب و ذخیره و بازیابی صحیح، دقیق و سریع هستند. درصورتی‌که این امر با استفاده از شیوه‌های خودکار متناسب با این منابع انجام شود، امکان دسترسی آسان، سریع و دقیق را برای کاربران فراهم می‌کند. در پژوهش حاضر با استفاده از الگوریتم «Rake» که یکی از الگوریتم‌های استخراج خودکار کلیدواژه برای داده‌های متنی است، اقدام به استخراج کلیدواژه‌های مجموعه داده‌هایی که متشکل از ۵۰۰ مقاله علمی در پنج گروه موضوعی در رشته‌های روانشناسی، فناوری اطلاعات و ارتباطات، معماری، مدیریت دولتی و حقوق عمومی

شد. سپس به بررسی میزان همپوشانی عنوان، چکیده و کلیدواژه‌های نویسندگان که عناصر واسط در بازنمایی اسناد متنی در قالب مقاله‌های پژوهشی است، پرداخته شد. در این پژوهش ابتدا میزان همپوشانی کلیدواژه‌های عنوان با کلیدواژه‌های نویسندگان بررسی شد. با توجه به معیار تمرکز میانه میزان همپوشانی بین کلیدواژه‌های عنوان با کلیدواژه‌های نویسندگان ۴۵ درصد همپوشانی داشت. سپس همپوشانی بین کلیدواژه‌های چکیده با کلیدواژه‌های نویسندگان موردبررسی قرار گرفت. در این بررسی میزان همپوشانی بین کلیدواژه‌های چکیده با کلیدواژه‌های نویسندگان بر اساس معیار تمرکز میانه ۱۸ درصد نشان داده شد و در ادامه این بررسی‌ها میزان پوشش کلیدواژه‌های چکیده توسط کلیدواژه‌های عنوان با پوشش ۲۲ درصد بر اساس معیار تمرکز میانه مشاهده شد. بررسی‌های صورت گرفته و نتایج به‌دست‌آمده در نمودار جعبه‌ای شکل ۲ جهت مقایسه و نمایش توزیع درصدهای همپوشانی قابل‌مشاهده است. بر این اساس می‌توان گفت میزان همپوشانی و پراکندگی بین کلیدواژه‌های چکیده و کلیدواژه‌های نویسندگان و همچنین همپوشانی بین کلیدواژه‌های چکیده و کلیدواژه‌های عنوان به‌طور متوازن و تقریبی حول میانه و به یک اندازه است. این در حالی است که میزان همپوشانی بین کلیدواژه‌های عنوان و کلیدواژه‌های نویسندگان این‌گونه نیست و مشاهده می‌شود مقاله‌هایی که بین چارک دوم (میانه) و چارک سوم (۷۵ درصد) قرار دارند نسبت به مقاله‌هایی که در چارک اول (۲۵ درصد) و چارک دوم (میانه) هستند دارای پراکندگی و درصد همپوشانی بیشتری هستند. می‌توان نتیجه گرفت که احتمال بیشتری وجود دارد که کلیدواژه‌های عنوان و کلیدواژه‌های نویسنده یک مقاله در مقایسه با کلیدواژه‌های چکیده و کلیدواژه‌های نویسنده و همچنین کلیدواژه‌های چکیده و کلیدواژه‌های عنوان همپوشانی بیشتری داشته باشند؛ بنابراین ازنظر همپوشانی، بین عنوان و نویسنده این امر بیشتر اتفاق افتاده است. یکی دیگر از مواردی که در این پژوهش به بررسی آن پرداخته شد، میزان دقت در انتخاب کلیدواژه توسط نویسنده به تفکیک پنج رشته دانشگاهی بود. نتایج حاصل در جدول‌های ۲ و ۳ و نمایش توزیع درصدهای همپوشانی آن‌ها در قالب نمودار جعبه‌ای در شکل ۳ قابل‌مشاهده است. در نتایج به‌دست‌آمده بیشترین میزان همپوشانی بین کلیدواژه‌های عنوان با کلیدواژه‌های نویسندگان به رشته روانشناسی و رشته مدیریت دولتی و کمترین میزان همپوشانی بین کلیدواژه‌های عنوان با کلیدواژه‌هایی نویسندگان به رشته معماری تعلق گرفت. در رشته فناوری اطلاعات و حقوق

عمومی نیز نیاز به بهبود و تقویت درک مفاهیم و مهارت‌های مرتبط با آن در بین پژوهشگران لازم به نظر می‌رسد. در رشته معماری با توجه به اینکه کمترین میزان همپوشانی را به خود اختصاص داده است، ضرورت تمرکز و توجه بیشتر بر مفاهیم تخصصی در این رشته حائز اهمیت است. همچنین بر اساس معیار تمرکز میانه، میزان همپوشانی بین کلیدواژه‌های چکیده با کلیدواژه‌های نویسندگان در سه رشته حقوق عمومی، فناوری اطلاعات و ارتباطات و روانشناسی یکسان بود و در مقابل دو رشته مدیریت دولتی و معماری به ترتیب با فاصله نسبتاً جزئی‌تری، کمترین میزان همپوشانی را داشتند. می‌توان گفت که میزان همپوشانی بین کلیدواژه‌های چکیده و کلیدواژه‌های نویسندگان به شکل تقریباً یکسان در پنج گروه موضوعی در یک سطح ۲۰ درصدی نسبت به‌کل مقاله است که می‌تواند بیان‌کننده محتوای اصلی مقاله در یک قالب کوچک به نام چکیده باشد. استفاده مناسب از کلیدواژه‌ها، نوشتن چکیده‌هایی با محتوای هماهنگ با موضوع موردنظر و انتخاب عناوین متناسب می‌تواند به چندین روش به بهبود فرایند استخراج مفاهیم، ذخیره‌سازی و بازیابی مقالات علمی کمک‌کننده باشد ازجمله اینکه کلیدواژه‌ها، چکیده‌ها و عناوین می‌توانند به‌عنوان ورودی برای الگوریتم‌های استخراج مفاهیم عمل کنند. این الگوریتم‌ها می‌توانند با استفاده از این اطلاعات به شناسایی مفاهیم اصلی مقاله بپردازند. همچنین ازآنجایی‌که کلیدواژه‌ها، چکیده‌ها و عناوین می‌توانند به‌عنوان بخش‌هایی از ساختار ذخیره‌سازی اطلاعات مورداستفاده قرار گیرند، این ساختارها در سرعت دسترسی کاربران به اطلاعات موردنیازشان نقش تأثیرگذاری خواهند داشت. در رابطه با بازیابی نیز کلیدواژه‌ها، چکیده‌ها و عناوین می‌توانند به‌عنوان ورودی برای الگوریتم‌های بازیابی اطلاعات عمل کنند. این الگوریتم‌ها می‌توانند از این اطلاعات برای دسترسی سریع به مقالات مرتبط کمک بسزایی داشته باشند. یافته‌های به‌دست‌آمده می‌توانند در تداوم پژوهش‌های مرتبط با این پژوهش مورداستفاده قرار گیرند. به‌عنوان‌مثال، می‌توان در پژوهش‌های آینده بر روی بهبود روش‌های نوشتن چکیده و انتخاب عنوان متمرکز شد تا بهبود یافته‌های استخراج، ذخیره‌سازی و بازیابی مقالات علمی را تسهیل کند. در انتها می‌توان گفت اگر چکیده‌ها به‌خوبی بیان شوند و عناوین متناسب با محتوای مقاله‌های پژوهشی انتخاب شوند، به‌طورقطع منابع مناسبی برای استخراج مفاهیم، ذخیره و بازیابی و دسترس‌پذیری سریع و آسان خواهند بود. در ادامه به‌منظور تداوم پژوهش‌های مرتبط با پژوهش حاضر بر اساس یافته‌های پژوهش

پیشنهاد می‌شود، با توجه به اینکه در این پژوهش میزان همپوشانی کلیدواژه‌ها در زبان فارسی موردبررسی قرار گرفته است، در آینده این بررسی در متون چندزبانه نیز موردپژوهش قرار بگیرد. همچنین میزان همپوشانی کلمات کلیدی در متون تخصصی و فنی و تأثیر آن بر دقت و کیفیت بازیابی اطلاعات متنی نیز به‌عنوان هدف پژوهشی در آینده مدنظر قرار گیرد. ارزیابی تأثیر تعداد و نوع کلیدواژه‌های استفاده‌شده (ساده و مرکب) توسط نویسندگان و مقایسه آن با تعداد و نوع کلیدواژه‌های استخراج‌شده بر دقت و کارایی بازنمایی اسناد متنی نیز می‌تواند در راستای پژوهش حاضر انجام شود. به‌کارگیری الگوریتم‌های دیگر متناسب با داده‌های متنی و مقایسه آن با کارکرد الگوریتم «Rake» استفاده‌شده در این پژوهش به‌عنوان یک پژوهش کاربردی پیشنهاد می‌شود. در این پیشنهاد می‌توان میزان همپوشانی کلیدواژه‌های دستی و ماشینی به‌منظور بررسی دقت و سرعت دستیابی به اطلاعات موردبررسی قرار گیرد. این امر درنهایت بر تحلیل مناسب و دقیق‌تر میزان همپوشانی کلیدواژه‌ها تأثیر مثبتی خواهد داشت.

منابع

انصاری، مریم (۱۳۷۹). *بررسی انطباق میان توصیفگرهای نمایه‌سازی و کلیدواژه‌های عنوان پایان‌نامه‌های دکترای تخصصی کودکان، زنان، قلب و عروق و روان‌پزشکی*. پایان‌نامه کارشناسی ارشد کتابداری و اطلاع‌رسانی پزشکی، دانشگاه علوم پزشکی و خدمات بهداشتی-درمانی ایران، تهران.

انصاری، مصطفی، روضه، محبوبه، مشایخ کندسکلایی، کبری و گوهری وثوق، صالحه (۱۴۰۰). بررسی میزان انطباق کلیدواژه‌های پایان‌نامه‌های پرستاری و مامایی دانشگاه‌های علوم پزشکی شهر تهران با سر عنوان‌های موضوعی پزشکی اصطلاح‌نامه MeSH. *نشریه پژوهش پرستاری*، ۱۶ (۳):۱-۸.

بنی اقبال، ناهید، خسروی، فریبرز و پیر هادی، صدیقه (۱۳۹۰). مقایسه واژه‌های عنوان و چکیده پایان‌نامه‌ها با توصیفگرهای تعیین‌شده در نمایه سازمان اسناد و کتابخانه ملی ایران. *مطالعات ملی کتابداری و سازمان‌دهی اطلاعات*، ۸۶ (۲): ۱۳۴-۱۴۷.

خطیر، اشکان و گنجه‌فر، سهیل (۱۳۹۷). تحلیل توزیع و تمرکز کلیدواژه‌های پایان‌نامه‌ها و رساله‌ها و میزان تطابق با توصیفگرها، عنوان و چکیده. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳۴ (۱): ۴۱۱-۴۲۸.

دانش، فرشید و رحیمی، فروغ (۱۴۰۲). داده‌کاوی متنی انتشارات کووید-۱۹ به‌منظور کشف و استخراج روندهای نوظهور. *مجله میکروب‌شناسی پزشکی ایران*، ۱۷ (۲): ۱۵۰-۱۶۰.

داورپناه، محمدرضا (۱۳۷۵). بررسی میزان سازگاری عناوین مقالات فارسی با محتوای آن‌ها. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۱۲ (۲): ۱-۱۲.

قاضی میر سعید، جواد و مسعودی، فاطمه (۱۳۹۸). بررسی حضور توصیفگرهای MeSH در مقالات مجلات ایرانی دندانپزشکی به زبان لاتین و نمایه شده درPubMed. *مجله دانشکده دامپزشکی مشهد*، ۴۳ (۲):۱۴۸-۱۵۴.

کریمی‌منش، مصطفی (۱۳۹۲). *کشف کلیدواژه‌های یک مستند بر مبنای آنالیز معنایی*. پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر نرم‌افزار، دانشگاه پیام نور استان تهران، تهران.

محرابی، الهه، محبی، آزاده و احمدی، عباس (۱۴۰۰). بهبود الگوریتم Rake برای استخراج کلیدواژه از متون علمی فارسی. مطالعه موردی: پایان‌نامه‌ها و رساله‌های فارسی. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳۷ (۱): ۱۹۷-۲۲۸.

یلوه، الهام، نوروزی، یعقوب و خطیر، اشکان (۱۴۰۰). مروری نظام‌مند بر پژوهش‌های بهبود الگوریتم کا-میانه برای خوشه‌بندی داده‌ها. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳۷ (۲): ۵۲۷-۵۵۶.

یلوه، الهام، نوروزی، یعقوب و خطیر، اشکان (۱۴۰۲). بهینه‌سازی سازمان‌دهی اسناد متنی فارسی با استفاده از تکنیک خوشه‌بندی. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳۸ (۳):۹۳۷-۹۶۸.

References

Ansari, M. (2018). *Investigation of the Compatibility between Indexing Descriptors and Keywords of Specialized Doctoral Theses Titles in Pediatrics, Women, Cardiology, and Psychiatry*. Master's Thesis in Medical Library and Information Science, Faculty of Medical Management and Information Science, Iran University of Medical Sciences and Health Services, Tehran. [In Persian]

Ansari, M., Rouzeh, M., Mashaekh Kandeskalaei, K., & Gohari Vosoogh, S. (2021). Investigation of the Compatibility of Keywords in Nursing and Midwifery Theses of Medical Sciences Universities in Tehran City with MeSH Medical Subject Headings. *Nursing Research Journal*, 16 (3):1-8. [In Persian]

Bani ghbal, N., Khosravi, F., & Pir Hadi, S. (2011). Comparison of Thesis Title and Abstract Words with Descriptors Determined in the Index of the National Library and Archives of Iran. *National Library and Information Organization Studies*, 86 (2): 134-147. [In Persian]

Baruni, J., & Sathiaseelan, J. (2020). Keyphrase Extraction from Document Using RAKE and TextRank Algorithms. *Int. J. Comput. Sci. Mob. Comput*, 9: 83-93.

Danesh, F., & Rahimi, F. (2023). Text Mining of COVID-19 Publications for Discovery and Extraction of Emerging Trends. *Iranian Journal of*

Medical Microbiology, 17 (2):150-160. [In Persian]

Davar Panah, M. (1996). Investigation of the Compatibility of Persian Article Titles with Their Content. *Information Processing & Management Journal*, 12 (2):1-12. [In Persian]

Derzi Khallordi, S., & Rezavi, A. A. (2018). The Concordance of Keywords in Articles of Sari University of Agricultural Sciences and Natural Resources with CAB Thesaurus. *Knowledge Studies Quarterly (Library and Information Science and Information Technology)*, 41(11): 48-57. [In Persian]

Ghazi Mir Saeed, J., & Masoudi, F. (2019). Investigation of the Presence of MeSH Descriptors in Latin Language Articles of Iranian Dental Journals Indexed in PubMed. *Journal of Mashhad Faculty of Veterinary Medicine*, 43 (2): 54-148. [In Persian]

Gil- Leiva, I., & Alonso- Arroyo, A. (2007). Keywords given by authors of scientific articles in database descriptors. *Journal of the American society for information science and technology*, 58(8): 1175-1187.

Hulth, A. (2003). Improved Automatic Keyword Extraction Given More Linguistic Knowledge. *Conference on Empirical Methods in Natural Language Processing*.216-223.DOI: https://doi.org/10.3115/1119355.1119383

Karimi Manesh, M. (2013). *Discovery of Keywords in a Documentary Based on Semantic Analysis*. Master's thesis in Computer-Software Engineering, Payam Noor University, Tehran Province. [In Persian]

Kharazi, H. (2015). Persian Stop Word List. https://github.com/kharazi/persian-stopwords (Retrieved: May 31, 2021).

Khatir, A., & Ganjehfar, S. (2018). Analysis of Distribution and Concentration of Keywords in theses and dissertations and their alignment with descriptors, title, and abstract. *Information Processing & Management Journal*, 34 (1): 411-428. [In Persian]

Kim, D., Lee, M. H., & Choi, M. (2016). Comparison and analysis of keywords in the Korean ophthalmic optics society articles to MeSH terms. *Journal of Korean Ophthalmic Optics Society*, 21(2): 83-90.

Kipp, M. E. (2011). Tagging of biomedical articles on CiteULike: A comparison of user, author and professional indexing. *Knowledge Organization*, 38(3): 245-261.

Liu, F., Huang, X., Huang, W., & Duan, S. X. (2020). Performance evaluation of keyword extraction methods and visualization for student online comments. *Symmetry*, 12(11): 19-23.

Mehrabi, E., Mohebbi, A., & Ahmadi, A. (2021). Improving the Rake Algorithm for Extracting Keywords from Persian Scientific Texts. Case Study: Persian Theses and Dissertations. *Information Processing & Management Journal*, 37 (1): 197-228. [In Persian]

Névéol, A., Doğan, R. I., & Lu, Z. (2010). Author keywords in biomedical journal articles. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2010, 537–541.

Parsaei Mohammadi, P., Ghasemi, A. H., & Hassanzadeh-Beheshtabad, R. (2017). A comparative study of the origin, structure, and indexing language of the Persian and English keywords of articles indexed in the IranMedex database and their compliance with the Persian medical thesaurus and Medical Subject Headings. *Journal of education and health promotion*, 6(1), 2. DOI:https://doi.org/10.4103/jehp.jehp_137_14

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual. documents. *Text mining: applications and theory*, 1: 1-2. DOI: https://doi.org/10.1002/9780470689646.ch1

Song, G., Ye, Y., Du, X., Huang, X., & Bie, S. (2014). Short text classification: a survey. *Journal of multimedia*, 9(5): 635-643.

Subramanian, L., & Karthik, R. (2017). Keyword Extraction: A Comparative Study Using Graph Based Model And Rake. Publication on International Journal of Advanced Research, Article.5(3):1133-1137.

Yalveh, E., Norouzi, Y., & Khatir, A. (2021). A Systematic Review of K-means Algorithm Improvement Research for Data Clustering. *Information Processing & Management Journal*, 37 (2): 527-556. [In Persian]

Yalveh, E., Norouzi, Y., & Khatir, A. (2023). Optimizing the Oranization of Persian Text Documents Using Clustering Technique. *Information Processing & Management Journal*, 38 (3): 937-968. [In Persian]

Zhang, C. (2008). Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems*, 4(3): 1169-1180.