Journal of Knowledge Studies
(Library and Information Science and Information Technology)
I. A. U. North Tehran Branch
Print ISSN: 2008-2754 Online ISSN: 2783-4115

Optimizing the provision of archival material services based on the design and application of the book locator tool

Reza Ahmadi Zamani[1] | **Zohreh Mirhosseini**[2]

1- Ph.D. student of Knowledge and Information Science, Islamic Azad University - North Tehran Branch. daneshzamani@yahoo.com
2- Associate Professor of Islamic Azad University, North Tehran Branch, PhD in Information Science. (Corresponding Author) z_mirhoseini@iau-tnb.ac.ir, zmirhosseini@yahoo.com

Article Info

Article type:
Research Article

Article history:
Received:
21 January 2023
Received in revised form:
14 February 2023
Accepted:
01 March 2023
Published online:
16 March 2023

Keywords:
Source Archive,
The Librarian,
Returning And Sorting Resources Carefully,
Shelf,
Book Locator Tool,
Archival Material.

ABSTRACT

Objective: "To investigate the optimization of the provision of archival material services based on the design and application of the book locator tool, in order to facilitate the restoration and accurate arrangement of resources. The existence of issues and problems in the archive of documents and documents, such as: confusion; depreciation; the disappearance of archival materials; Moving and misplacing; The accuracy in returning the sources and other things was investigated and led to the invention of the "book locator tool".

Methodology: It is a semi- experimental method and a type of applied and fundamental research. The statistical population of the research included 10 libraries in Tehran, which have several characteristics in common. In this research, it was done by distributing "pre-test" and "post-test" questionnaires among 50 employees working in the repository and archive of the target community. number of fifty "book locator tools" were made and tested for two months, then the frequency distribution was done and analyzed using Excel and EPSS software.

Results: The results showed that by replacing the "book locator tool" the problem of confusion of other sources has been solved by 96%. More than 71% of the sample people consider the use of "new device" to be effective in reducing the depreciation of archival materials. More than 87% have stated that by testing the proxy device, it is fast and takes less time (between 30 and 60 seconds). According to 98% of the people in the group, using the locator tool, the amount of "displacement and misarrangement" of archival materials was resolved when restoring and arranging. And 96% have seen the "flashing light" of the locator on the shelf, which has caused the attention of librarians to be attracted quickly and has also been effective in saving time.

Conclusion: The results of the research showed that 100% of the studied libraries are without tools and devices such as "Book Locator" and the sorting of documents is done by the employees and librarians working in the archive. 90% of the people in the sample have evaluated the use of the "book locator tool" useful, which indicates the efficiency of the replacement device. Accordingly, it can be used generally for all libraries.

Cite this article: Ahmadi Zamani, R. & Mirhosseini, Z. (2022). Optimizing the provision of archival material services based on the design and application of the book locator tool. *Journal of Knowledge Studies*, 15(59), 54-73.
DOR :20.1001.1.20082754.1401.15.59.1.1

© The Author(s).
Publisher: Islamic Azad University North Tehran Branch

دانش شناسی
(علوم کتابداری و اطلاع رسانی و فناوری اطلاعات)
شاپا الکترونیکی: ۴۱۱۵-۲۷۸۳ شاپا چاپی: ۲۷۵۴-۲۰۰۸

شناسایی روابط موضوعی بین منابع مورد استفاده توسط کاربران مرکز منطقه‌ای اطلاع رسانی علوم و فناوری با استفاده از تکنیک متن کاوی

خجسته شعبانی[۱] | عاصفه عاصمی[۲]

۱- کارشناس ارشد رشته علم اطلاعات و دانش‌شناسی_دانشگاه اصفهان (نویسنده مسئول) khojastehshabani@yahoo.com
۲- عضو هیئت علمی دانشگاه کورینوس بوداپست، دانشیار دانشگاه اصفهان af_asemi@yahoo.com

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۴۰۱/۱۱/۰۱
تاریخ بازنگری: ۱۴۰۱/۱۱/۲۵
تاریخ پذیرش: ۱۴۰۱/۱۲/۱۰
تاریخ انتشار آنلاین: ۱۴۰۱/۱۲/۲۵

واژه‌های کلیدی:
داده کاوی،
متن کاوی،
تحلیل هم رخدادی واژگان،
پایگاه اطلاعاتی رایسست.

چکیده

هدف: هدف اصلی پژوهش حاضر بررسی روابط موضوعی در عناوین منابع مورد استفاده توسط کاربران رایسست با استفاده از تکنیک متن کاوی بود. بنابراین، به بازتاب چگونگی روابط موضوعی در منابع اطلاعاتی کاربران در مرکز رایسست مبادرت شده، تا از طریق شناخت به رفتار و احساس استفاده کنندگان دست یابند.

روش پژوهش: روش پژوهش مبتنی بر متن کاوی بود، که به داده کاوی بر روی متن، تحلیل متن و به منظور فرایند استخراج اطلاعات با کیفیت از متن اشاره دارد. دسترسی اطلاعات به متن کامل مقالات مجلات علمی – پژوهشی، علمی – ترویجی، مجموعه مقالات کنفرانس‌ها و همایش‌های علمی، کتاب‌های لاتین و فارسی جامعه آماری پژوهش را تشکیل داده، که با استفاده از روش سرشماری، کلیه داده‌های حاصل از گزارش‌گیری توسط رایسست بررسی گردید. به منظور تجزیه و تحلیل داده‌ها و تحلیل متن از نرم افزار ویانت، و برای پاکسازی و نرمال سازی داده‌ها از نرم افزار پایتون بهره جویی گردید.

یافته‌ها: براساس یافته‌ها از داده‌های حاصل شده، ۲۱ کلمه و ۱۶۰ کلمه موضوعی پرتکرار از منبع مورد استفاده در پایگاه اطلاعاتی رایسست مشخص گردید. دور نمای لوم از چگونگی توزیع کلمات موضوعی با تکرار بالا تهیه شده و ضریب همبستگی تکرار موضوعات پر استفاده در عنوان‌های منابع اطلاعاتی تدوین شد. به منظور تدوین نمایه درهم کرد کلمات موضوعی پر تکرارترند (Trend) استفاده شد.

نتیجه‌گیری: نتایج نشان داد که تدوین پژوهش در مجموعه سازی منابع الکترونیکی پایگاه‌های اطلاعاتی و پیش‌نگری در آینده این دسته از منابع به مدیران مراکز اطلاع رسانی و کاربران آنها مفید است.

استناد: شعبانی، خ. و عاصمی، ع. (۱۴۰۱). شناسایی روابط موضوعی بین منابع مورد استفاده توسط کاربران مرکز منطقه‌ای اطلاع رسانی علوم و فناوری با استفاده از تکنیک متن کاوی. *دانش‌شناسی*، ۱۵ (۵۹)، ۵۴-۷۳.

DOR: 20.1001.1.20082754.1401.15.59.4.4

ناشر: دانشگاه آزاد اسلامی واحد تهران شمال

حق‌مؤلف © نویسندگان.

مقدمه

فعالیت‌های پژوهشی در قالب‌های مختلف مانند طرح پژوهشی، پایان نامه، مقاله علمی، و اشکال دیگر ارائه می‌شود. در طول سال‌های اخیر در حوزه تولید علم، برای پژوهشگران و سازمان‌های آنها، آنچه هم عرض با افزایش تعداد کمی تولیدهای علمی به ویژه انتشار مقاله اهمیت یافته، افزایش کیفیت مقالات است. اهمیت کیفیت و کمیت مقالات پژوهشی مورد توجه پژوهشگران و مؤسسات آموزش عالی و تحقیقاتی شاخص‌های کمی مانند تعداد استنادات، ضریب تأثیر مجلات، و مواردی از این قبیل اندازه‌گیری می‌شود؛ شاخص‌های دیگری همچون افزایش شاخص اچ و مانند آنها، برای پژوهشگران و سازمان‌ها اهمیت زیادی دارد. از راه‌های افزایش عملکرد پژوهشی، گسترش استناد مقالات و انتشارات علمی است. چاپ مقاله در یک مجله با ضریب تأثیر مطلوب، ضمانتی برای افزایش استنادات بیشتر به مقاله مورد نظر نیست. جدای از کیفیت یک مقاله، افزایش مشاهده پذیری مقالات و اشتراک علمی میان پژوهشگران موجب افزایش تعداد استنادات مؤثر خواهد شد (بتولی،۱۳۹۶). با این احوال، موضوع استفاده از مقالات و پرتکراری در بهره‌جویی از منابع به منزله یکی از اصول مورد ارجاع در سنجش جنبه‌های کمی محسوب می‌شود.

متن کاوی[1] نخستین بار توسط فلدمن و داگان عنوان شد. متن کاوی زمینه چند رشته‌ای از بازیابی اطلاعات[2]، پردازش زبان طبیعی[3]، آمار، و یادگیری ماشینی[4] است. کاربردهای این حوزه دسته‌بندی، خوشه‌بندی، خلاصه‌سازی و یافتن روابط میان مفاهیم در متون می‌باشد (االهیاری[5] و همکاران، ۲۰۱۷). متن‌کاوی، که به آن تحلیل متن[6] نیز گفته می‌شود، فرایند تبدیل داده‌های متنی غیر ساخت‌یافته به اطلاعات با معنا و عملی است. که از طریق شناسایی «موضوعات[7]»، «الگوها[8]»، و «کلمات کلیدی[9]» مرتبط به کاربران اجازه می‌دهد بدون نیاز به بررسی دستی حجم عظیمی از اطلاعات، دانش، و اطلاعات مفیدی از داده‌های متنی غیر ساخت یافته به دست آورند (جادارياں، ۱۳۹۸). تالیب و همکاران فرایند متن‌کاوی را شامل آماده سازی، پردازش متن، و تحلیل متن می‌دانند (تلیب و دیگران[10]، ۲۰۱۶). هر مرحله به شرح زیر بیان شده است:

با توجه به اینکه متن کاوی به استخراج اطلاعات مفید می‌پردازد، لازم است که معانی واژه و کلید واژه نیز مورد بررسی قرار گیرد.

واژه یا کلمه به مجموعه حروفی که یک واحد را تشکیل داده، اطلاق می‌شود. در دستور زبان فارسی، معمولاً واژه را در نُه نوع دسته بندی می‌کنند: اسم، صفت، عدد، کنایه، فعل، قید، حرف اضافه، حرف ربط، و صوت. «واژه» کوچک‌ترین شکل معنادار از حرفها می‌باشد اگر بتواند به‌تنهایی به‌کار رود. برای نمونه، «–انه» در واژه‌هایی مانند مردانه، زنانه، مهربانانه، دارای معنی ویژه خود است، ولی از آن جا که نمی‌توان آن را به‌تنهایی به‌کار برد، واژه نامیده نمی‌شود. بسیاری از واژه‌ها به بخش‌های کوچکتری بخش‌پذیرند که به آن‌ها تک واژه گفته می‌شود. تکواژ کوچکترین بخش واژه است که در بسیاری از موارد یک واژه مستقل محسوب شده و در برخی موارد نیز واژه به حساب نمی‌آید (اگرای[11] و همکاران، ۱۹۹۳).

کلید واژه به کلماتی گفته می‌شود که از متنی طولانی استخراج شده و در بسامدی بالا و البته غیراتفاقی گویای موضوع‌های مهم آن متن می‌باشد (اسکات و تریبل[12]، ۲۰۰۶). اصطلاح کلیدواژه‌های موضوعی معمولاً برای مقالات و تولیدهای علمی بکار می‌روند. درواقع کلیدواژه‌ها موضوعی کلمات ساده یا مرکبی هستند که بیانگر محتوا هستند. از دیدگاه (اس. ای. ا.[13]) آن‌ها کلماتی هستند که کاربر

---

1. Text mining
2. Information Retrieval
3. Natural Language Processing(NLP)
4. Machine learning
5. Allahyari c
6. Text Analysis
7. Topics
8. Patterns
9. Keywords
10. Talib & etc
11. O'Grady, Dobrovolsky, & Aronoff
12. Scott & Tribble
13. SEO

جست‌وجو کننده را وارد موتورهای جست‌وجو می‌کند. نویسندگان و پدیدآورنده گاه بایستی از کلیدواژه و کلماتی در عنوان استفاده کنند که به خوبی بیانگر محتوای تولید علمی آنها باشد. از طرفی بهره‌جویی از کلیدواژه‌های مناسب نقش مهمی در بهینه سازی وب‌سایت دارد و بایستی با زبان کاربر همخوانی داشته باشد. بدین ترتیب استفاده کنندگان از مطالب می‌توانند به آسانی اطلاعات مورد نیاز خود را بازیابی کرده و پدیدآورندگان بدین ترتیب بازیابی نیاز اطلاعاتی خوانندگان تولیدهای علمی خود را امکان پذیر می‌سازند. به همین دلیل توسعه فهرست کلیدواژه‌های یک پایگاه اطلاعات علمی اولین و مهمترین گام در بهینه‌سازی آن پایگاه اطلاعاتی است.

در پایگاه‌های اطلاعات علمی در دو قسمت مهمترین واژه‌های بیانگر موضوع منبع اطلاعاتی وجود دارد. یکی در عنوان منبع و دیگری در قسمت کلیدواژه‌هایی که توسط نویسنده معرفی می‌شود. کلمات کلیدی نویسنده بلافاصله بعد از چکیده نوشته می‌شود. این کلمات در برگیرنده موضوعات اصلی و فرعی مقاله است. این کلیدواژه‌ها نقش مهمی در نمایه‌سازی مناسب و مؤثر مقالات و مجلات توسط پایگاه‌های اطلاعاتی نمایه کننده مقالات دارند. بر همین اساس شانس دستیابی به مقاله در جست‌وجوی راحت و سریع توسط پژوهشگران افزایش می‌یابد. در نهایت میزان استناد به آن مقاله یا منبع علمی بالا رفته که نهایتاً تعداد استناد نشان دهنده محبوبیت و اعتبار علمی آن منبع است. کلمات کلیدی همیشه در عنوان و چکیده مقاله استفاده می‌شوند.

طبق استانداردهای جاری، نمایه سیاهه نظام‌یافته مدخل‌هایی است که به‌منظور کمک به استفاده‌کنندگان در جایابی اطلاعات سند ساخته می‌شود. در واقع نمایه خط ارتباطی بین منابع اطلاعاتی و استفاده‌کننده یا کاربر است. نمایه‌ها به دو دسته کلی توصیفی و موضوعی طبقه‌بندی می‌شوند. نمایه توصیفی، نمایه‌ای است که اسناد را بر اساس ویژگی‌های اطلاعاتی غیرموضوعی (که معمولاً به اطلاعات کتابشناختی موسوم هستند) فهرست و طبقه‌بندی می‌نماید. نمایه موضوعی، نمایه‌ای است که اسناد را براساس زمینه‌های اطلاعاتی آن که بار موضوعی دارد، فهرست نموده و دسترسی به اسناد را از طریق موضوع امکان‌پذیر می‌سازد. در نمایه‌سازی موضوعی، نخستین گام، شناخت نوع و دامنه مطالب سند و تحلیل موضوعی آن است. هدف از نمایه‌سازی آماده‌کردن اسناد با هدف بازیابی است. در واقع هدف نمایه‌سازی افزایش میزان دسترسی مراجعه‌کنندگان به مطالب موضوعی مورد جست‌وجو است. از این نظر، ممکن است که نمایه‌سازی‌ها با توجه به جامعه استفاده‌کننده و نیازهای اطلاعاتی آنها با هم متفاوت باشند. گزینش مواد از میان مطالب منتشره، توجه به موضوعی خاص، سوگیری زبانی، توجه به مدرک و منبعی خاص و ... از جمله عواملی هستند که می‌توانند بر هدف نمایه‌سازی تأثیر بگذارند.

رایسست (مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری)[1] یک مرکز اطلاعاتی و پژوهشی در شهر شیراز است که با هدف تأمین نیازهای اطلاعاتی و تأمین مدارک علمی مورد نیاز اعضای هیئت علمی، پژوهشگران و دانشجویان کشورهایی که در حوزه جغرافیایی ایران قرار دارند، تأسیس شده است. رایسست، در طول فعالیت خود با تأسیس پایگاه‌های اطلاعاتی مختلف سعی در رفع نیازهای اطلاعاتی جامعه علمی کشور کرده است. این پایگاه حاوی مقالات اکثر نشریات داخلی به زبان فارسی بوده و مقالات آن به صورت تمام متن و بدون محدودیت موضوعی تهیه شده که می‌توان بر اساس گزینه‌های مختلف مانند عنوان، موضوع، نویسنده، و سایر گزینه‌ها در آن جست‌وجو نموده و اطلاعات مورد نیاز را بازیابی نمود (رایسست، ۱۳۹۹).

با توجه به اینکه در زمینه تحلیل کمی روابط موضوعی بین عناوین منابع اطلاعاتی مورد استفاده توسط کاربران رایسست با استفاده از تکنیک متن کاوی پژوهش جامع و کافی انجام نشده است و تحقیق در این زمینه برای ارتقا رایسست مهم است، این پژوهش با هدف تعیین روابط موضوعی در عناوین منابع اطلاعاتی مورد استفاده توسط کاربران رایسست با استفاده از تکنیک متن کاوی انجام شده است. مسئله از آنجا ناشی می‌شود که اطلاعاتی از وضعیت میزان استفاده از موضوعات مورد توجه در منابع اطلاعاتی فارسی در دسترس عموم نبود. از طرفی رایسست یکی از مهمترین پایگاه‌های اطلاعاتی در ایران است که منابع اطلاعات علمی را در اختیار کاربران قرار می‌دهد. بنابراین برای آگاهی از وضعیت موضوعات مورد استفاده از این پایگاه اطلاعات علمی، تصمیم بر آن شد تا بر اساس میزان استفاده و جست‌وجوی کاربران از این پایگاه اطلاعاتی، موضوعات مورد توجه در عنوان منابع اطلاعاتی بررسی شود. در نهایت نیز یک نمایه درهم کرد موضوعی

از این نتایج ارائه شود. در حال حاضر میزان استفاده از کلیدواژه‌های نویسنده براحتی قابل گزارش‌گیری است ولی چگونگی وضعیت استفاده از موضوعات مورد کاربرد در عنوان نیاز به تحقیق جامعی دارد تا با استفاده از نتایج آن بتوان واژه نامه‌های تخصصی و موضوعی کاربر پسند را بهینه سازی نمود. در پژوهش جاری مقصود از منابع اطلاعاتی مورد استفاده به مواردی اطلاق شده که به توسط کاربران و در جهت اعمال بهره‌جویی از آنها مورد جست‌وجو قرار گرفته و در حقیقت با طلب و استفاده پژوهشی کاربر مواجه شده است. بر این قرار در این پژوهش، به بازتاب چگونگی روابط موضوعی در منابع اطلاعاتی کاربران در مرکز رایسست مبادرت شده، تا از طریق شناخت به رفتار و احساس استفاده‌کنندگان و مراجعین در بهره‌جویی از مواد مورد نیاز دسترسی حاصل شود. چنین رویکردی می‌تواند ملاحظات بهره‌جویی از منابع و مقالات را برای سیاست گذاری آینده آشکار نماید. بنابراین، پرسش‌های پژوهش، به شرح ذیل است:

۱. نمای متنی موضوعات پر استفاده در عنوان‌های منابع اطلاعاتی مورد استفاده توسط کاربران رایسست با استفاده از تکنیک متن کاوی چگونه است؟

۲. توزیع موضوعات پر استفاده در عنوان‌های منابع اطلاعاتی مورد استفاده توسط کاربران رایسست با استفاده از تکنیک متن کاوی چگونه است؟

۳. همبستگی تکرار موضوعات پر استفاده در عنوان‌های منابع اطلاعاتی مورداستفاده توسط کاربران رایسست با استفاده از تکنیک متن کاوی چگونه است؟

نتایج پژوهش سهرابی و غفاری (۱۳۹۸) مربوط به خوشه‌بندی سلسله‌مراتبی به روش «وارد» منجر به شکل‌گیری پنج خوشه در این حوزه گردید که از مهم‌ترین خوشه‌ها می‌توان به «علم ارتباطات»، «دسترس‌پذیری علم» و «سنجش علمی» اشاره نمود. نتایج نمودار راهبردی نشان داد که خوشه «دسترس‌پذیری علم» جزء خوشه‌های بالغ و مرکزی به حساب می‌آید و نقش محوری و اساسی در حوزهٔ ارتباطات علمی دارد. همچنین خوشه «سنجش علمی» جزء خوشه‌های مرکزی ولی توسعه‌نیافته و به موضوعات در حال ظهور مثل «تحویل مدرک» و «دسترسی آزاد به انتشارات» اشاره شد. بر اساس نتایج پژوهش سهیلی و همکاران (۱۳۹۷) مربوط به نمودار راهبردی، مباحث علم‌سنجی، بهترین جایگاه را در پژوهش‌های علم اطلاعات و دانش‌شناسی ایران دارند و عناوینی نظیر رابط کاربر، معماری اطلاعات، موتورهای جست‌وجو، کتابخانه دیجیتالی، ابرداده، جست‌وجوی اطلاعات، حفاظت اطلاعات، مدیریت دانش، هستی‌شناسی، مصوّرسازی، و شبکه‌های اجتماعی جزء موضوعات نوظهور در مطالعات علم اطلاعات و دانش‌شناسی ایران هستند. نتایج پژوهش رحیمی و همکاران (۱۳۹۷) به ارائه مدل موضوعی احتمالاتی مبتنی بر روابط محلّی واژگان در پنجره‌های هم پوشان منجر شد که روش پیشنهادی، موضوعات منسجم‌تری را تولید و در کاربرد خوشه‌بندی اسناد، دقیق‌تر از دو مدل آل. دی آ. و بی. تی. ام.[1] عمل می‌کند. مسعودی و راحتی قوچانی (۱۳۹۴) به ارائه مدلی برای رفع ابهام از واژگان مبهم فارسی به وسیلهٔ استخراج ویژگی‌های جدید پرداختند. صدیقی (۱۳۹۳) در پژوهشی نشان داد که بر اساس روش تجزیه و تحلیل هم‌رخدادی واژگان می‌توان موضوعات علمی را استخراج و ارتباط میان آنها را به صورت مستقیم از محتوای موضوعی کشف کرد. چن و همکاران ترکیبی از الگوریتم‌های مدل‌سازی موضوع و کتاب سنجی چالش‌های جدیدی در تفسیر و درک نتیجه مدل‌سازی موضوع ایجاد و رویکرد شناسایی رابطه موضوع برای مدل‌سازی کمی روابط بین موضوعات ارائه کرده است (چن[2]، ۲۰۱۹). مانتیلا و همکاران به سیر تکامل و تجزیه و تحلیل احساسات در بررسی مباحث تحقیق، مکان‌ها، و مقالات برتر با توجه به استناد به دیگر مقالات منتشره مبادرت کرده و نشان دادند که در سال‌های اخیر، تجزیه و تحلیل احساسات به تجزیه و تحلیل بررسی محصولات پیوسته به متن رسانه‌های اجتماعی از توییتر و فیس بوک تغییر یافته است (مانتیلا و همکاران[3]، ۲۰۱۹). بسیاری از موضوعات فراتر از بررسی محصولات مانند بازارهای سهام، انتخابات، بلایای طبیعی، پزشکی، مهندسی نرم افزار، و حمله سایبری به استفاده از تجزیه و تحلیل احساسات مبادرت ورزیده‌اند. میلر پروژه‌های متن کاوی دیجیتال در علوم انسانی را با قابلیت‌های تجزیه و تحلیل محتوا با ابزارهای ویانت[4]

---

1. BTM و LDA
2. Chen
3. Mantyla & etc
4. Voyant

ارزیابی کردند و نشان دادند که ابزارهای ویانت، نرم افزاری با دسترسی آزاد بوده که متن کاوی کاربرپسند با مستندات خوب دارد. گروسی و مانتیلا در مطالعه‌ای با تأکید بر جنبه‌های کتابشناسی به نقل قول‌ها، مباحثات تحقیق، و کشورهای فعال در حوزه مهندسی نرم افزار مبادرت کردند و موضوعات داغ تحقیق در مهندسی نرم افزار مانند خدمات وب، موبایل و فضای ابری، مطالعات (موردی) در حوزه‌های صنعتی، کد منبع باز، و آزمون نسل‌ها را شناسایی کردند (گروسی و مانتیلا[1]، ۲۰۱۶). لیدسدورف و نرقس به ترسیم نقشه‌های هم واژگانی و مدل‌سازی موضوعی با الگوبرداری از "داده‌های بزرگ" و"مدل‌سازی موضوعی" به گزینه‌ای جذاب برای نقشه برداری هم واژگانی در مدت زمان معین از نظر هم رخدادی و هم فقدانی با استفاده از تکنیک‌های شبکه‌ای پرداختند (لیدسدورف و نرقس[2]، ۲۰۱۶).

ورونتسو و همکاران[3] (۲۰۱۵) به تنظیم مدل‌سازی موضوع‌های چند مفهومی منظم از مجموعه‌های بزرگ کتابخانه با منبع باز"، پروژه منبع باز بیگارتم[4] را برای مدل‌سازی منظم چند موضوع از مجموعه‌های بزرگ پرداختند (ورونتسو و همکاران[5]، ۲۰۱۹). جا-هیون و مین نیز، روند پژوهش در علوم کتابداری و اطلاع رسانی در کره با استفاده از مدل‌سازی موضوعی را نشان دادند (جاهین و مین[6]، ۲۰۱۳).

از بررسی پیشینه‌های داخل و خارج از ایران عیان شد، در طول چند سال گذشته توجه به تحلیل هم رخدادی، جنبه‌های پر کاربرد تولیدهای علم، زمینه‌های متن کاوی، و ساخت مدل برای موضوع‌های مختلف کتابداری، ادبیات، زبان شناسی و حتی حوزه‌های مهندسی فزونی گرفته، این کوشش از سویی به دلیل توسعه فناوری در مباحث آموزشی و از جنبه‌ای دیگر به توانمندی گسترده پژوهشگران مربوط بوده است. هر چند زمینه مورد استفاده و پر استفاده، به روندی نو در پژوهش‌های ایرانی منتسب است، ولی در این خصوص به ویژه بر سامانه‌های خاص کار جدی و قابلی صورت نپذیرفته و پژوهش حاضر سعی دارد با استفاده از تکنیک متن کاوی به چالش این مطلب در مرکز رایسیست اهتمام آورد.

روش پژوهش

روش پژوهش حاضر متن کاوی است. متن‌کاوی، به داده‌کاوی بر روی متن اشاره دارد. همچنین به عنوان تحلیل متن نیز شناخته می‌شود که منظور از آن فرایند استخراج اطلاعات با کیفیت از متن است. متن‌کاوی یا تحلیل متن فرایند تبدیل داده‌های متنی غیر ساخت‌یافته به اطلاعات با معنا و عملی است که از طریق شناسایی «موضوعات»، «الگوها»، و «کلمات کلیدی» به کاربران امکان می‌دهد بدون نیاز به بررسی دستی حجم زیادی از اطلاعات و دانش، اطلاعات مفیدی از داده‌های متنی غیرساخت‌یافته به دست آورند. غالباً مفاهیم متن کاوی و تحلیل متن مترادف هستند. مفهوم تحلیل کمی متن، تا حدی خاص‌تر است. به اختصار، مدل‌های متن‌کاوی و مدل‌های تحلیل کمی متن سعی دارند مسئله‌ای یکسان (تحلیل خودکار داده‌های متنی خام) را به وسیله تکنیک‌های متفاوت حل کنند. تکنیک‌های متن کاوی، اطلاعات مرتبط درون یک متن را شناسایی کرده و در این روند نتایج کیفی تولید می‌کنند. در این پژوهش از تکنیک‌های کمی تحلیل متن به عنوان روش متن کاوی عنوان استفاده شده است. روش‌های مورد استفاده برای متن کاوی یا تحلیل کمی عناوین به شرح زیر است:

روش مبتنی بر تناوب کلمات[7]: از روش‌های مبتنی بر تناوب کلمه برای شناسایی متناوب‌ترین لغات یا مفاهیم موجود در مجموعه‌ای از داده‌های متنی استفاده می‌شود.

---

1. Garousi & Mantyla
2. Leydesdorff & Nerghes
3. Vorontsov & etc.
4. http://bigartm.org
5. Vorontsov & etc
6. Ja-Hyun & Min
7. Word Frequency

روش‌های مبتنی بر باهم‌گذاری یا هم‌اتفاقی کلمات[1]: اصطلاح باهم‌گذاری یا هم‌اتفاقی کلمات، به دنباله‌ای از کلمات یا مفاهیم اطلاق می‌شود که معمولاً در یک داده متنی در کنار هم‌دیگر (همسایگی یکدیگر) ظاهر می‌شوند. شایع‌ترین نوع کلمات یا مفاهیم باهم‌گذاری (هم‌اتفاقی)، «دو کلمه‌ای‌ها» و «سه کلمه‌ای‌ها» هستند. دو کلمه‌ای‌ها، عباراتی دو کلمه‌ای هستند که معمولاً در کنار یکدیگر اتفاق می‌افتند.

روش‌های مبتنی بر کشف لغات[2]: اصطلاح کشف لغات، به فهرستی از لغات یا مفاهیم موجود در یک سند به همراه مشخصه محل ظاهر شدن آنها اطلاق می‌شود. از روش‌های مبتنی بر راهنمای لغات، برای بازشناسی یک «زمینه محتوایی» خاص استفاده می‌شود که یک کلمه یا مجموعه‌ای از کلمات در آن ظاهر شده‌اند (متن کاوی-به زبان ساده، ۱۳۹۸، بازیابی شده در ۶ مرداد ۱۳۹۹). در این پژوهش از روش تحلیل کمی و از تکنیک‌های داده کاوی شامل متن کاوی به شرح بالا استفاده شد. جامعه آماری پژوهش حاضر مجموعه لاگ فایل کاربران رایسست شامل ۱۴۶۰۴۴ رکورد حاصل از خروجی گزارش‌گیری از این سیستم است. در این پژوهش برای نمونه‌گیری از روش سرشماری استفاده شد و کلیه داده‌های حاصل از گزارش‌گیری توسط سیستم رایسست بررسی گردید. بطور کلی نمونه‌گیری این منابع با استفاده از روش سرشماری بررسی گردید. نمونه کاربران هم بصورت هدفمند و کسانی که از منابع در دوره زمانی تحقیق استفاده کرده بودند بر اساس گزارش رایسست در نظر گرفته شد.

در این پژوهش از پایگاه‌های موجود در مرکز منطقه‌ای اطلاع رسانی علوم و فناوری شیراز با حمایت مسئولان بر اساس مکاتبه شماره ۹۸/۳۱۳۹/د در تاریخ ۱۳۹۸/۱۱/۱۹ در بازه زمانی دو ساله استفاده شده است.

در این پژوهش از ابزار ویانت[3] به منظور تحلیل متن استفاده شد، ویانت یک محیط خواندن و تفسیر علمی متون مبتنی بر وب است که به منظور تسهیل خواندن و شیوه‌های تفسیری برای دانشجویان و دانشمندان علوم دیجیتال و همچنین برای عموم مردم طراحی شده است. از این ابزار می‌توان برای تحلیل رایانه‌ای، جست‌وجو و مطالعه متونی که در وب یا در رایانه خود ویرایش شده، افزودن قابلیت‌ها به مجموعه‌های پیوسته، ژورنال‌ها، وبلاگ‌ها، یا وب سایت‌ها و نیز افزودن شواهد تعاملی به مقالات که به صورت پیوسته منتشر شده است و جهت توسعه ابزارها با استفاده از قابلیت و کد استفاده کرد (عاصمی، ۲۰۲۰).

پاکسازی و نرمال‌سازی داده‌ها با استفاده از نرم افزار پایتون[4] به قرار ذیل انجام شد: دیتاست پژوهش جاری مستخرج از گزارش اطلاعات منابع به زبان‌های مختلف و مورد استفاده کاربران پایگاه اطلاعاتی رایسست در محدوده زمانی دوساله، ۱۳۹۸/۱۱/۱۹-۱۳۹۶/۱۱/۱۹ است. به منظور تحلیل داده‌ها از سه فرمت برای پاکسازی داده‌ها از سه نوع فرمت محبوب (سی. اس. وی و جی. اس. ا. ان. و تی. ایکس. تی.)[5] داده استفاده شد. پاکسازی داده‌های اضافی و زاید با استفاده از نامپی[6] و پانداس[7] در پایتون انجام گرفت. در ابتدا، کار با انتقال (ایمپورت) ماژول‌های مورد نیاز (یعنی دو کتابخانه یاد شده) آغاز گردید و عملیات زیر انجام شد:

حذف ستون‌های غیر لازم از دیتافریم؛

- تغییر اندیس یک دیتافریم؛
- استفاده از متدهای (اس. تی. آر.[8]) برای پاک کردن ستون‌ها؛
- استفاده از تابع (دیتافریم اپلای مپ[9]) برای پاک‌سازی کل مجموعه داده به صورت مؤلفه به مؤلفه؛
- نام‌گذاری مجدد ستون‌ها به منظور ایجاد مجموعه برچسب‌های قابل تشخیص‌تر؛
- گذر کردن از سطرهای غیر لازم در فایل سی. اس. وی.؛

---

1. Word Collocation
2. Concordance
3. Voyant
4. Python
5. CSV, JSON, TXT
6. NumPy
7. NumPy
8. STR
9. DataFrame.applymap

- تبدیل فایل سی. اس. وی. به تی. ایکس. تی.

بطور کلی اطلاعات غیر لازم از مجموعه عنوان‌های منابع مورد استفاده و گزارش شده در دیتاست، با استفاده از تابع (دراپ[1]) حذف گردید. تنظیم اندیس برای مجموعه داده به نوعی که آیتم‌های آن به سادگی قابل ارجاع بوده، انجام شد. علاوه بر این، فیلدهای آبجکت[2] با اکسسور (اس. تی. آر.) پاک سازی گردید و در نهایت کل مجموعه داده با استفاده از متد (اپلای مپ) پاکسازی نهایی گردید. در پایان، گذر از سطرها در فایل (سی. اس. وی.) و تبدیل به تکست انجام گرفت. پاکسازی داده‌ها در ده مرحله در پایتون انجام گرفت. در هر محله با تست داده‌های پاکسازی شده در نرم افزار ویانت، داده‌های زاید آشکار و مجدداً عملیات پاکسازی و نهایتاً نرمال سازی داده‌ها انجام شد. پس از مرحله دهم داده‌ها آماده تجزیه و تحلیل در نرم افزار ویانت گردید. در پاکسازی داده‌ها عملیاتی نظیر یکسان سازی کاراکترهای فارسی و عربی (به عنوان مثال کنترل "ی" و "ک" به صورت تایپ فارسی و عربی)، حذف فاصله‌های اضافی بین کلمات، حذف فاصله اضافی بین پارگراف ها (اینتر)، حذف نشانه‌ها، حذف استاپ وردها، حذف حروف اضافه، حذف حروف ربط، حذف کلمات انگلیسی (با توجه به اینکه میزان تکرار آنها بسیار کم بود)، حذف کلمات مبهم و بی‌معنی مانند بررسی، تعیین و...، یکسان سازی اشکال مختلف کلمات موضوعی، جایگزینی مترادف‌ها، بررسی کلمات موضوعی هم نویسه و هم آوا، یکسان سازی اشکال جمع و مفرد و موارد دیگر انجام پذیرفت.

در مرحله پاکسازی و نرمال سازی داده‌ها بعضاً ترکیب شده‌اند. به عنوان مثال کلمه "ایران" و " ایرانی"، با کلمه "ایران" نشان گذاری شده است. کلمه "آموزش" و "آموزشی" با کلمه "آموزش" نشان گذاری شده است. همچنین تمام عناوینی که کلمه "رشد" و "توسعه" در آنها بود، تحت عبارت تک کلمه‌ای با کاراکترهای متصل "جنبه‌های رشدوتوسعه" در نظر گرفته شد. به همین ترتیب این روند در مورد "نظام"، "سیستم" و "سیستمی" بکار برده شد. کلمه (منظور کارکترهای متصل به هم است) "جنبه‌های اجتماعی" نیز در مورد تمام عناوینی صادق است که در آنها کلمه "اجتماعی" بکار رفته است. کلمات "سازمان" و "سازمانی" هم به همین ترتیب "سازمان سازمانی" در نظر گرفته شد. چند عنوان هم کلمه "ارگان" را به کار برده بودند که آنها نیز جایگزینی و به حساب آمده است. تمام عناوینی که در آن کلمه موضوعی دانشگاه، دانشکده، موسسه یا مؤسسات آموزش عالی، آموزشکده عالی بکار رفته بود با در نظر گرفتن یک تکرار برای عناوینی که کلمات دانشگاه و دانشکده با هم بکار رفته شده بود به صورت مؤسسات آموزش عالی دانشگاه. دانشکده" در نظر گرفته شد تا در تحلیل تمام موارد مورد نظر محسوب شود. همچنین کلمات "دانش آموز" و"دانش آموزان" نیز بصورت "دانش آموز" در نظر گرفته شد. این عمل برای "کودک" و "کودکان"، "دختر"، "زن"، "زنان" و "دختران"، "دانشجو" و "دانشجویان"، "سلامت"، "سلامتی" و "بهداشت" و "بهداشت" همراه با کنترل واژه‌ها، نیز به همین صورت اعمال شد. در مورد کلمه موضوعی "اسلام" و "اسلامی" نیز کنترل اعمال گردید و عباراتی که در آن کلمه "اسلام" موضوع محسوب نمی‌شد مانند "حجت الاسلام" و کلماتی مانند "جمهوری اسلامی ایران" و "دانشگاه آزاد اسلامی" در نظر گرفته نشد. نهایتاً فرمت تکست شامل یک سند ۶۶۰ صفحه‌ای با حدود ۱۱۴۰۰۰۰ هزار کلمه مورد بررسی قرار گرفت. برای پاسخ به برخی از سؤال‌های پژوهش نیز از فرمت (جی. اس. ا. ان.[3]) استفاده گردید و داده‌ها با استفاده از این فرمت در ویانت مورد تجزیه و تحلیل قرار گرفت.

یافته‌ها

-تعیین چگونگی نمای متنی موضوعات پر استفاده در عنوان‌های منابع اطلاعاتی توسط کاربران رایسست با استفاده از تکنیک متن کاوی

در پاسخ به این سؤال (توده) موضوعات پرتکرار بازیابی شده در عنوان مورد تحلیل قرار گرفت. جدول شماره ۱ بیانگر نمای کورپوس ۲۱ موضوعی است که بیش از ۲۰۰۰ بار در پایگاه اطلاعاتی رایسست در فاصله زمانی دوساله (۱۳۹۸/۱۱/۱۹- ۱۳۹۶/۱۱/۱۹) مورد استفاده قرار

---

1. Drop
2. Object
3. JSON

گرفته‌اند. حد ۲۰۰۰ بار به این دلیل تعیین گردید که میزان استفاده از موضوعات زیر ۲۰۰۰ بار بسیار زیاد شده و از محدوده این بخش خارج بود. به همین دلیل حد ۲۰۰۰ بار به عنوان یک محدودیت در ارائه یافته‌ها تعیین گردید، تا داده‌های جاری به نحو مناسب در ارائه یافته‌های قابل کنترل و بررسی شوند. جدول [1] نمایش فرکانس‌های (تکرارهای) موضوعات پراستفاده در کل عنوان‌های منابع مورد استفاده در پایگاه مورد نظر را نشان می‌دهد. این جدول چند ستون داده را نشان می‌دهد و اصطلاحات مورد نیاز برای خواننده به قرار ذیل است:
در تصویرهای بعد دو ستون با عنوان ترند[1] و مرتبط[2] نیز ارائه شده است که ترندها نمایانگر روند یک موضوع است. روندها یک نمودار اسپارک لاین[3] هستند که توزیع فرکانس‌های نسبی را در بین عناوین منابع در مجموعه دیتاست نشان می‌دهد. مرتبط نیز فرکانس نسبی موضوعات پراستفاده در مجموعه عنوان‌های منابع در هر یک میلیون کلمه است.

جدول ۱. کورپوس کلمات موضوعی در عنوان‌های منابع مورد استفاده در پایگاه اطلاعاتی رایسست (با بیش از ۲۰۰۰ بار تکرار) در بازه زمانی (۱۳۹۸- ۱۳۹۶)

ردیف	کلید واژه	فراوانی موضوع در کل مجموعه عنوان‌های منابع
۱.	ایران	۱۰۰۵۴
۲.	آموزش	۷۶۲۲
۳.	جنبه‌های رشد و توسعه	۵۹۴۸
۴.	نظام. سیستم	۵۳۶۴
۵.	جنبه‌های اجتماعی	۵۱۹۶
۶.	سازمان. سازمانی	۵۱۳۵
۷.	مدیریت	۵۰۳۸
۸.	مؤسسات آموزش. عالی. دانشگاه. دانشکده	۴۰۲۵
۹.	دانش آموز	۳۷۹۷
۱۰.	اسلام و اسلامی	۳۷۰۹
۱۱.	کودک و کودکان	۳۶۲۴
۱۲.	تهران	۳۳۸۴
۱۳.	زن. دختر	۳۳۸۱
۱۴.	معماری	۳۲۸۲
۱۵.	اطلاعات	۳۲۶۱
۱۶.	تولید	۲۷۸۶
۱۷.	زندگی	۲۵۳۸
۱۸.	زبان	۲۴۹۰
۱۹.	دانشجو	۲۴۲۱
۲۰.	یادگیری	۲۳۲۰
۲۱.	سلامت بهداشت	۲۲۴۶

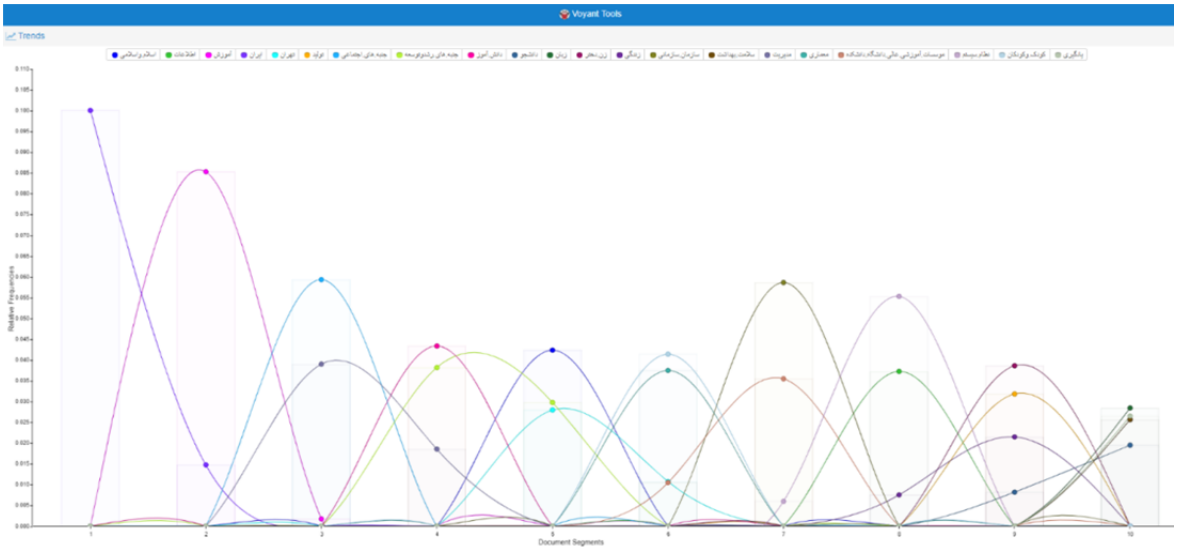
[1]. Trend
[2] . Relative
[3]. Sparkline

براساس جدول شماره ۱، بیشترین کلمه موضوعی پر تکرار در بین کلیه عنوان‌های منابع مورد استفاده پایگاه رایسست، کلمه "ایران" با ۱۰۰۰۵۴ بار تکرار است. کلمه ایران مربوطه به تمام منابعی می‌شود که در عنوان آنها کلمه "ایران" یا "ایرانی" بکار برده شده است. بر اساس جدول ۱، کلمه "آموزش" ۷۶۲۲ بار، کلمه "جنبه‌های رشد و توسعه" ۵۹۴۸ بار، کلمه "نظام سیستم" ۵۳۶۴ بار، و کلمه "جنبه‌های اجتماعی" ۵۱۹۶ بار تکرار شده است؛ بقیه موارد از کورپوس کلمات موضوعی در جدول ۱ نشان داده شده است.

شکل ۱. ابر موضوعی کلمات پرتکرار عنوان‌های منابع مورد استفاده در پایگاه اطلاعاتی رایسست (با بیش از ۲۰۰۰ بار تکرار) در بازه زمانی (۱۳۹۸- ۱۳۹۶)

یکی از تکنیک‌های نمایش نمای متنی کلمات موضوعی بر اساس میزان تکرار در عناوین منابع استفاده از ابرِ کلمات[1] بود. شکل ۱، ۲۱ ابر موضوع پرتکرار که بیانگر پراستفاده‌ترین منابع پایگاه اطلاعاتی رایسست را نشان می‌دهد. شکل ۱ نشان دهنده کلمات موضوعی با تکرار بالا در مجموعه عناوین منابع است. نمودار ابر کلمات موضوعی به گونه‌ای است که اصطلاحاتی که بیشترین اتفاق یا حضور در عنوان‌ها را داشته به صورت مرکزی قرار گرفته و آنها را در مقیاس بزرگتر به نمایش می‌گذارد. در شکل ۱، طبق الگوریتم آن نشان می‌دهد تلاش شده کلمات موضوعی با بیشترین تکرار را تا حد امکان به مرکز نزدیک کند؛ قابل توجه است که رنگ کلمات و موقعیت آنها بصورت مطلق نیست. به عنوان مثال اگر در نرم افزار اندازه پنجره را تغییر دهیم یا صفحه را مجدد بارگذاری نماییم ممکن است کلمات در مکان دیگری ظاهر شوند.

[1].Word Cloud



شکل ۲. روند ۲۱ موضوع پرتکرار عنوان‌های منابع مورد استفاده در پایگاه اطلاعاتی رایسست (با بیش از ۲۰۰۰ بار تکرار) نسبت به یکدیگر در مجموعه عنوان‌های منابع پر استفاده در بازه زمانی (۱۳۹۸- ۱۳۹۶)

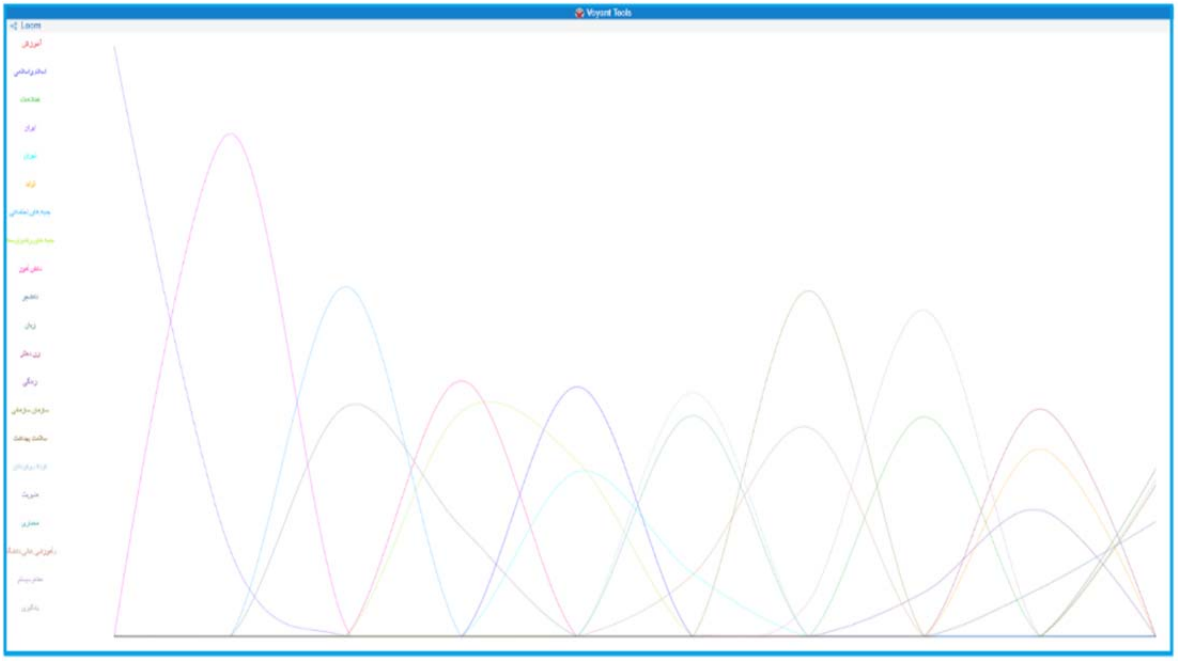
شکل ۲، یک نمودار خطی است که روند ۲۱ موضوع پرتکرار (بیش از ۲۰۰۰ بار) نسبت به یکدیگر را در مجموعه عناوین منابع پر استفاده نشان می‌دهد. این روند کلمات موضوعی پر استفاده برحسب توزیع وقوع یک کلمه در مجموعه عناوین منابع پر استفاده پایگاه رایسست و نسبت به یکدیگر را نشان می‌دهد. روند هر موضوع در شکل شماره ۲ با یک رنگ نشان داده شده است؛ مثلاً از سمت چپ شکل، نمودار کلمه "ایران" به رنگی آبی قابل مشاهده می‌باشد، که نسبت به دیگر عنوان‌ها گستره وسیعی را به خود تخصیص داده است بر همین اساس می‌توان توسط راهنمای بالای نمودار رنگ موضوعات را با رنگ خطوط در شکل انطباق داد و گستره ۲۱ موضوع را مشاهده کرد. در محور Y فرکانس نسبی موضوعات نسبت به یکدیگر نمایش داده شده است.

شکل ۳. روند ۱۶۰ کلمه موضوعی پرتکرار عنون‌های منابع مورد استفاده در پایگاه اطلاعاتی رایسست (با بیش از ۲۰۰۰ بار تکرار) نسبت به یکدیگر در مجموعه عنوان‌های منابع پر استفاده در بازه زمانی (۱۳۹۸- ۱۳۹۶)

در شکل شماره ۳، نرم افزار ویانت با دقت و جامعیت مبادرت به ترسیم روند ۱۶۰ موضوع پر تکرار نسبت به یکدیگر را در مجموعه عنوان‌های منابع پر استفاده از پایگاه اطلاعاتی رایسست نموده و به واقع گستره موضوعی کاربران را برای پژوهشگر با توجه به بیان موضوعات در راهنمای نمودار آورده است این یافته جانبی پژوهش به بیان میزان عملکرد اجرایی از نرم افزار مزبور و تأثیر هر یک از موضوعات نسبت به ۱۶۰ کلمه موضوعی در نمره Z و نسبت نمره Z بیان می کند.

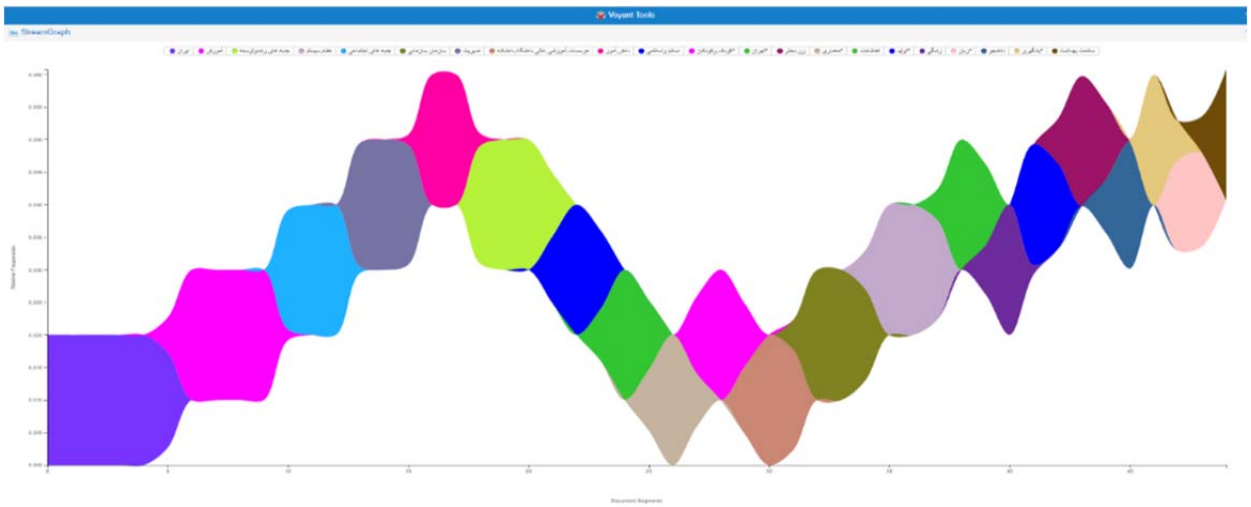
- تعیین چگونگی توزیع موضوعات پر استفاده در عناوین منابع اطلاعاتی توسط کاربران رایسست با استفاده از تکنیک متن کاوی

در این بخش به تعیین چگونگی توزیع موضوعات پر استفاده در عنوان‌های منابع اطلاعاتی که به توسط کاربران رایسست استفاده شده مبادرت می‌شود. به منظور این که نمای مناسبی از موضوعات در تصور خواننده ظاهر شود این بخش را با دو شکل متفاوت همراه با تکنیک‌های بصری آغاز می‌نماییم.



شکل ۴. دورنمای لوم از چگونگی توزیع کلمات موضوعی با فرکانس بالا در عناوین منابع پر استفاده از پایگاه رایسست

براساس شکل شماره ۴، دورنمای لوم از توزیع کلمات موضوعی با تکرار (فرکانس) بالا در عنوان‌های منابع پر استفاده از پایگاه رایسست را نشان می‌دهد. شکل شماره ۴، هر موضوع را با رنگ خاص خود بر حسب میزان توزیع در مجموع عنوان‌ها نشان می‌دهد. در سمت چپ شکل کلمات موضوعی بر حسب حروف الفبا مرتب شده که رنگ هر کلمه در شکل با همان رنگ نمایش داده شده است. کلمه موضوعی "ایران" با رنگ بنفش دارای بیشترین تکرار است که اگر خط بنفش رنگ در شکل دنبال شود چگونگی توزیع آن نسبت به دیگر رنگ‌ها یا موضوعات مشخص شده است. چنانچه رنگ بنفش به منزله موضوع "ایران" در نمودار جاری دنبال شود، به راحتی می‌توان میزان تکرار این موضوع را نسبت به موضوعات دیگر مشاهده نمود به نوعی که کارایی این گونه منابع در پایگاه رایسست قابل توجه است.



شکل ۵. استریم گراف کلمات موضوعی پرتکرار و موقعیت آنها نسبت به یکدیگر در عنوان‌های منابع اطلاعاتی پر استفاده توسط کاربران رایسست

براساس شکل شماره ۵، نمودار جریان کلمات موضوعی پرتکرار نسبت به یکدیگر در مجموعه عنوان‌های منابع پر استفاده را نشان می‌دهد استریم گراف[1] مربوطه، تغییر فرکانس ۲۱ کلمه موضوعی پر تکرار را در مجموعه عنوان‌ها بیان می‌کند. هر سطح زیر نمودار که با رنگ‌های متفاوت درج شده، یک کلمه موضوعی پر تکرار را نشان می‌دهد؛ مثلاً نخستین سطح زیر نمودار که با رنگ بنفش نشان داده شده بازتاب کلمه "ایران" است.

براساس جدول شماره ۲، چگونگی توزیع هر کلمه موضوعی از ۲۱ کلمه موضوعی پرتکرار را در عناوین منابع اطلاعاتی استفاده شده توسط کاربران رایسست با بهره‌جویی از تکنیک‌های متن کاوی نشان می‌دهد. طبق جدول شماره ۴-۲، فراوانی مطلق کلمه "ایران" ۱۰۰۵۴ بار و فراوانی نسبی آن ۱۰۹۷۴/۷۳۴ بار؛ کلمه "آموزش" با فراوانی مطلق ۷۶۲۲ بار و فراوانی نسبی ۸۳۱۶۷/۸۴ بار؛ کلمه "جنبه‌های رشد و توسعه"، با فراوانی مطلق ۵۹۴۸ بار و فراوانی نسبی ۶۴۹۰۱/۹۰۲ بار؛ کلمه"نظام سیستم" با فراوانی مطلق ۵۳۶۴ بار و فراوانی نسبی ۵۸۵۲۹/۵۶ بار و کلمه "جنبه‌های اجتماعی" با فراوانی مطلق ۵۱۹۶ بار و فراوانی نسبی ۵۶۶۹۶/۴۱۸ بار در بین ۲۱ کلمه موضوعی تکرار شده است.

جدول ۲. توزیع فراوانی مطلق و فراوانی نسبی ۲۱ کلمه موضوعی پرتکرار در عناوین منابع اطلاعاتی

Docindex	Term	Raw Frequency[2]	Relative Frequency[3]	Z Score[4]	Z Score Ratio	TF-IDF[5]
0	ایران	۱۰۰۰۵۴	۱۰۹۷۰۴/۷۳	/۷۹۶	-۴۵۶۰/۹۸۰۵	۰/۰
0	آموزش	۷۶۲۲	۸۳۱۷۶/۸۴	۷۶۱۹/۷۹۵۴	-۳۴۵۷/۴۶۵۳	۰/۰
0	جنبه. های. رشدوتوسعه	۵۹۴۸	۶۴۹۰۱/۹۰۲	۵۹۴۵/۷۹۵۴	-۲۶۹۷/۸۹۱۶	۰/۰

[1]. Stream Graph

۲. شمارش حسابی تعداد یک ویژگی زبانی (یک کلمه، یک ساختار و غیره) مستقیم‌ترین داده‌های کمی ارائه شده توسط یک پیکره.

۳. توزیع فرکانس نسبی نسبت تعداد کل مشاهدات مرتبط با هر مقدار یا کلاس مقادیر را نشان می‌دهد و به توزیع احتمال مربوط می‌شود که به طور گسترده در آمار استفاده می‌شود.

۴. امتیاز z را می‌توان روی یک منحنی توزیع نرمال قرار داد. همچنین، A Z-score یک اندازه گیری عددی است که رابطه یک مقدار را با میانگین گروهی از مقادیر توصیف می‌کند.

۵. TF-IDF (فرکانس معکوس-سند فرکانس)

0	نظام سیستم	۵۳۶۴	۵۸۵۲۹/۵۶	۵۳۶۱/۷۹۵۴	-۲۴۳۲/۹۰۲۸	۰/۰
0	جنبه‌های اجتماعی	۵۱۹۶	۵۶۶۹۶/۴۱۸	۵۱۹۳/۷۹۵۴	-۲۳۵۶/۶۷۳۳	۰/۰
0	سازمان سازمانی	۵۱۳۵	۵۶۰۳۰/۸۱۲	۵۱۳۲/۷۹۵۴	-۲۳۲۸/۹۹۴۶	۰/۰
0	مدیریت	۵۰۳۸	۵۴۹۷۲/۳۹۵	۵۰۳۵/۷۹۵۴	-۲۲۸۴/۹۸۱۲	۰/۰
0	دانشگاه دانشکده	۴۰۲۵	۴۳۹۱۸/۹۹۲	۴۰۲۲/۷۹۵۷	-۱۸۲۵/۳۳۴۷	۰/۰
0	مؤسسات آموزشی عالی	۴۰۲۵	۴۳۹۱۸/۹۹۲	۴۰۲۲/۷۹۵۷	-۱۸۲۵/۳۳۴۷	۰/۰
0	دانش آموز	۳۷۹۷	۴۱۴۳۱/۱۶	۳۷۹۴/۷۹۵۷	-۱۷۲۱/۸۸۰۱	۰/۰
0	اسلام واسلامی	۳۷۰۹	۴۰۴۷۰/۹۴	۳۷۰۶/۷۹۵۷	-۱۶۸۱/۹۵۰۳	۰/۰
0	کودک وکودکان	۳۶۲۴	۳۹۵۴۳/۴۶	۳۶۲۱/۷۹۵۷	-۱۶۴۳/۳۸۱۸	۰/۰
0	تهران	۳۳۸۴	۳۶۹۲۴/۶۹	۳۳۸۱/۷۹۵۷	-۱۵۳۴/۴۸۲۳	۰/۰
0	زن دختر	۳۳۸۱	۳۶۸۹۱/۹۵۳	۳۳۷۸/۷۹۵۷	-۱۵۳۳/۱۲۱۱	۰/۰
0	معماری	۳۲۸۲	۳۵۸۱۱/۷۱	۳۲۷۹/۷۹۵۷	-۱۴۸۸/۲۰۰۱	۰/۰
0	اطلاعات	۳۲۶۱	۳۵۵۸۲/۵۷	۳۲۵۸/۷۹۵۷	-۱۴۷۸/۶۷۱۴	۰/۰
0	تولید	۲۷۸۶	۳۰۳۹۱/۵۸۲	۲۷۸۳/۷۹۵۷	-۱۲۶۳/۱۴۱۱	۰/۰
0	زندگی	۲۵۳۸	۲۷۶۹۳/۵۱۶	۲۵۳۵/۷۹۵۷	-۱۱۵۰/۶۱۱۷	۰/۰
0	زبان	۲۴۹۰	۲۷۱۶۹/۷۶۲	۲۴۸۷/۷۹۵۹	-۱۱۲۸/۸۳۱۸	۰/۰
0	دانشجو	۲۴۲۱	۲۱۶۸۱۶/۸۶۳	۲۴۱۸/۷۹۵۷	-۱۰۹۷/۵۲۳۲	۰/۰
0	یادگیری	۲۳۲۰	۲۵۳۱۴/۷۹۹	۲۳۱۷/۷۹۵۷	-۱۰۵۱/۶۹۴۷	۰/۰
0	سلامت. بهداشت	۲۲۴۶	۲۴۵۰۷/۳۴۴	۲۲۴۳/۷۹۵۷	-۱۰۱۸/۱۱۷۴	۰/۰

براساس جدول شماره ۳، چگونگی توزیع ۱۶۰ کلمه موضوعی پر استفاده به انتخاب نرم افزار ویانت و بر اساس پیش فرض‌های تعریف شده در این نرم افزار، در عنوان‌های منابع اطلاعاتی توسط کاربران رایسست را با استفاده از تکنیک‌های متن کاوی نشان می‌دهد. طبق جدول ۳، توزیع کلمه‌های ۲۱ گانه پر تکرار، نسبت به کلمه‌های موضوعی در گستره ۱۶۰ کلمه موضوعی تغییر یافته است. چنانچه کلمه "ایران" با فراوانی مطلق ۱۰۰۵۴ بار تکرار و فراوانی نسبی ۶۰۴۸/۹۳۴، "آموزش" با فراوانی مطلق ۵۲۲۷ بار تکرار و فراوانی نسبی ۳۱۲۶/۷۴۶؛ "اجتماعی" با فراوانی مطلق ۵۱۹۷ بار تکرار و فراوانی نسبی ۳۱۲۶/۷۴۶۸؛ "مدیریت" با فراوانی مطلق ۵۰۳۸ و فراوانی نسبی ۳۰۳۱/۰۸۵۲ و کلمه "دانش" با فراوانی مطلق ۴۱۶۶ و فراوانی نسبی ۲۵۰۶/۴۵۱۲ بار تکرار نمای دیگری را نسبت به شمار کلمات موضوعی در این گستره را نشان می‌دهد. تغییر در گستره جست‌وجو و افزایش به ۱۶۰ کلمه تغییر قابل توجهی در جدول شماره ۳ پدید آورده است: فراوانی نسبی کلمات موضوعی دگرگون شده، و دیگر ترتیب کلمات از منظر فراوانی نسبی موجب تحرک کلمات موضوعی در فهرست جدول شماره ۴۲ شده است؛ مثلاً کلمه‌های "اجتماعی"، "مدیریت"، "دانش" از نظر رتبه‌ای به جایگاه بالایی در جدول شماره ۳ رسیده است که به نحو طبیعی بر تصمیم گیری سیاستگذاران مرکز رایسست تأثیرگذار خواهد شد (پیوست ۱).

-تعیین چگونگی همبستگی تکرار موضوعات پر استفاده در عنوان‌های منابع اطلاعاتی توسط کاربران رایسست با استفاده از تکنیک متن کاوی

براساس جدول شماره ۴ (پیوست ۲)، ضریب همبستگی پیرسون بین جفت کلمات موضوعی پرتکرار از عناوین منابع پراستفاده از پایگاه رایسست را نشان می‌دهد. جدول شماره ۴ همبستگی بین کلمه‌های موضوعی و میزان تغییر همزمان تکرارهای (فرکانس‌های) این موضوعات را نشان می‌دهد (اصطلاحاتی که فرکانس‌های آنها با هم یا برعکس بالا رفته و تنزل می‌کنند.). جدول شماره ۴ به طور پیش فرض در نرم افزار ویانت، ستون‌های زیر را نشان می‌دهد:

ترم ۱: ترم اول جفت (ترتیب یا اصطلاحات مهم نیست)؛

ترم ۲: ترم دوم جفت (ترتیب یا اصطلاحات مهم نیست)؛

همبستگی: ضریب همبستگی پیرسون برای این جفت کلمه.

ضریب همبستگی با مقایسه فرکانس‌های نسبی کلمه‌های موضوعی (نسبت به هر قسمت از دیتاست) محاسبه می‌شود. ضریبی که به ۱ نزدیک می‌شود، نشان می‌دهد که مقادیر با هم ارتباط مثبت دارند، با هم افزایش یافته و سقوط می‌کنند. ضریبی که به ۱ نزدیک می‌شود، نشان می‌دهد که مقادیر با هم ارتباط منفی دارند و فرکانس‌ها برای یک‌ترم افزایش یافته در حالی که برای قسمت دیگر افت می‌کنند. ضرایبی که به صفر نزدیک می‌شوند، همبستگی مثبت یا منفی کمی را نشان می‌دهند. این مقدار همبستگی پیرسون توسط کلاس (سیمپل رگرسیون[1]) کتابخانه (اپچ مس کامانس[2]) محاسبه شده است. این مقدار معنی‌داری معیار اطمینان از مقدار همبستگی است. غالباً معنی‌دار بودن ۰/۰۵ یا کمتر بیانگر همبستگی قوی است (که امکان می‌دهد این فرضیه صفر را که مقادیر به طور تصادفی توزیع می‌شوند را رد کنیم). اعتبار این معیار بستگی به فرضیات مربوط به توزیع عادی داده‌ها دارد.

نمودارهای ۱ و ۲، ترندها را به ترتیب در ۲۱ کلمه موضوعی پرتکرار و ۱۶۰ کلمه موضوعی پرتکرار به انتخاب نرم افزار نشان می‌دهد. خطها نمایانگر توزیع فرکانس‌های نسبی کلمه‌های موضوعی در مجموعه عناوین منابع اطلاعاتی پر استفاده در پایگاه اطلاعاتی رایسست است.

Term	Count	Relative	Trend
ایران	10054	109,705	
آموزش	7622	83,168	
جلیه.های.رشدوتوسعه	5948	64,902	
نظام.سیستم	5364	58,530	
جلیه.های.اجتماعی	5196	56,696	
سازمان.سازمانی	5135	56,031	
مدیریت	5038	54,972	
دانشگاه.دانشکده	4025	43,919	
موسسات.آموزشی.عالی	4025	43,919	
دانش.آموز	3797	41,431	
اسلام.واسلامی	3709	40,471	
کودک.وکودکان	3624	39,543	
تهران	3384	36,925	
زن.دختر	3381	36,892	
معماری	3282	35,812	
اطلاعات	3261	35,583	
تولید	2786	30,400	
زندگی	2538	27,694	
زبان	2490	27,170	
دانشجو	2421	26,417	
یادگیری	2320	25,315	
سلامت.بهداشت	2246	24,507	

نمودار ۱. توزیع فرکانس نسبی ۲۱ کلمه موضوعی پرتکرار و ترند آنها به توسط کاربران رایسست

[1]. SimpleRegression
[2]. Apache Math Commons

نمودار ۲. توزیع فرکانس نسبی ۱۶۰ کلمه موضوعی پرتکرار و ترند آنها به توسط کاربران رایسست

بحث و نتیجه گیری

براساس پاسخ به یافته‌های سؤال اول، مبتنی بر موضوعات پرتکرار، آشکار شد کاربران مرکز رایسست گرایش عمیقی به کلمات ویژه دارند؛ این کلمات از کلمه "ایران" با ۱۰۰۵۴ بار، "آموزش" با ۷۶۲۲ بار، "جنبه‌های رشد و توسعه"با ۵۹۴۸ بار، "نظام. سیستم" با ۵۳۶۴ بار، "جنبه. های. اجتماعی" با ۵۱۹۶ بار تکرار آغاز شده، تا اینکه کلماتی چونان "سازمان سازمانی"، "مدیریت"، "مؤسسات. آموزشی. عالی. دانشگاه. دانشکده"، "دانش. آموز"، "اسلام. و اسلامی"، "کودک. و کودکان"، "تهران"، "زن. دختر"، "مساوی"، "اطلاعات"، "تولید"، "زندگی"، "زبان"، "دانشجو"، "یادگیری"، "سلامت. بهداشت" را از ۵۱۳۵ تا ۲۲۶۴ بار تکرار شامل می‌شود. بررسی کلمات مزبور مبین این نکته است که گرایش کاربران به زمینه‌های اجتماعی و آموزشی در واژگان استیلا داشته که مبین رویکرد بازیابی ویژه‌ای است که از کاربران به منابع بروز یافته است؛ نخستین بازتاب از داده‌ها بر این موضوع روایت دارد که کاربران از توجه به مدارک علوم تجربی، پزشکی، و مهندسی خودداری ورزیده که خود بیانگر گرایش کاربران در بهره جویی از مواد پژوهشگران مزبور از منابع لاتین تلقی شده؛ همچنین فقدان واژگانی از حوزه ادبیات، فلسفه، ورزش و رشته‌های پیرامونی به قسمی نشان از رویکرد این دسته از پژوهشگران به منابع کتابی داشته که خود به نوعی غفلت از منابع روز آمد علمی در مجلات جاری آموزش عالی است.

نتایج مرتبط با ابر موضوعی کلمات پر تکرار بیانگر توجه زاید الوصف پژوهشگران ایرانی به موضوعات ملی و تبیین آن در نظام و سیستم کلان محسوب می‌شود؛ استقرار واژه و کلمه "ایران" در مرکزیت ابر به انضمام کلماتی مانند "نظام. سیستم"، "جنبه. های. رشدوتوسعه"، "مدیریت"، "مؤسسات. آموزشی. عالی. دانشگاه. دانشکده" به واقع دل مشغولی فراگیر کاربران را در جامعه‌ای نشان می‌دهد که نیاز به

ثبات نسبی در ارکان خود احساس می‌کند. اینک واژه "اسلام. و اسلامی"که در ابر موجود در حاشیه عمودی مشاهده می‌شود، بازتابی است از حرکت‌های جدید دانش پژوهان ایرانی که مقوله بندی‌های جدیدی را از منظر ذهنی پذیرا شده‌اند.

گستره واژگان و کلمات بر حسب نمودار خطی برای روند ۲۱ موضوع پر تکرار نشان داده می‌شود. چنانچه به محور Y در شکل شماره، ۲ رجوع شود، حجم و گستره غالب کلمه "ایران" و "آموزش" را به کلماتی چون "تولید" و " زندگی" نشان می‌دهد؛ به واقع رجوع به متون آموزش ناشی از نقصان کاربران رشته‌های مرتبط نسبت کلمه‌ای چونان جنبه‌های تولید و اقتصاد در متون خارجی دارد؛ حجم واژگان مزبور بر محورY مبین توجهی است که بایستی نسبت به بنیادهای آموزشی و فراهم آوری مواد الکترونیکی در مراکز و مؤسسات آموزش عالی و مراکز اطلاع رسانی صورت پذیرد.

تعیین چگونگی توزیع موضوعات پر استفاده در عناوین منابع اطلاعاتی توسط کاربران رایسست با استفاده از تکنیک‌های نوین داده کاوی استفاده شد. استفاده از این تکنیک وجوه کلمه را در متن نسبت به حضور آن در توزیع موضوعات منابع اطلاعاتی نشان داده، به واقع اهمیت حیاتی فراوانی مطلق کلمه را نسبت به جایگاه آن در مدارک مرتبط نشان می‌دهد؛ چنانچه "ایران" با ۱۰۰۵۴ فراوانی مطلق از ۱۰۹۷۰۴/۷۳۴ بار فراوانی نسبی در عناوین موضوعی برخوردار است، اما کلمه "اطلاعات" با ۳۲۶۱ بار فراوانی مطلق فقط ۳۵۵۸۲/۲۵۷ بار فراوانی نسبی در عناوین موضوعی شاهد می‌باشد. چنانچه به تعریف و توصیف فراوانی نسبی در موضوع آمار توصیفی توجه شود که اگر دسته i ام دارای فراوانی مطلق fi حاصل از n داده باشد، فراوانی نسبی این دسته به صورت کسر fi/n تعریف می‌شود؛ جدول ۴-۲ فراوانی نسبی هر کلمه را در مدارک در ستون چهارم به نمایش آورده است. که استنتاج می‌شود فزونی فراوانی مطلق برای هر کلمه در مدرک با افزایش فراوانی نسبی آن در طبقات موضوعی همراه است لیکن قابلیت پیش بینی ندارد. در ستون پنجم ۲، نمره Z موضوعات پر تکرار گزارش شده؛ این مطلب که در آمار، بر این معنا می‌باشد که فاصله هر داده از مقدار متوسط اعداد یک داده مجموعه چقدر است و این فاصله برحسب انحراف معیار بیان می‌شود و به نوعی از روش‌ها نیز در نمره Z وجود دارد که می‌تواند داده‌های پرت را حذف کند. از نمرات Z در جدول شماره ۲ آشکار است که واژه‌های با فراوانی مطلق بالا، نمره Z بالایی داشته؛ چنانچه در مشابهت دو واژه "دانشگاه. دانشکده" یا "مؤسسات آموزش عالی" با ۴۰۲۵ بار تکرار نمرهZ همسان و ۴۰۲۲/۷۹۵۷ بار بازتاب یافته است. در نهایت این که نسبت نمره Z که نوعی پیش‌بینی پذیری را در موضوعات با خود حمل کرده و اول بار آلتمن در مباحث امور مالی و ورشکستگی شرکت‌ها بیان نمود؛ به نوعی در ستون ششم جدول شماره ۲ متخذ از تکرار کلماتی تصور شده که با فراوانی مطلق و فراوانی نسبی بالاتری بهرمند می‌باشند. پیش بینی پذیری در این ستون به نوعی رخداد تنظیم یافته در پایگاه‌های اطلاعاتی و مراکز اطلاع رسانی منتهی شده که تا چه حد باب اطمینان را از تهیه مواد چونان گذشته در دستور کار مدیران مراکز اطلاع رسانی قرار دهند.

چگونگی همبستگی تکرار موضوعات پر استفاده در عنوان‌های منابع اطلاعاتی توسط کاربران رایسست نیز بررسی شد نتایج نشان داد که همبستگی بین کلمه‌های "معماری" و "کودک. وکودکان"؛ "تولید" و "زن. دختر"؛ "زبان" و "یادگیری"، به ضریبی معادل ۱ دست یافته، که نشان می‌دهد مقادیر با هم ارتباط منفی دارند؛ در همین رابطه کلمه‌های "آموزش"؛ "ایران"؛ "سازمان. سازمانی"؛ "نظام. سیستم"؛ "تهران" و "مؤسسات. آموزشی. عالی. دانشگاه. دانشکده" که ضریبی کمتر از ۰/۰۵ را نشان می‌دهد، بیانگر همبستگی قوی است؛ این قوت همبستگی بیانگر چند مطلب است؛ چگونه می‌توان با توجه به منابعی که همبستگی معنی‌داری دارد به رویکرد مدیران سوی و جهت داده، توان مجموعه‌های الکترونیکی، استحکام منابع الکترونیکی، و هزینه و فایده مندی را تقویت می‌کند. به واقع نکته مثبت از این بخش می‌توان به ضعف همبستگی‌های کلمات موضوعی مورد شناسایی قرار داد که بایستی در گام‌های آتی مدیریت مجموعه سازی الکترونیکی به آن توجه نماید؛ به نحو طبیعی ضرایب همبستگی ضعیف از کلمه‌های موضوعی بازتابی از خنثی بودن منابعی است که به نوعی فضای کتابخانه‌ها و پایگاه‌های اطلاعاتی را اشغال کرده و کمتر مورد بهره‌جویی واقع می‌شود. در ادامه تحلیل داده‌ها ابتدا فراوانی مطلق و فراوانی نسبی ۲۱ کلمه موضوعی پرتکرار و بازتاب کمی آن درج است؛ بلافاصله به منظور بیان گسترده شماری از کلمات پرتکرار از استفاده کاربران رایسست انعکاس یافته است. مبحث در هم کرد، به منزله بحثی فنی در تکنیک و فنون علوم کتابداری و اطلاع رسانی قابل بررسی است؛ به زبان ساده و شفاف نمایه در هم کرد، نمایه‌ای است به شکل ادواری که در فاصله‌های زمانی معین مطالب جدید را با مطالب یک

یا چند شماره قبل در هم آمیخته و نمایه واحد جدیدی پدید می‌آورد (سلطانی و راستین، ۱۳۷۹، ص ۴۲۸). در آخر به گزارش تصویری نمایه هر کلمه پرتکرار مبادرت نموده است؛ بر این قرار که‌ترند نمایانگر روند یک موضوع است، روندها یک نمودار اسپارک لاین هستند که توزیع فرکانس‌های نسبی را در بین عناوین منابع در مجموعه دیتاست نشان می‌دهد. بر این قرار که نقطه اوج نمودار بر محور عمودی و حاصل یافته‌های کمی ترندها را به ترتیب در ۲۱ کلمه موضوعی پرتکرار به انتخاب نرم افزار نشان می‌دهد. خطها نمایانگر توزیع فرکانس‌های نسبی کلمه‌های موضوعی در مجموعه عناوین منابع اطلاعاتی پر استفاده در پایگاه رایسست است. هرچه اعداد مزبور به سمت ۰.۰۰۱ حرکت نموده در شمار بیشتری از منابع مشاهده شده و به نحو طبیعی‌ترند آن در انعکاس منابع بیشتر جاری می‌شود. بر این قرار محور خطوط بر امتداد موازی با محور افقی استقرار می‌یابد؛ این مطلب بیانگرترند مذکور در کلمه‌های پرتکرار «ایران» با ۰/۰۰۶؛ «آموزش» با ۰/۰۰۳، و اجتماعی با ۰/۰۰۳ ترند به همراه کشش خطوط قابل مشاهده است.

این پژوهش از منظر اجرا و نرم افزارهای استفاده شده و عملیات اجرایی در کشور بداعت داشته و مشابه آن یافت نشد. پژوهش جاری در بخشی با تحقیق سهیلی و همکاران (۱۳۹۷)، تحت عنوان "روند موضوعی مفاهیم حوزه علم اطلاعات و دانش شناسی ایران براساس تحلیل هم‌رخدادی واژگان" همسویی دارد. آن پژوهش سعی داشته در دو بازه زمانی پنج ساله (۱۳۸۴-۱۳۸۹) و (۱۳۹۰-۱۳۹۴) هم رخدادی واژگان را برحسب کلیدواژه‌های مقالات، مورد سنجش قرار دهد، و به این استنتاج دست یافت که مرور زمان موجب گرایش نویسندگان به جنبه‌های فناوری شد؛ موضوعی که از کاربران رایسست نیز در علوم انسانی به منظور بازیابی اطلاعات در این تحقیق قابل ملاحظه است.

پژوهش جاری با تحقیق احمدی و عصاره (۱۳۹۶) تحت عنوان "مروری بر کارکردهای تحلیل هم واژگانی" همسویی دارد. آن پژوهش سعی داشته تحلیل هم واژگانی را با خلاصه سازی مدارک در واژه‌هایی قدرتمند و محاسبه رخداد هم رخدادی نسبت به حوزه موضوعی شناسایی کند؛ موضوعی که کاربران رایسست با کلیدهای موضوعی قدرتمند در دستیابی به خواسته اطلاعاتی خود در مجموعه کلمه‌های موضوعی ۲۱ موضوعی و ۱۶۰ موضوعی دنبال کرده و تکرار پذیری آن از کلمات موضوعی جفتی در ضرایب همبستگی جفتی آشکار شد. همچنین پژوهش جاری با تحقیقات داخلی، از جمله صدیقی (۱۳۹۳) در زمینه استخراج موضوعات علمی و ارتباط میان آنها و کشف محتوای موضوعی، و سهرابی و غفاری (۱۳۹۸) در زمینه هم رخدادی واژگان موضوعات پر کاربرد تولیدات علمی در حوزه ارتباطات علمی نمایه شده در پایگاه اطلاعاتی وب‌آوساینس طی سال‌های ۲۰۰۰-۲۰۱۷ همسویی دارد؛ به واقع کلمه‌های موضوعی در گستره ۲۱ کلمه و ۱۶۰ کلمه‌ای بهرمند از تکرار پذیری بوده و در وجوه کلمه‌های جفتی ضریب همبستگی مطلوبی را احراز کرده‌اند.

در راستای پژوهش‌های بین المللی، پژوهش جاری با تحقیق میلر (۲۰۱۸)، تحت عنوان "پروژه‌های متن کاوی دیجیتال در علوم انسانی: ارزیابی قابلیت‌های تجزیه و تحلیل محتوا با ابزارهای ویانت" که به مطالعه آزمایشی پروژه‌ها با محتوای جدید و مضامین فنی روز آمد پرداخته، همسو بوده و با فرایند تحلیلی به متن کاوی کاربرپسندی مبادرت نموده است. پژوهش جاری همچنین با تحقیق لیدسدرف و نرقس (۲۰۱۶) تحت عنوان "ترسیم نقشه‌های هم واژگانی و مدلسازی موضوعی: مقایسه‌ای با استفاده از شرکت‌های کوچک و متوسط" که با الگو برداری از داده‌های بزرگ و مدلسازی موضوعی به گزینه‌ای برای نقشه برداری هم واژگانی در مدت زمان معین از نظر هم رخدادی پرداخته، همسو بوده و به واقع تحقیق جاری در مدت زمان معین مجموعه گسترده‌ای از کلمه‌های موضوعی را مورد سنجش قرار داده است. در همین راستا با پژوهش ورونتسو و همکاران (۲۰۱۵) به منظور اجرای پژوهش در چند موضوع از مجموعه‌های بزرگ همسویی دارد، و نیز پژوهش، مجموعه‌های درخواست کاربران را به جد با امکانات فنی مورد ملاحظه قرار داد.

با توجه به نتایج پژوهش جاری پیشنهاد می‌شود

- مدیران مجموعه سازی مرکز رایسست به منظور صرفه جویی و هزینه – فایده مندی از مجموعه کلمه‌های جاری برای ساخت پایگاه‌های اطلاعاتی بهره مند شوند؛

- با توجه به نتایج پژوهش جاری پیشنهاد می‌شود مدیران فناوری مرکز رایسست انطباق اطلاعات جمعیت شناختی را با کلمه‌های موضوعی کاربران تسهیل نمایند؛

- با توجه به نتایج پژوهش جاری پیشنهاد می‌شود مدیران آموزش عالی و محتوای برنامه‌های درسی رشته علم اطلاعات و دانش شناسی اهمیت بهره جویی از نرم افزارهای نو ابداع را در گستره فناوری در دستور کار قرار دهند.

منابع

بتولی، ز. (۱۳۹۶). رابطه بین شاخص‌های پایگاه استنادی علوم و ریسرچ گیت: مطالعه موردی مقاله‌های داغ و پر استناد پژوهشگران ایرانی. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳۳(۱۶۱)، ۱۸۳-۱۹۱

رحیمی، م.، زاهدی، م.، و مشایخی، ه. (۱۳۹۷). یک مدل موضوعی احتمالاتی مبتنی بر روابط ملی واژگان در پنجره‌های هم پوشان. *فصلنامه پردازش علائم و داده‌ها*، (۴)۳۸، ۵۷-۷۰.

سلطانی، پ. و راستین، ف. (۱۳۷۹). *دانشنامه کتابداری و اطلاع رسانی*. فرهنگ معاصر.

سهرابی، ط. و غفاری، س. (۱۳۹۸). شناسایی موضوعات پر کاربرد تولیدات علمی حوزه ارتباطات علمی با استفاده از روش هم رخدادی واژگان. *دو فصلنامه علمی دانشگاه شاهد*، (۲)۵، ۴۵-۶۱.

سهیلی، ف.، خاصه، ع.، و کرانیان، پ. (۱۳۹۷). روند موضوعی مفاهیم حوزه علم اطلاعات و دانش شناسی ایران براساس تحلیل هم رخدادی واژگان. *فصلنامه مطالعات ملی کتابداری و سازماندهی اطلاعات*، (۲)۲۹، ۱۷۲-۱۹۰.

صدیقی، م. (۱۳۹۳). بررسی کاربرد روش تحلیل هم رخدادی واژگان در ترسیم ساختار حوزه‌های علمی (مطالعه موردی: حوزه اطلاع سنجی). *پژوهشنامه پردازش و مدیریت اطلاعات*، (۲) ۳۰، ۳۷۳-۳۹۶.

کشاورزیان، س. و برداردخت، ح. (۱۳۹۶). جایگاه کتاب و کتابخوانی در سایت تبیان با رویکرد متن کاوی و تحلیل شبکه‌های اجتماعی. *فصلنامه مدیریت کسب و کار هوشمند*، (۲۱)۶، ۱۶۹-۱۸۸.

متن کاوی (Text Mining) *به زبان ساده (۱۳۹۸)*، بازیابی شده در ۱۲ فروردین ۱۳۹۹، از https://blog.faradars.org/introduction-to-text-mining/.

مرکز منطقه‌ای اطلاع رسانی علوم و فناوری (رایسست). (بی تا). *تاریخچه مرکز منطقه‌ای اطلاع رسانی علوم و فناوری (رایسست)*. بازیابی شده در ۱۷ تیر ۱۳۹۹، از https://ricest.ac.ir/ricest-history/.

مسعودی، ب. و راحتی قوچانی، س.(۱۳۹۴). رفع ابهام معنایی واژگان مبهم فارسی با مدل موضوعی LDA.. *فصلنامه پردازش علائم و داده‌ها*، (۴)۲۶، ۱۱۷- ۱۲۵.

مقدسی، ح.، حسینی، ا.، اسدی، ف.، و جهانبخش، م.(۱۳۹۱). داده کاوی و کاربرد آن در سلامت. *مدیریت اطلاعات سلامت*، (۲)۹، ۲۹۷-۳۰۴.

References

Allahyari, M., Pourieh, A., Assefi, M., Safaei, S., Trippe, J.B., Gutierrez, E., & Kochut, k. (2017). *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques.* E-Printes, KDD Bigdas.

Asemi, A. (2020). Unstructured Data Analysis Recommender System (RSS) (text analysis by voyant). Corvinus university of Budapest.

Batuli, Z. (2017). The relationship between science citation database indicators and research Gate: a case study of hot and highly cited articles by Iranian researcher. Information Processig and Management, 33(161), 91-183. [In Persian].

Chen, H., Wang, X., Pan, s., xiong, F. (2019). Identify Topic Relation In Scientific Literature Using Topic Modeling. *IEEE Transactions on Engineering Management*, 1-13.

Garousi, V. & Mantyla, M.V. (2016). Citations, Research Topics and Active Countries in Software Engineering: Bibliometric study. *Computer Science Review*, 19(2), 56- 77.

Gholamhosseini, L. & Damarvi, M. (2015). Examining the applications of data mining in the health system. *Journal of Paramedical Sciences and Military Health*, 10(1), 39-48. [In Persian].

https://blog.faradars.org/introduction-to-text-mining/. [In Persian].

Ja-hyun, P. & Min, S. (2013). A Study on the Research Trends in Library & Information Science in Korean Using Topic Modeling. *Journal of the Korean Society for Information Management*, 30(1), 7-32.

Keshavarzian, S. & Barardokht, H. (2017). The position of books and reading on the tebian site with the approach of text mining and social network analysis. *Smart Business Management Quarterly*, 6(21), 169-188. [In Persian].

Leydesdorff, L., & Nerghes, A. (2016). Co-Word Maps And Topic Modeling: A Comparison Using Small And Medium- Sized Corpora (N<1000). *Journal of the Association for Information Science and Technology.*

Mantyla, M., Graziotin, & D., Kutila, M. (2018). The Evolution of Sentiment Analysis- A Review of Research Topics, Venues, and Top Cited Papers. *Computer Science Review*, 16-32.

Masudi, M., & Rahati Ghuchani, S. (2015). Resolving the semantic ambiguity of ambiguous Persian words with thematic model LDA. *Quarterly Journal of Signal and Data Processing*, 26(4), 117-125. [In Persian].

Miller, A. (2018). Text Mining Digital Humanities Projects: Assessing Content Analysis Capabilities of Voyant Tools. *Journal of Web Librarianship*, 12(3), 169-197.

Moghadassi, H., Hosseini, A., Asadi, F., & Jahanbakhshm M. (2012). Data mining and its application in health. *Health Information Management,* 9(2), 297-304. [In Persian].

Ogrady, W., Dobrovolsky, M., Aronoff, M. (1993). Contemporary Linguistics, an Introduction, 2nd Ed, St. Martin Press, INC.

Rahimi, M., Zahedi, M., & Mashayekhi, H. (2017). A probabilistic topic model based on national vocabulary relations in overlapping windows. *Quarterly journal of signal and data processing*, 38(4), 57-70. [In Persian].

Regional science and technology information center (RICEST) (without data). History Regional science and technology information center (RICEST). Retrieved 8 July 2020 https://ricest.ac.ir/ricest-history/. [In Persian].

Scott, M. & Tribble, C. (2006). *Textual Patterns: Keyword and Corpus Analysis in Language Education.* Benjamins.

Seddighi, M. (2014). Investigating the application of vocabulary co-occurrence analysis method in drawing the structure of scientific fields (Case study: field of scientometrics information). *Information Processing And Management Research Paper*, 30(2), 373-396. [In Persian].

Soheili, F., Khaseh, A., & Karanian, P. (2018). Thematic trend of concepts in the field of information science and epistemology of Iran based on co-occurrence analysis of words. *Quarterly Journal of National Library Studies and Information Organization*, 29(2), 172-190. [In Persian].

Sohrabi, T., & Ghafari, S. (2019). Identifying the frequently used topic of scientific productions in the field of scientific communication using the co-occurrence method of words. *Two Scientific Quarterly Journals of Shahed University*, 5(2), 45-61. [In Persian].

Soltani, P., & Rastin, F. (2000). *Encyclopedia of librarianship and information*. contemporary culture. [In Persian].

Talib, R., Kashif, M., Ayesha, S., & Fatima, F. (2016). Text Mining: Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Application*, 4(3), 56-78.

Vorontsov, K., Frei, O., Romov, P., & Dudarenko, M. (2015). Open Source Library For Regularized Multimodal Conference On Analysis Of Images. *Social Network and Texts*, 370-381.

Zhou, X., Liu, B., Wu, Z. & Feng, Y. (2007). Integrative Mining of Traditional Chinese Medicine Literature and Medline for Functional Gene Networks. *Artificial Intelligence in Medicine*, 41(2), 87-104.