Machine Indexing of Documents in the Field of Information Retrieval Using Text Mining in the RapidMiner Software

Hamideh Jafari Pavarsi
PhD Candidate in Knowledge and Information Science; Science and Research Branch; Islamic Azad University; Theran, Iran Email: hmdh.jfr@gmail.com

Nadjla Hariri*
PhD in Knowledge and Information Science; Professor; Department of Science and Research Branch; Islamic Azad University; Tehran, Iran Email: nadjlahariri@gmail.com

Mehdi Alipour-Hafezi
PhD in Knowledge and Information Science; Assistant Professor; Department of Knowledge and Information Science; Allameh Tabataba'i University; Tehran, Iran Email: Meh.hafezi@gmail.com

Fahimeh Bab Al-Hawaeji
PhD in Knowledge and Information Science; Associate Professor of Communication and Knowledge Sciences; Department of Science and Research Branch; Islamic Azad University; Tehran, Iran Email: f.babalhavaeji@gmail.com

Maryam Khademi
PhD in Applied Mathematics; Associate Professor; Department of Tehran-South Branch; Islamic Azad University; Tehran, Iran; Email: dr.maryam.khademi@gmail.com

Received: 10, Feb. 2019 **Accepted: 18, Sep. 2019**

Abstract: Machine indexing Provides compatibility between classification codes and indexing terms, extracted expressions and words automatically from a Compiled thesaurus.. In designing an auto-indexing system, computer completely replaces humans. The purpose of this research was to identifying and extracting keywords and the subject trends of articles in the field of information retrieval and the subject's specificity of the author of each article by using the text mining and categorizing (classifying) with the help of concurrence vocabularies. The method of this research is applied and based on the CRISP model of data mining and text mining algorithms are used. The research population consists of 313 articles in the field of information retrieval indexed in the Normmags database. After normalizing the text of the

* Corresponding Author

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 35 | No. 2 | 349-374
Winter 2020

articles by the Virastyar software, and after text mining of the articles by version 7.1 of the RapidMiner software, the keywords were extracted by calculating their weight and were analyzed using two classical classification algorithms consisting of KNN and Naïve Bayse. In this study, the computer automatically indexed the readable machine text by using the frequency of the words with the help of the text mining tools of RapidMiner software. For this purpose, we use N-gram operators and calculate the weight of the words according to TF-IDF method. Terms and key concepts and subject and specialization of author of each article are extracted in the form of 16 categories. Finally, the superiority of the KNN model in categorization of the core subjects of the papers, this study is proving to be 85% more accurate than the Naïve bayse model. Finding the results of calculating the accuracy of the models indicate the acceptable performance of the RapidMiner software in machine indexing of texts. Indexing texts by using this method can help improve the results of information retrieval and prevent false dropping of information in databases.

Keywords: Machine Indexing, Classifying, RapidMiner, Text Mining, Information Retrieval (IR)

نمایه‌سازی ماشینی مدارک حوزهٔ بازیابی اطلاعات با استفاده از متن‌کاوی در نرم‌افزار «رپیدماینر»

حمیده جعفری پاورسی
دانشجوی دکتری علم اطلاعات و دانش‌شناسی؛ دانشگاه آزاد اسلامی؛ واحد علوم و تحقیقات؛ تهران، ایران hmdh.jfr@gmail.com

نجلا حریری
دکتری علم اطلاعات و دانش‌شناسی؛ استاد؛ دانشگاه آزاد اسلامی؛ واحد علوم و تحقیقات؛ تهران، ایران؛ پدیدآور رابط nadjlahariri@gmail.com

مهدی علیپور حافظی
دکتری علم اطلاعات و دانش‌شناسی؛ استادیار؛ دانشگاه علامه طباطبائی؛ تهران، ایران؛ Meh.hafezi@gmail.com

فهیمه باب‌الحوائجی
دکتری علم اطلاعات و دانش‌شناسی؛ دانشیار؛ دانشگاه آزاد اسلامی؛ واحد علوم و تحقیقات؛ تهران، ایران f.babalhavaeji@gmail.com

مریم خادمی
دکتری ریاضی کاربردی؛ دانشیار؛ دانشگاه آزاد اسلامی؛ واحد تهران جنوب؛ تهران، ایران؛ dr.maryam.khademi@gmail.com

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۷/۱۱/۲۱ | **پذیرش:** ۱۳۹۸/۰۶/۲۷ | مقاله برای اصلاح به مدت ۳۷ روز نزد پدیدآوران بوده است.

چکیده: سازگاری کدهای رده‌بندی و اصطلاحات نمایه‌سازی از یک اصطلاح‌نامهٔ مدون با عبارات و کلماتی که به‌طور خودکار استخراج شده، با استفاده از نمایه‌سازی ماشینی ایجاد می‌شود. در طراحی نظام نمایه‌سازی خودکار، کامپیوتر به‌طور کامل جایگزین انسان می‌شود. این پژوهش با هدف استخراج کلمات کلیدی و شناسایی گرایش‌های موضوعی مقالات نمونهٔ آماری در حوزهٔ بازیابی اطلاعات و تخصص موضوعی نویسندهٔ هر مقاله با روش متن‌کاوی و دسته‌بندی آن‌ها با استفاده از هم‌رخدادی واژگان صورت گرفته است. روش این پژوهش از نوع کاربردی است و بر اساس مدل «کریسپ» از مدل‌های فرایند داده‌کاوی و الگوریتم‌های متن‌کاوی انجام گرفته است. جامعهٔ پژوهش، ۳۱۳ مقالهٔ حوزهٔ بازیابی اطلاعات نمایه‌شده

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۵ | شمارهٔ ۲ | ص‌ص ۳۴۹-۳۷۴
زمستان ۱۳۹۸

در «پایگاه نورمگز» است. پس از نرمال‌سازی متن مقالات با نرم‌افزار ویراستیار، طی متن‌کاوی مقالات با نسخهٔ ۷/۱ نرم‌افزار «رپیدماینر»، واژگان کلیدی از طریق محاسبهٔ وزن آن‌ها استخراج و داده‌ها با استفاده از دو الگوریتم کلاسیک دسته‌بندی یعنی «کی‌ان‌ان» و «نایو بیز» تجزیه و تحلیل شدند. در پژوهش حاضر، کامپیوتر با کمک ابزارهای متن‌کاوی نرم‌افزار «رپیدماینر»، متن ماشین‌خوان را با استفاده از بسامد واژه‌ها به‌طور خودکار نمایه‌سازی کرده است. بدین‌منظور، با کمک عملگرهای «ان-گرام» و محاسبهٔ وزن کلمات بر اساس روش «تی‌اف-آی‌دی‌اف»، اصطلاحات و مفاهیم کلیدی و تخصص موضوعی نویسندهٔ هر مقاله در قالب ۱۶ دسته‌بندی استخراج شده است. سرانجام، برتری مدل «کی‌ان‌ان» در دسته‌بندی موضوعات هستهٔ مقالات این پژوهش با دقت ۸۵ درصدی نسبت به مدل «نایو بیز» تأیید شد. مشاهدهٔ نتایج محاسبهٔ دقت‌های مأخوذه از مدل‌ها، گواه کارایی قابل قبول نرم‌افزار «رپیدماینر» در نمایه‌سازی ماشینی متون است. نمایه‌سازی متون با استفاده از این روش می‌تواند به بهبود نتایج بازیابی اطلاعات و جلوگیری از ریزش کاذب اطلاعات در پایگاه‌های اطلاعاتی کمک کند.

کلیدواژه‌ها: نمایه‌سازی ماشینی، دسته‌بندی، رپیدماینر، متن‌کاوی، بازیابی اطلاعات

۱. مقدمه

همگام با افزایش تولید اطلاعات و دانش و پیشرفت صنعت نشر و افزایش حجم مخازن کتابخانه‌ها، طراحی نرم‌افزارهای کتابخانه‌ای نیز رشد چشمگیری داشته است. این رشد در پاسخ به نیاز کتابداران جهت پاسخگویی به پرسش‌های کاربران از طریق منابع موجود در مخزن کتابخانه و نیز تمایل کاربران به دستیابی سریع و انتخاب منبع دلخواه از بین منابع انبوه کتابخانه اتفاق افتاده است؛ به‌طوری که اکنون شاهد پیدایش نرم‌افزارها و برنامه‌های سازماندهی و ذخیره و بازیابی اطلاعات در سطح جهان و کشور هستیم. متون علمی حاوی اطلاعات ارزشمندی است و تبدیل اطلاعات به دانش نیازمند صرف زمان و هزینه‌های بسیار است. اما سیستم‌های اطلاعاتی و نرم‌افزارهای موجود می‌توانند این کار را با استفاده از فناوری‌ها و رویکردهای جدید و در زمان و با هزینه‌ای کمتر به انجام برسانند. در نهایت، با دانش حاصل‌شده می‌توان روند و گرایش‌های تولیدات علمی را در هر حوزه‌ای مشخص ساخته و به‌عنوان ابزاری برای تصمیم‌گیری‌های مدیران در سطوح کلان سیاست‌گذاری‌های علمی و همچنین آینده‌پژوهی مورد استفاده قرار داد. نکتهٔ مهم این‌که نرم‌افزار مورد استفاده باید در ارتقای سطح کیفی نتایج بازیابی‌شده در سیستم‌های اطلاعاتی و به حداقل رساندن ریزش کاذب اطلاعات مؤثر باشد و از طرف دیگر، تلاش برای آزمون قابلیت‌های این نرم‌افزارها از طرف متخصصان علم اطلاعات و دانش‌شناسی

ضروری است. یکی از مواردی که در ارزیابی نرم‌افزارها مورد بررسی قرار می‌گیرد استفاده از شیوه‌های نمایه‌سازی است که می‌تواند نقش عمده‌ای در کارایی نظام ذخیره و بازیابی اطلاعات داشته باشد (سلگی ۱۳۸۲).

امروزه، روش‌های متن‌کاوی با هدف ایجاد ارزش افزوده در نمایه‌سازی خودکار مورد استفاده قرار می‌گیرند. «اکرسون» در مقالهٔ خود در کنفرانس ایفلا ۲۰۱۳ در «سنگاپور»، از اختصار TDM[1] برای مجموعهٔ متن‌کاوی و داده‌کاوی - اکتشاف و تحلیل حجم زیادی از داده‌ها برای کشف الگوها و قواعد معنادار (غضنفری، علیزاده و تیمورپور ۱۳۹۳) - استفاده کرده (Okerson 2013). «اکرسون» از آن به‌عنوان «پردازش خودکار مقادیر زیاد محتوای متنی دیجیتالی ساختاریافته با اهداف بازیابی، استخراج، تفسیر و تجزیه و تحلیل اطلاعات» یاد می‌کند (Okerson 2013). امروزه، کاربرد فنون خودکار تدوین اصطلاح‌نامه‌ها هنوز به‌طور گسترده به اجرا درنیامده است. اگرچه تحقیق با نظام‌های آزمایشی ادامه دارد، اما در تولید خودکار اصطلاح‌نامه، رابطه‌ها بین اصطلاحات نمایه‌سازی عمدتاً به‌صورت آماری شناسایی می‌شوند تا به‌صورت معناشناختی. در تولید خودکار اصطلاح‌نامه برای ایجاد رده‌های اصطلاح‌نامه، تحلیل میزان حضور اصطلاحات در اسناد نظام صورت می‌گیرد و در این کار از بررسی وزن اصطلاحات بر مبنای الگوی فاصله‌برداری که توسط «سالتن»[2] طراحی شده، و از یک الگوریتم دسته‌بندی استفاده می‌شود (حسینی بهشتی ۱۳۸۲).

استخراج واژگان کلیدی یک مسیر مهم تحقیق در زمینهٔ استخراج متن، پردازش زبان طبیعی و بازیابی اطلاعات است. با رشد روز افزون اسناد، به‌کارگیری روش‌های سریع و ارزان برای پیداکردن و استخراج کلمات معنادار از مجموعهٔ وسیع مستندات، اهمیت بیشتری می‌یابد. برای رسیدن به این هدف، پیداکردن الگوهای مکرر، کاوش و استخراج قوانین انجمنی[3] یکی از زمینه‌های مهم در داده‌کاوی است؛ چراکه کشف دانش از یک پایگاه اطلاعاتی جهت فراهم‌کردن اطلاعات مورد نیاز برای فعالیت‌های تصمیم‌گیری و پیش‌بینی، بسیار مهم است.

سنجش تشابه معنایی از اقدامات مهمی است که کاربردهای چشمگیری همچون پردازش زبان طبیعی، اصلاح پرس‌وجو، غلط‌یابی معنایی، مقایسهٔ اسناد و دیگر زمینه‌های

1. Text & Data Mining 2. Salton

۳. قوانین انجمنی روابط و وابستگی‌های متقابل بین مجموعه بزرگی از اقلام داده‌ای را نشان می‌دهند.

کاربـردی در حـوزهٔ موتورهـای جسـت‌وجو به‌منظـور اسـتخراج اصطلاحـات هم‌معنـا دارد. در ایـن زمینـه همچنـان چالش‌هایـی وجـود دارد (صلواتـی ۱۳۹۵).

بـا توجـه بـه گسـترش متـون و مسـتندات الکترونیکـی، اسـتفاده از روشـی کارآمـد جهـت بازیابـی اطلاعـات ضـروری اسـت. نمایه‌سـازی ماشـینی نوعـی نمایه‌سـازی اسـت کـه در آن بـا اسـتفاده از الگوریتـم رایانـه‌ای، واژگان کلیـدی یـک مـدرک از عنـوان یـا متـن اسـتخراج شـده و سـپس، در قالـب مدخل‌هـای نمایـه (در قالـب اسـتخراجی یـا تخصیصـی)، مرتـب و سـازماندهی می‌شـوند. بـرای بازیابـی اطلاعـات، پـی بـردن بـه مفهـوم اصلـی متـن، رده‌بنـدی متـون، و یافتـن کلمـات مناسـب بـرای جسـت‌وجوی مقـالات، اسـتخراج کلمـات کلیـدی بهتریـن روش اسـت. سـاده‌ترین روش بـرای یافتـن واژگان کلیـدی، اسـتفاده از فراوانـی واژه اسـت. به‌منظـور وزن‌دهـی بـه واژگان، از شـیوه «تـی‌اف ـ آی‌دی‌اف»[1] اسـتفاده می‌شـود (علیمـرادی ۱۳۹۶). ایـن روش معمـولاً بـه نتایـج ضعیفـی منجـر می‌شـود، امـا روش‌هـای مختلفـی مـورد بررسـی قـرار گرفته‌انـد و در حـال حاضـر، بهتریـن نتایـج مشاهده‌شـده مربـوط بـه سـامانه‌های یادگیرنـده بـا نظـارت اسـت کـه بـر پایـهٔ ویژگی‌هـای لغـوی و دسـتوری طراحـی شـده‌اند. اسـتخراج کلمـات کلیـدی از مسـتندات، عملیـاتی مهـم در فراینـدهایی ماننـد خوشه‌بندی[2]، طبقه‌بندی[3] و یـا اسـتخراج اطلاعـات معنایـی اسـت (عربـی نرئـی، وحیـدی اصـل و مینایـی بیدگلـی ۱۳۸۶). در پژوهـش حاضـر، بـا توجـه بـه نامشـخص‌بودن مـرز دقیـق کلمـات موجـود به‌دلیـل ابهـام، جدایـی ذاتـی و وندهـای آزاد، اسـتفاده از توالـی کلمـات کلیـدی، بـه جـای خـود کلمـات در اسـتخراج کلمـات کلیـدی موجـود در مسـتندات، به‌منظـور طبقه‌بنـدی کارآمـد آن‌هـا در موتورهـای جسـت‌وجو پیشـنهاد و اسـتفاده شـد. هـدف اصلـی پژوهـش حاضـر اسـتخراج کلمـات کلیـدی و گرایش‌هـای موضوعـی مقـالات نمونـهٔ آمـاری و تخصـص موضوعـی نویسـندهٔ هـر مقالـه اسـت. در ایـن راسـتا اهـداف جزئـی نظیـر شناسـایی روش‌هـای نرمال‌سـازی متـون، میـزان تکـرار و توزیـع فراوانـی کلیدواژه‌هـای هـر مقالـه، پربسـامدترین گرایش‌هـای موضوعـی هـر مقالـه، الگوریتـم مناسـب بـرای شناسـایی کلیدواژه‌هـای متـون، و مقایسـه و اعتبارسـنجی الگوریتم‌هـا و مدل‌هـای دسـته‌بندی به‌دسـت‌آمده نیـز مـد نظـر قـرار گرفته‌انـد.

بـا توجـه بـه مـوارد فـوق، ایـن پژوهـش در راسـتای پرسـش اصلـی آن (چـه گرایش‌هـای موضوعـی در مقـالات نمونـهٔ آمـاری وجـود دارد و تخصـص موضوعـی نویسـندهٔ هـر مقالـه

1. TF-IDF
2. clustering
3. classification

چیست؟) در صدد پاسخگویی به چهار پرسش‌های فرعی زیر است:

۱. از چه روش‌هایی می‌توان برای نرمال‌سازی متون و داده‌ها استفاده کرد؟
۲. میزان تکرار و توزیع فراوانی کلیدواژه‌های مقالات حوزهٔ بازیابی اطلاعات بر اساس میزان هم‌رخدادی واژگان چگونه است؟
۳. الگوریتم مناسب برای شناسایی کلیدواژه‌های متون چیست؟
۴. چگونه می‌توان اعتبار و دقت مدل‌های دسته‌بندی متون را سنجید؟

۲. مروری بر پیشینهٔ پژوهش

تاکنون پژوهش‌هایی متنوع با ابزارهای گوناگون در حوزهٔ نمایه‌سازی انجام گرفته است که از زوایای متفاوت از جمله حوزهٔ معنایی با نرم‌افزارهای متعدد به این موضوع پرداخته‌اند. پیشینه‌های زیر به‌ترتیب به نحوهٔ برچسب‌گذاری اجزای متون و همچنین، به نحوهٔ نمایه‌سازی معنایی پرداخته‌اند.

«ایرانپور و مینایی بیدگلی» در مقالهٔ ارائه‌شدهٔ خود در دومین کنگرهٔ مشترک سیستم‌های فازی و سیستم‌های هوشمند با عنوان «یک روش جدید برای استخراج کلمات و عبارات کلیدی تک‌سندهٔ فارسی با استفاده از تعیین حدود جمله» از یک روش جدید برای استخراج واژه‌های کلیدی سند فارسی استفاده کرده‌اند. ویژگی‌های آماری برای واژه‌های مختلف محاسبه شده و با استفاده از اعمال قواعد، واژه‌های کلیدی محتمل انتخاب می‌شوند. گام بعدی محاسبهٔ رخداد همزمان[1] پیشین و پسین واژه‌های کلیدی محتمل با واژه‌های تکرارشونده در جملات سند است. با استفاده از این روش و بر خلاف اکثر روش‌های آماری که فقط واژه‌های کلیدی یک واژه‌ای را استخراج می‌کنند، واژه‌های کلیدی بیش از یک‌واژه‌ای نیز استخراج می‌شوند. استفاده از روش رخداد همزمان بهبود خوبی را نشان می‌دهد و واژه‌های کلیدی بامعنا را پیشنهاد می‌کند. عمل حذف واژه‌های عمومی[2] که به‌عنوان پیش‌پردازش روی اسناد انجام می‌شد، در این روش به‌صورت پس‌پردازش انجام شده که واژه‌های کلیدی چندواژه‌ای نیز به‌دست آیند (۱۳۸۷).

«خون سیاوش» هدف خود را از انجام رساله‌اش با عنوان «ارائهٔ یک روش نمایه‌سازی معنایی بر پایهٔ هستی‌شناسی برای بازیابی متون و اسناد علمی»، شناسایی مفاهیم مستتر در

1. co-occurrence
2. stopwords

دامنهٔ معنایی و متون و اسناد به‌منظور استفاده در نمایه‌سازی و بهبود عملکرد سیستم‌های بازیابی اطلاعات معرفی می‌کند. برای انجام این کار دامنهٔ معنایی متن با استفاده از دامنهٔ معنایی مفاهیم که در پایگاه دانش سیستم تعریف شده، شناسایی می‌شود. سپس، مفاهیم مستتر در دامنهٔ معنایی متن استخراج و بر اساس ارتباط معنایی که با متن (مفاهیم موجود) دارند، رده‌بندی می‌شوند. مفاهیم موجود در صدر رده‌بندی فوق به‌عنوان مهم‌ترین مفاهیم مستتر در دامنهٔ معنایی متن به نمایهٔ متن افزوده می‌شوند تا در پرس‌وجوها با نمایه مد نظر قرار بگیرند. پیاده‌سازی ایدهٔ مذکور به ابداع دو روش اکتشافی، یکی در زمینهٔ مهندسی دانش و هستی‌شناسی و دیگری در زمینهٔ پردازش زبان طبیعی انجامید. در زمینهٔ هستی‌شناسی یک روش جدید برای نمایش مفاهیم توسط یک بردار معنایی در فضای n بعدی دامنه و برای نگاشت متن به پایگاه دانش سیستم، مفهوم هسته‌های معنایی متن بر اساس زنجیره‌های معنایی ارائه و مورد استفاده قرار گرفت (۱۳۸۹).

«موسوی‌زاده» در پایان‌نامهٔ کارشناسی ارشد خود با عنوان «بررسی ساختار گرایش‌های موضوعی مقالات تألیفی فارسی و انگلیسی حوزهٔ سازماندهی اطلاعات از طریق تحلیل عنوان، چکیده و کلیدواژه‌ها: وزن‌دهی و تحلیل هم‌رخدادی اصطلاحات» هدفش را دستیابی به نقشهٔ معنایی حوزهٔ سازماندهی اطلاعات مطرح کرده است. در این پایان‌نامه ساختار گرایش‌های موضوعی حوزهٔ سازماندهی اطلاعات در دو بخش بررسی شده است: تعیین میزان استفاده از هر اصطلاح موضوعی و تعیین میزان ارتباط بین آن‌ها. برای بخش نخست از تحلیل محتوا و وزن‌دهی اصطلاحات و برای وزن‌دهی، از فرمول مدل فضای برداری و تعیین ضریب وزن برای اصطلاحات از طریق نظرسنجی از متخصصان استفاده شد. برای بخش دوم، تحلیل هم‌رخدادی به‌کار رفت، بدین‌معنا که تعداد دفعاتی که دو اصطلاح موضوعی با هم در یک مقاله ظاهر شده‌اند، محاسبه شد. سپس، ماتریس رخداد اصطلاح/مدرک این اصطلاحات ایجاد و به ماتریس همبستگی «پیرسون»[1] تبدیل شد تا ارتباط‌های معنادار بین اصطلاحات مشخص گردد. با ورود اطلاعات ماتریس همبستگی و همچنین، وزن اصطلاحات به نرم‌افزار یوسینت[2]، تصویری از ساختار اصطلاحات مقالات حوزهٔ سازماندهی اطلاعات ایجاد شد (۱۳۸۹).

«دانش» هدف اساسی پایان‌نامهٔ کارشناسی ارشد خود با عنوان «بهبود طبقه‌بندی متن

1. Pearson 2. UCINET

با استفاده از روش‌های ترکیب» را افزایش میزان صحت طبقه‌بندی متون بیان می‌دارد. وی با توجه به مشکل بودن افزایش کارایی طبقه‌بندهای منفرد در پژوهش‌های پیشین، رهیافت خود را برای رسیدن به این هدف، استفاده و بهبود روش‌های ترکیب طبقه‌بندها قرار داد. در این پایان‌نامه برای بهبود صحت طبقه‌بندی متن، و بر مبنای روش ترکیب رأی‌گیری وزن‌دار، دو رهیافت جدید برای وزن‌دهی طبقه‌ها و طبقه‌بندها پیشنهاد شده است. رهیافت اول، مبتنی بر در نظر گرفتن وزن مستقل برای هر طبقه و هر طبقه‌بند، و رهیافت دوم، تعمیم رهیافت اول است؛ بدین شکل که برای جواب مثبت یا منفی هر طبقه‌بند در مورد هر طبقه وزن مستقلی در نظر گرفته شد. برای محاسبهٔ وزن‌ها در هر دو رهیافت، علاوه بر الگوریتم ژنتیک، معادلهٔ تجربی خاصی هم پیشنهاد شده که در زمان بسیار کمتری نسبت به الگوریتم ژنتیک اجرا می‌شود. نتایج طبقه‌بندی بر مبنای محاسبهٔ وزن‌ها با استفاده از معادلهٔ پیشنهادی، کاملاً با نتایج استفاده از الگوریتم ژنتیک قابل مقایسه و حتی گاهی بهتر هم بود. آزمایش‌ها با استفاده از طبقه‌بندهای «رُکیو»[1]، نزدیک‌ترین همسایه و «بیز»[2]، و سه روش انتخاب ویژگی شامل اطلاعات متقابل، خی ۲ و MCFS انجام گرفت. نتایج تجربی حاصل از اعمال الگوریتم‌های ترکیب پیشنهادی بر روی مجموعه داده‌های آموزشی رایج و مقایسه با نتایج حاصل از سایر روش‌های ترکیب طبقه‌بندها، مانند رأی‌گیری وزن‌دار، عملگر میانگین وزن‌دار رتبه‌یافته و روش قالب تصمیم نشان داد که رهیافت‌های پیشنهادی دقت طبقه‌بندی را به نحو چشمگیری افزایش دادند. این نتایج از آزمایش بر روی چهار مجموعهٔ داده‌های آموزشی متفاوت و رایج به‌دست آمد (۱۳۹۱).

«رمضانی اومالی» در رسالهٔ خود با عنوان «دسته‌بندی اسناد وب با استفاده از گراف نمایه‌سازی اسناد» از مدل جدید نمایه‌سازی سند بر اساس عبارت با عنوان Document Index Graph یک روش دسته‌بندی مبتنی بر گراف که در سال ۲۰۰۴ مطرح شده، استفاده کرده است. در این روش انطباق عبارات جهت بررسی شباهت بین اسناد به‌صورت مؤثر انجام می‌شود. به‌دلیل استفاده از عبارات نسبت به مدل‌های مبتنی بر کلمات منفرد و ساختار گراف کاراتر و فاقد افزونگی بوده و در دسته‌بندی از هر تعداد سند پشتیبانی می‌کند. این مدل همچنین، به‌دلیل ساختار افزایشی الگوریتم دسته‌بندی، قابلیت به‌کارگیری به‌صورت آنلاین در وب را نیز دارد. استفاده از این مدل، نتایج دسته‌بندی اسناد وب را در مقایسه با

1. Rocchio 2. Bayes

روش‌های سنتی تا حد چشم‌گیری بهبود می‌بخشد. این پایان‌نامه به بررسی روش‌های مختلف دسته‌بندی اسناد و نقاط قوت و ضعف هر کدام پرداخته و با تمرکز بر روش دسته‌بندی مبتنی بر گراف به بررسی این روش و مزایای آن نسبت به روش‌های قبلی می‌پردازد. در ادامه، با توجه به این‌که این سیستم قابلیت استفاده در موتور جست‌وجو را جهت دسته‌بندی اسناد بازیابی‌شده دارد، با نگاهی دیگر از زاویهٔ موتور جست‌وجو به بررسی عملکرد این سیستم پرداخته و سعی در بهبود کارایی این سیستم در قالب موتور جست‌وجو دارد. اسناد بازیابی‌شده توسط موتور جست‌وجو غالباً بر اساس میزان بازدید کاربران در لیست نتایج مرتب شده و در اختیار کاربر قرار می‌گیرند. با به‌کارگیری سیستم معرفی‌شده و اضافه‌کردن وزن‌هایی به گره‌ها و یال‌های گراف می‌توان وزن عبارت مورد جست‌وجو را در اسناد مختلف محاسبه و آن‌ها را بر اساس وزن عبارت مورد جست‌وجو مرتب کرد. این کار سبب می‌شود کاربر با دقت و سرعت بیشتر به اطلاعات مورد نظر خود دست یابد. برای اضافه‌کردن وزن با اصلاح ساختار گراف به‌ازای هر سند وزن گره‌ها را با شمارش و وزن یال‌ها را با استفاده از یک شبکهٔ عصبی پرسپترون[1] محاسبه و عملکرد سیستم به‌عنوان بخشی از یک موتور جست‌وجو بهبود می‌یابد (۱۳۹۲).

«ژیانگ و ژنگ» نمایهٔ معنایی پنهان[2] (LSI) را مقوله‌ای می‌دانند که در بسیاری از زمینه‌های پردازش زبان طبیعی استفاده شده و می‌توان ویژگی‌های هم‌رخداد را با انتقال روابط میان مدارک و درون آن‌ها به‌دست آورد. ویژگی‌های با بسامد بالاتر مدرک احتمالاً برخی از ویژگی‌های پنهان، غیرمنطقی و نامشخص آن را معرفی می‌کند که بر روی انتقال روابط به فضای معنایی پنهان و شباهت بین ویژگی‌ها و مدارک در مجموعه مدارک این پژوهش اثر می‌گذارد. در این مقاله ویژگی‌ای در بهینه‌سازی فناوری نمایه‌سازی معنایی پنهان مؤثر است که دارای ویژگی انتقال ارتباط درون و میان مدارک باشد. با الگوریتم پیوند کامل، نتایج آزمایش نشان می‌دهد که این روش به‌طور مؤثر عملیات نمایه‌سازی معنایی پنهان را بهبود می‌بخشد. روش DF (فراوانی اصطلاح) مورد استفاده در این مقاله می‌تواند ویژگی‌های انتخاب‌شده در مجموعهٔ مدارک و تعداد انتقال بین ویژگی‌ها را به‌سادگی فیلتر کند (Jiang and Zheng 2012).

«تکلی، چیبر و ترینا» در مقالهٔ پذیرفته‌شدهٔ خود با عنوان «مدل کامل نمایه‌سازی و

1. perceptron
2. latent semantic index

پرس‌وجوی معنایی طراحی‌شده برای یکپارچگی ناپیوسته در میراث RDBMS»[1] به مسئلهٔ پرس‌وجوی معنایی-آگاهانه پرداخته و یک چارچوب کلی برای مدل‌سازی و پردازش کلمات کلیدی مبتنی بر معنا در پایگاه داده‌های متنی (برای مثال، با در نظر گرفتن شباهت‌ها/ تفاوت‌های واژگانی و معناشناختی هنگام مطابقت با پرس‌وجوی کاربر و اصطلاحات نمایه داده) ارائه می‌دهند. برای انجام این کار، یک ساختار نمایهٔ مقلوب معنایی-آگاهانه به نام SemIndex طراحی و ساخته شد که استاندارد نمایهٔ مقلوب را با ساخت یک نمودار نمایهٔ مقلوب جفتی گسترش می‌دهد که دو مهم را ترکیب کند: یکی شبکهٔ معنایی و دیگری نمایهٔ مقلوب استاندارد در مجموعه‌ای از داده‌های متنی. سپس، یک مدل پرس‌وجوی کلیدواژه با الگوریتم پردازش پرس‌وجوی خاص بر روی SemIndex ساخته شده تا نتایج معنایی-آگاهانه را تولید کرده و به کاربر اجازه دهد که نتایج معنایی را بر اساس نیازهایش انتخاب کند. برای بررسی کاربرد و اثربخشی SemIndex نیز در مورد طراحی فیزیکی آن در یک استاندارد RDBMS تجاری که اجازهٔ ایجاد، ذخیره، و پرس‌وجوی گراف‌مدار را می‌دهد، می‌پردازند. این سیستم قادر است به‌راحتی در مقیاس بزرگ و به حجم زیادی از داده‌ها رسیدگی کند. در نهایت، مرحلهٔ آزمایش SemIndex انجام شده تا زمان، اندازهٔ ذخیره‌سازی، زمان پردازش پرس‌وجو و کیفیت نتیجه در مقایسه با نمایهٔ مقلوب ارزیابی شود. نتایج، حاکی از اثربخشی و مقیاس‌پذیری رویکرد مذکور است (Tekli, Chbeir, & Traina 2018).

تحقیق‌های فوق با تکنیک‌های خوشه‌بندی اسناد، مانند مدل فضابرداری و تحلیل کلمات انجام شده که اغلب با درصدی از خطا نیز همراه است. این پژوهش با هدف سازماندهی و دسته‌بندی اطلاعات و امور مربوط به بازیابی اطلاعات انجام شده و تلاش کرده تا با استفاده از تکنیک‌های پیشرفته‌تر خوشه‌بندی (و دسته‌بندی) نظیر ویژگی‌های وزنی و سنجش عبارات هم‌رخداد، راهی نو به سوی افزونگی کمتر و دقت بیشتر نتایج بازیابی اطلاعات در پایگاه‌های اطلاعاتی بگشاید. اما هیچ‌یک از پژوهش‌ها از منظر متن‌کاوی به حوزهٔ نمایه‌سازی خودکار نپرداخته‌اند. از سوی دیگر، نتایج این تحقیقات اغلب حاصل رساله‌های حوزهٔ فنی و مهندسی است که رویکردی سیستم‌محور داشته تا کاربرمحور. جای خالی این دست پژوهش‌ها در رشته‌های علوم انسانی و علم اطلاعات،

1. Full-fledged Semantic Indexing and Querying Model designed for Seamless Integration in Legacy RDBMS

که رویکردی کاربرمحور نیز می‌تواند داشته باشد، کماکان از سوی محقق احساس شده است. پژوهش حاضر سعی کرده تا با استفاده از نرم‌افزار «رپیدماینر»[1] و الگوریتم‌های متناسب، به دسته‌بندی مدارک حوزهٔ بازیابی اطلاعات (به‌مثابهٔ نمونهٔ آماری) و شکاف حوزهٔ نمایه‌سازی خودکار متون با دیدگاه هم‌رخدادی واژگان با کمک ابزارهای نوین در این جریان بپردازد.

۳. روش پژوهش

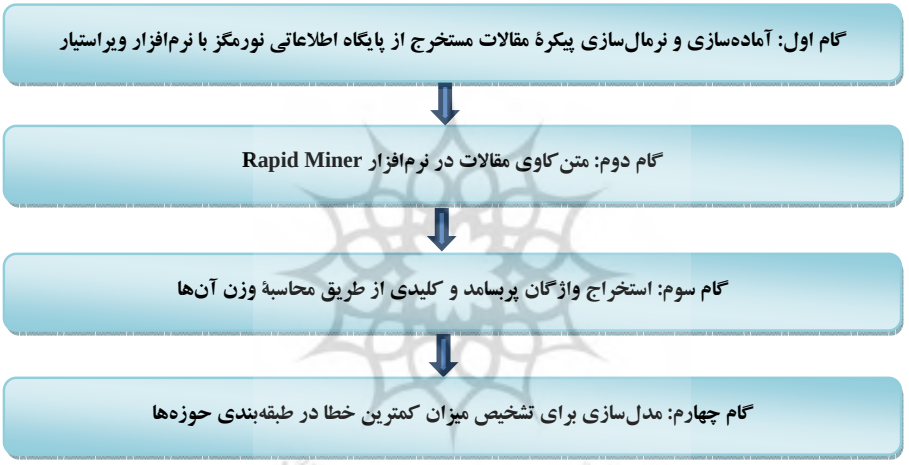
پژوهش حاضر از دستهٔ مطالعات کاربردی تحلیل متن است که با استفاده از روش تحلیل هم‌رخدادی واژگان انجام شده است. تحلیل هم‌رخدادی واژگان یک فن تحلیل محتواست که از الگوهای هم‌رخدادی جفت‌هایی (مانند واژگان یا گروه‌های اسمی) در مجموعه‌ای از متون برای شناسایی ارتباط بین اندیشه‌ها در یک حوزهٔ موضوعی درون این متون استفاده می‌کند (کربلاآقایی کامران، باقری و موسوی‌زاده ۱۳۹۳). این تحلیل، روشی مناسب برای کشف ارتباطات حوزه‌های پژوهشی علم است و پیوندهای مهمی را نشان می‌دهد که ممکن است کشف آن‌ها با روش‌های دیگر مشکل باشد. به‌دلیل هزینه‌بر و زمان‌بربودن برچسب‌زنی داده‌ها با روش‌های داده‌کاوی (به‌طور خاص متن‌کاوی) از روش‌های یادگیری با ناظر در این پژوهش استفاده شده است. در نتیجه، می‌توان آن را در زمرهٔ پژوهش‌هایی از جنس دسته‌بندی مدارک به حساب آورد؛ چرا که در روش‌های با نظارت، ابزارهای خوشه‌بندی برای طبقه‌بندی منابع به‌کار گرفته می‌شود. ایدهٔ این روش آن است که هم‌رخدادی واژگان در یک مدرک را نشان‌دهندهٔ محتوای آن می‌داند (سهیلی، خاصه و کرانیان، [زودآیند]؛ علیپور حافظی، رمضانی و مؤمنی ۱۳۹۶) که پژوهشگر با کمک روش‌های ان-گرام[2] (به‌طور خاص بایگرام[3] و تریگرام[4]) هدف خویش را محقق ساخته است. این پژوهش، با مطالعهٔ روش‌های موجود به استخراج واژگان معنادار هر یک از متون بر اساس معیار رخداد با استفاده از تکنیک TF-IDF برای تعیین وزن کلمات کلیدی و تنظیم ان-گرام‌ها به تعیین واژگان هم‌رخداد در محیط نرم‌افزار «رپیدماینر» مبادرت ورزیده است.

میان نمایه‌سازی تخصیصی (اسنادی) و نمایه‌سازی استخراجی تمایز وجود دارد.

1. Rapid Miner
2. N-gram
3. Bigram
4. Trigram

اغلب نمایه‌سازی‌های انسانی که از نوع نمایه‌سازی تخصیصی هستند و در آن‌ها از نوعی واژگان کنترل‌شده استفاده می‌شود، محتوای موضوعی مدرک ارائه می‌شود. در نمایه‌سازی استخراجی، کلمات و یا عبارات موجود در متن استخراج شده و از آن‌ها برای نشان‌دادن محتوای موضوعی متن استفاده می‌شود. نمایه‌سازان انسانی می‌کوشند تا آن دسته از اصطلاحات متنی را انتخاب کنند که به نظر راهنماهای مناسب از آن چیزی هستند که مدرک دربارهٔ آن بحث می‌کند. در این پژوهش سعی شده تا تلفیقی از معیارهای نمایه‌سازی تخصیصی و استخراجی استفاده شود (لنکستر ۱۳۸۸؛ نوروزی و ولایتی ۱۳۸۹).

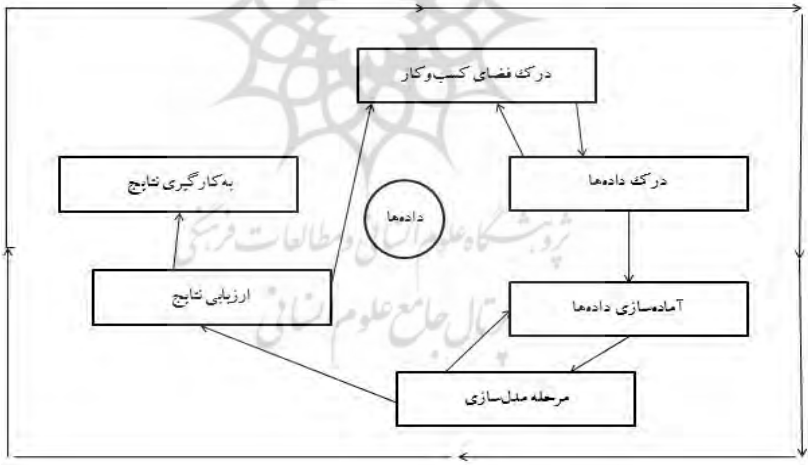
به‌طور خلاصه، مراحل اجرای این پژوهش در شکل زیر ارائه شده است:



شکل ۱. مراحل اجرای پژوهش

داده‌های مورد نیاز این پژوهش حاصل نتایج بازیابی از جست‌وجوی اطلاعات اولیه با عین عبارت بازیابی اطلاعات در ابتدای ۱۶۹۷ مقاله بود که به‌عنوان جامعهٔ پژوهش در نظر گرفته شد. علت انتخاب پایگاه اطلاعاتی نورمگز امکان اخذ خروجی متنی HTML بود که برای انجام عملیات متن‌کاوی مناسب به نظر می‌رسید. طی نمونه‌گیری بر اساس جدول «مورگان» در نهایت، نمونهٔ ۳۱۳ مقاله‌ای به‌صورت تصادفی برای این پژوهش انتخاب شد. اطلاعات هر یک از مقالات (از جمله کد مقاله، عنوان، پدیدآور، نشریه) در جدولی در فضای نرم‌افزار «اکسل» وارد شد. به‌منظور نرمال‌سازی متون پژوهش (داده‌های ساختارنیافته) و برای پیش‌پردازش داده‌ها و یکپارچه‌سازی کلیدواژه‌ها و نشانه‌گذاری‌های متون نمونهٔ آماری از نرم‌افزار «ویراستیار» استفاده شد. (گام اول در شکل ۱)

از آنجا که استخراج خودکار عبارت‌های نمایه‌ای متون نشریات و صفحات وب، خواندن و جست‌وجوی اطلاعات نشریات را برای خوانندگان تسهیل می‌کند و حضور عبارت‌های کلیدی در نتایج جست‌وجو می‌تواند به اصلاح و تعریف مجدد فرمول جست‌وجو و حتی تغییر دیدگاه کاربران از ساختار موجود در یک زمینهٔ خاص و بالابردن ضریب دقت و بازیابی کمک کند (گزنی ۱۳۸۵) و نیز هدف پژوهش حاضر که شناسایی میزان دقت جست‌وجوی نتایج بازیابی‌شده در پایگاه اطلاعاتی مبتنی بر متن کامل مقالات در مقایسه با جست‌وجوی کلیدواژه‌های هم‌رخداد بود، لازم شد تا از مدل «کریسپ»[1] که به معنای فرایند میان‌صنعتی برای داده‌کاوی است، استفاده شود. روش «کریسپ» رویکردی ساختاری برای برنامه‌ریزی یک پروژهٔ داده‌کاوی ارائه می‌دهد. این روش، قدرتمند است و عملکرد آن به‌خوبی اثبات شده است. مراحل آن شامل درک فضای کسب‌وکار، درک داده‌ها، آماده‌سازی داده‌ها، مدل‌سازی، ارزیابی و به‌کارگیری نتایج است که در این پژوهش به‌کار گرفته شده است (اسماعیلی ۱۳۹۵؛ سهرابی، رئیسی وانانی و طالبیان ۱۳۹۵؛ نصیری و اسماعیلی ۱۳۹۵).



شکل ۲. مراحل مدل کریسپ

به‌منظور یکپارچگی داده‌های ساختارنیافتهٔ این پژوهش، از ابزار ویراستیار برای یکپارچه‌سازی نام‌ها و واژگان، نشانه‌گذاری‌ها، واژگان هم‌آوا و هم‌معنا استفاده شد.

1. CRISP-DM (cross-industry process for data mining)

سپس، برای به‌دست آوردن میزان تکرار کلیدواژه‌های موجود در هر مقاله از نرم‌افزار «رپیدماینر» و عملگرهای مربوطه استفاده شد. از آنجا که هدف، به‌دست‌آوردن میزان تکرار کلیدواژه‌های متون بود، نه حروف اضافه یا افعال، یا حروف ربطی، بنابراین، فهرست کلمات حذفی به اپراتورهای این فرایند افزوده شد تا نتایج دقیق‌تری را به‌دنبال داشته باشد. به‌منظور استخراج موضوع و یا اصطلاح هستهٔ هر مقاله از مدل‌های زبانی دوکلمه‌ای[1] و سه‌کلمه‌ای[2] و عملگر وزن‌دهی ان-گرام‌ها استفاده شد و در بازهٔ وزنی دو تا سه کلمه‌ای به تناسب هر مقاله بررسی شد. به‌دنبال آن، کلیدواژه‌های هم‌رخداد هر مقاله به‌ترتیبِ نزولی میزان تکرار هر یک مرتب‌سازی و استخراج شد. آن جفت هم‌رخداد از هر مقاله که بیشترین میزان تکرار را داشتند، به‌عنوان موضوع هستهٔ آن مقاله در نظر گرفته شد (گام دوم و سوم در شکل ۱). به‌دنبال آن، هر یک از مقالات در دایرهٔ شمول موضوعی دسته‌بندی شد که در یافته‌های پژوهش نیز دقت این دسته‌بندی‌ها با مدل‌های متناسب خود مورد سنجش و بحث قرار می‌گیرد (گام چهارم در شکل ۱).

٤. یافته‌های پژوهش

اخیراً نیاز به جست‌وجو و نمایه‌سازی داده‌های معنادار در پایگاه‌های متنی (ساختاریافته و NoSQL) افزایش یافته است. امروزه اغلب کاربران غیرمتخصص کلمات کلیدی خود را در سیستم‌های جست‌وجوی متن‌کامل جست‌وجو می‌کنند. اما این کلمات کلیدی غالباً با آنچه که توسط نویسندگان در نمایه‌سازی اسناد مربوطه استفاده می‌شود، متفاوت است. در نتیجه، این مغایرت به بازیابی نتایج ناخواسته و گاهی غیرمرتبط منجر می‌شود.

متن‌کاوی، مبحث تقریباً جدیدی است که مدتی است به کمک متخصصان علم اطلاعات آمده و با در اختیار گذاشتن ابزارهای مربوط، به تحلیل و کاربردی‌تر ساختن اطلاعات ساختارنیافته (متون) کمک کند. متن‌کاوی به‌دنبال استخراج دانش و الگو از داده‌های متنی ساختارنیافته و به‌عبارتی، کشف اطلاعات جدید و از پیش ناشناخته، به‌وسیلهٔ استخراج خودکار دانش از منابع مختلف نوشتاری است (عابدین ۱۳۹۶). متن‌کاوی مؤثر همراه با سایر تکنیک‌های تجزیه و تحلیل کشف می‌تواند کارایی و مزایای قابل توجهی برای حرفه‌مندان اطلاعات داشته باشد. این مزایا را می‌توان به چندین حوزهٔ گسترده

1. Bigram (بایگرام)
2. Trigram (تریگرام)

تقسیم کرد؛ از جمله: ارزیابی مجموعه (خوشه‌بندی)، اکتشافات گسترده (جست‌وجوی پیچیده)، و ارتباط مفهومی (همان هدفی را که وردنت[1] دنبال می‌کند) (Fox 2010). تصور کنید که دانشجوی تحصیلات تکمیلی در یکی از پایگاه‌های اطلاعاتی، نیاز اطلاعاتی خود را با هدف بازیابی تمام متون مرتبط جست‌وجو می‌کند. او اغلب، نتایج زیادی را بازیابی می‌کند که باید از میان آن‌ها مرتبط‌ترین را انتخاب کرده و برای مرور سایر نتایج بازیابی‌شده با کمبود وقت یا حوصله مواجه می‌شود. متن‌کاوی به‌عنوان تکنیکی کاربردی با خلاصه‌سازی متن[2] و یا شناسایی روابط میان مفاهیم، برای استخراج موضوع و مفهوم اصلی متون مورد استفاده قرار می‌گیرد که در این میان از ابزارهای هوش مصنوعی نظیر پردازش زبان طبیعی[3] نیز استفاده می‌شود.

عناصر متن‌کاوی عبارت‌اند از ماژول (IE[4] + NER[5] + IR[6] + TS[7] + QA[8]) و گام اول شناسایی و بازیابی اسناد مربوطه (IR) (Rzhetsky, Seringhaus, and Gerstein, 2008).

روش وزن‌دهی به لغات بر اساس تکرار لغت، معکوس فراوانی سند (TF-IDF):

به‌طور کلی، میزان اهمیت یک لغت در مجموعهٔ اسناد و وزن‌دهی به لغات بر اساس تکرار لغت، معکوس فراوانی سند (TF_IDF) به دو روش مشخص می‌شود:

۱. فراوانی نسبی تکرار آن لغت در سند که فراوانی لغت نام دارد؛
۲. تعداد اسناد دربرگیرندهٔ آن لغت که فراوانی سند نامیده می‌شود.

در این بین، اِشکال وارده این است که همیشه میزان تکرار بالای کلمه الزاماً ممکن است در دسته‌بندی نقشی را ایفا نکند. لذا، برای حل این مسئله از IDF – به‌عنوان معیار کاستن از درجهٔ اهمیت واژگان بسیار متداول در متن – استفاده می‌شود. هدف این معیار اهمیت دادن به واژگانی است که قدرت تفاوت‌گذاری بیشتری بین گروه‌های مختلف متنی را دارند. در این روش از حاصل‌ضرب لگاریتمی تعداد اسناد در تعداد متون حاوی واژه استفاده می‌شود تا وزن متعادل برای هر واژه استخراج گردد. هدف این ترکیب استفاده از مزایای دو روش و حذف نارسائی‌های آن‌هاست.

$$Tf\text{-}idf_{t,d} = tf_{t,d} \times idf_{t}$$

1. WordNet
2. text summarization (TS)
3. natural language processing (NLP)
4. information extraction (IE)
5. named entity recognition (NER)
6. Information retrieval (IR)
7. text summarization (TS)
8. question answering (QA)

در فرمول بالا چگونگی ترکیب دو مفهوم فوق ذکر شده است. اگر به‌دنبال TF-IDF کلمهٔ t در سند d هستیم، باید فرکانس کلمهٔ t در سند d در idf کلمهٔ t ضرب شود.

TF-IDF برای یک واژه وقتی مقدار بالایی خواهد داشت که هر واژه در تعداد کمی از اسناد رخ داده باشد (کلائی دارابی ۱۳۹۶). به‌عبارت دیگر، هرچه میزان TF-IDF بیشتر باشد، نشانگر اهمیت بیشتر کلیدواژهٔ مورد نظر در محتواست و برعکس.

در این قسمت، به بررسی سؤالات پژوهش و چگونگی پاسخ‌دهی به این سؤالات از طریق نتایج این پژوهش پرداخته شده است.

پرسش ۱. از چه روش‌هایی می‌توان برای نرمال‌سازی متون و داده‌ها استفاده کرد؟

برای انجام متن‌کاوی از هر روشی که استفاده می‌شود، مراحلی شامل پیش‌پردازش داده‌ها، استخراج و انتخاب ویژگی، ایجاد مدل و ارزیابی آن وجود دارد که طی آن دادهٔ خام به دادهٔ قابل پردازش (داده‌های ساختارنیافته به داده‌های ساختاریافته) تبدیل می‌شود و در پایان این مرحله، متن به ویژگی‌های قابل استفاده تبدیل می‌شود. از آنجا که متون مورد بررسی (داده‌های ساختارنیافته) شامل مقالات فارسی نشریات مختلف و در قالب فایل HTML است، و علاوه بر آن، هر نشریه‌ای قالب‌بندی خاص خود را دارد که حاوی تکرار نام نشریه، شماره، صفحات، عنوان مقاله و پدیدآور آن است (به عبارتی، طی مراحل متن‌کاوی جزء داده‌های پرت[1]، مقادیر از دست‌رفته و یا تکراری و فاقد یکدستی محسوب می‌شود)، لذا لازم است که در مرحلهٔ پاک‌سازی داده‌ها برای ویرایش و یا حذف آن‌ها اقدامات لازم به عمل آید.

در این پژوهش، به‌منظور نرمال‌سازی داده‌ها – متن کامل مقالات – از نرم‌افزار «ویراست‌یار» استفاده شده است. طی این مرحله، نویسه‌ها، دستورخط‌ها، نشانه‌گذاری‌ها، واژگان و عبارات مترادف، هم‌آوا، فواصل و نیم‌فاصله‌های کلمات، عبارت‌های جایگزین و یکپارچگی اجزای متن هر مقاله به تفکیک مورد بررسی و ویرایش قرار گرفت (گام اول در شکل ۱).

قطعه‌سازی[2]، نام مرحله‌ای است که طی آن جداسازی اجزای جملات به واحدهای تشکیل‌دهندهٔ آن انجام می‌شود. هر یک از این واحدها در اصطلاح یک توکن[3] نامیده می‌شود. طی پردازش زبان طبیعی بر روی متون، تمام کلمات مورد پردازش قرار

1. noise (نویز) 2. tokenizing 3. Token

می‌گیرند. اما از آن رو که تمامی کلمات متن، از جمله حروف اضافه، ربط و ...، برای پردازش مفید نیستند، بنابراین، به‌منظور پردازش مفیدتر و کامل‌تر بهتر است در مرحلهٔ پیش‌پردازش حذف شوند. این حروف و کلمات زائد و اضافی نظیر «و»، «که»، «ولی»، «اگر» و غیره – به‌عنوان ایست‌واژه‌ها[1] – به نرم‌افزار معرفی می‌شوند تا به بهبود عملکرد متن‌کاوی منجر شود. این مرحله را فیلترکردن[2] می‌نامند (گام دوم در شکل ۱).

برای استخراج ویژگی امتیازدهی از روش پرکاربرد TF-IDF استفاده شده است. در فرکانس واژه، هر واژهٔ سند وزنی خواهد داشت. این وزن بر اساس تعداد تکرار واژه در سند مشخص می‌شود و هر واژه‌ای که بیشترین بسامد تکرار را داشته باشد شانس بیشتری برای انتخاب به‌عنوان یک ویژگی و معیار دسته‌بندی خواهد داشت.

به‌منظور نیل به هدف هم‌رخدادی واژگان از الگوریتم‌های یادگیری ماشینی «بایگرام» استفاده شده است تا بتوان پربسامدترین اصطلاحات هر مقاله را به‌عنوان ویژگی هر متن، که در این پژوهش ویژگی هستهٔ هر مقالهٔ مد نظر است، استخراج کرد. در مدل زبانی «بایگرام»، متون علمی به‌صورت دوکلمه دوکلمه بررسی می‌شوند (شیروانی، وطن‌خواه خوزانی و یغمایی ۱۳۹۳). در این پژوهش با توجه به نوع کاربرد زمینه‌ای آن، مقدار «ان-گرام» برابر با عدد ۲ در نظر گرفته شده است. در روش یادگیری ماشینی با مجموعه‌ای از اسناد آموزشی و کلمات کلیدی مشخص برای آن‌ها، فرایند استخراج کلمات کلیدی به‌عنوان یک مسئلهٔ طبقه‌بندی مدل‌سازی می‌شود. کلمات بر اساس مشخصه‌هایشان، به «کلمات کلیدی» و «کلمات غیرکلیدی» طبقه‌بندی می‌شوند (گام سوم در شکل ۱).

پیش از مطرح‌شدن نظریهٔ فازی، الگوریتم‌های سنتی خوشه‌بندی به‌گونه‌ای عمل می‌کردند که هر قالب و در نتیجه، هر مدرک صرفاً به یک خوشه اختصاص می‌یافت. به همین جهت این قبیل راهبردها خوشه‌بندی‌های سخت نام گرفتند. با استفاده از نظریهٔ فازی و اختصاص درجهٔ عضویت به هر قالب، انعطاف‌پذیری نسبتاً مناسبی برای گروه‌بندی مدارک به‌وجود آمد. بر این اساس، هر قالب و بهتر بگوییم هر مدرک، درجه‌ای از عضویت در خوشه‌های مختلف خواهد داشت. بنابراین، می‌توان نتیجه گرفت که خوشه‌های حاصل از روش‌های خوشه‌بندی سخت کاملاً از یکدیگر جدا بوده و وجوه مشترک ندارند؛ در حالی که در خوشه‌بندی فازی، خوشه‌های حاصل وجوه و

1. stop words
2. filtering

فصول مشترک خواهند داشت (ارسطوپور ۱۳۸۸). بنابراین، با توجه به دسته‌بندی‌های انجام‌شده در این پژوهش می‌توان گفت که علاوه بر تخصیص هر موضوع به یک مقاله، با توجه به خاصیت چندوجهی، برخی مقالات می‌توانند در دسته‌های دیگر نیز قرار گرفته و با برخی مقالات دیگر هم‌موضوع و مشترک باشند. به‌عبارت دیگر، راهبرد خوشه‌بندی نرم نیز می‌تواند در این پژوهش مورد استفاده قرار گیرد.

پرسش ۲: میزان تکرار و توزیع فراوانی کلیدواژه‌های مقالات حوزهٔ بازیابی اطلاعات بر اساس میزان هم‌رخدادی واژگان چگونه است؟

پیوست ۱، حاوی فهرست موضوعات هستهٔ هر یک از مقالات نمونهٔ آماری است. پربسامدترین گرایش‌های موضوعی هر یک از مقالات حوزهٔ بازیابی اطلاعات در فهرست زیر ارائه شده است.

جدول ۱. پربسامدترین گرایش‌های موضوعی

گرایش موضوعی	میزان بسامد	گرایش موضوعی	میزان بسامد
فلسفهٔ کتابداری	۶	فناوری اطلاعات	۶
بازیابی اطلاعات	۳۲	اصطلاح‌نامه	۱۰
جست‌وجو	۲۹	نمایه‌سازی	۸
مدیریت اطلاعات	۶	ربط	۶
سازماندهی اطلاعات	۱۶	نرم‌افزارهای کتابخانه‌ای	۶
موتورهای جست‌وجو	۱۸	کتابخانه‌های دیجیتال	۸
وب معنایی	۶	سواد اطلاعاتی	۶
رابط کاربر	۶	رفتار اطلاع‌یابی	۴

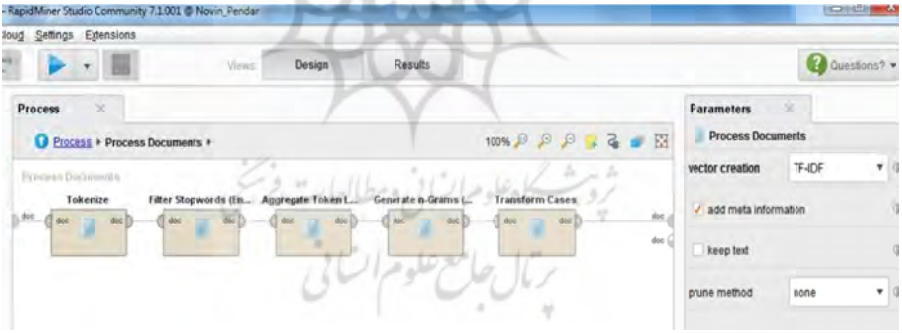
در جدول بالا فهرست پربسامدترین گرایش‌های موضوعی را مشاهده می‌کنید. طبق این جدول، موضوعات «بازیابی اطلاعات» با ۳۲ مقاله و «جست‌وجو» با ۲۹ مقاله از پربسامدترین مقالات حوزهٔ بازیابی اطلاعات نمونهٔ آماری این پژوهش محسوب می‌شوند. همان‌گونه که در روش‌شناسی پژوهش اشاره شد، نتایج پرسش دوم این پژوهش برگرفته از به‌کارگیری عملگرهای مربوط به هم‌رخدادی واژگان و ابزارهای «ان-گرام» (به‌طور خاص عملگر رایج دوکلمه‌ای) است (گام دوم در شکل ۱).

پرسش ۳: الگوریتم مناسب برای شناسایی کلیدواژه‌های متون چیست؟

در راستای تشریح گام سوم در شکل ۱، برای استخراج ویژگی امتیازدهی (وزن‌دهی) از روش پرکاربرد TF-IDF استفاده شده است. TF به معنای Term Frequency یعنی تعداد تکرار یک کلمه در یک متن و IDF به‌معنای Inverse Document Frequency است که می‌توان آن را به «برعکس تعداد تکرار در متون» ترجمه کرد. به‌عبارت دیگر، TF-IDF صرفاً میزان تکرار یک کلمهٔ کلیدی یا عبارت را در متن نشان نمی‌دهد، بلکه هدف آن نشان‌دادن اهمیت کلمهٔ کلیدی مورد نظر از طریق مقایسهٔ تعداد تکرار کلمه در متن با تکرار آن کلمه در مجموعه‌ای بزرگ‌تر از مستندات است.

لازم به توضیح است که به‌منظور سهولت در انجام محاسبات آماری از عملگرها و اپراتورهای هر یک از الگوریتم‌ها و محاسبه‌گرهای مربوطه استفاده شده است. این اپراتورها بدین‌ترتیب به‌کار گرفته شده‌اند: read document، process document، tokenize، filter stopwords، filter token (by length)، transform cases، که در نهایت، پس از اجرا میزان رخداد هر کلیدواژه قابل نمایش خواهد بود (گام سوم در شکل ۱).

شکل ۳، نشان‌دهندهٔ نحوهٔ عملیات انجام‌شده در نرم‌افزار «رپیدماینر» است.

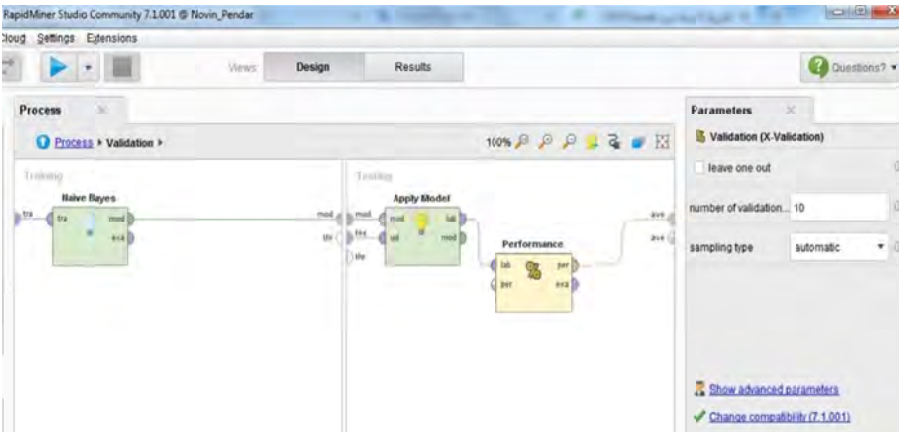


شکل ۳. اپراتورهای به‌کار گرفته‌شده در متن‌کاوی با نرم‌افزار «رپیدماینر»

پرسش ۴: چگونه می‌توان اعتبار و دقت مدل‌های دسته‌بندی متون را سنجید؟

در این پژوهش برای ارزیابی و بررسی دقت دسته‌بندی متون از دو مدل «نایو بیز»[1] و «کِ‌ان‌ان»[2] استفاده شده که شکل زیر نشانگر تنظیمات و فراخوانی اپراتورهای مربوطه در نرم‌افزار «رپیدماینر» است.

1. Naive Bayes
2. KNN stands for k-Nearest Neighbors



شکل ٤. عملیات «نایو بیز» در نرم‌افزار «رپیدماینر»

احتمالات طبقه‌بندی به‌صورت آماری از مجموعهٔ آموزشی یاد گرفته می‌شوند. این روش‌ها دارای انعطاف‌پذیری زیادی هستند. به‌عنوان نمونه، یکی از این روش‌های به‌کاررفته مدل یادگیری «نایو بیز» است. این الگوریتم بر اساس احتمال شرطی طراحی شده است. مبنای اصلی این الگوریتم نظریهٔ «بیز» است؛ به‌طوری که اگر رخداد B وابسته به رخداد A باشد، یعنی B در شرایطی که رخداد A اتفاق افتاده باشد. برای محاسبهٔ احتمال وقوع B به شرط A الگوریتم ابتدا تمامی مواردی را که رخدادهای A و B به‌صورت همزمان اتفاق افتاده‌اند، می‌شمارد و سپس، به تعداد رخدادهای A که به‌تنهایی اتفاق افتاده، تقسیم می‌کند تا احتمال شرطی مورد نظر حاصل گردد. همچنین، ابتدا درصد رخداد دوبه‌دو بررسی می‌شود و به درصد رخداد تک‌به‌تک تقسیم می‌گردد (شیروانی، وطن‌خواه خوزانی و یغمایی ۱۳۹۳).

«کی‌ان‌ان» شاید ساده‌ترین الگوریتم در بحث طبقه‌بندی باشد. «کی‌ان‌ان» (نزدیک‌ترین همسایگی)، یک روش ناپارامتریک است که در داده‌کاوی، یادگیری ماشین و تشخیص الگو مورد استفاده قرار می‌گیرد و یکی از ده الگوریتمی است که بیشترین استفاده را در پروژه‌های گوناگون یادگیری ماشین و داده‌کاوی، هم در صنعت و هم در دانشگاه، داشته است.

لازم به توضیح است که به‌منظور سهولت در انجام محاسبات آماری از عملگرها و اپراتورهای هر یک از الگوریتم‌ها و محاسبه‌گرهای مربوطه استفاده شده است. این اپراتورها بدین‌ترتیب به‌کار گرفته شده‌اند: process document from files، tokenize، filter

naïve ،x-validation ،generate n-gram (terms) ،stem (porter) ،transform cases ،stopwords ،bayes ،KNN ،apply model و performance که در نهایت، پس از اجرا میزان دقت هر مدل قابل نمایش خواهد بود (گام چهارم در شکل ۱).

پس از انجام محاسبات و اجرای مدل‌های مربوطه، دقت هر یک به‌تفکیک به‌شرح زیر است.

جدول ۲. میزان دقت محاسبه‌شده مدل‌های اجرایی در نرم‌افزار رپیدماینر

مدل	میزان دقت (Accuracy)
نایو بیز	٪۷۸/۳۳ +/- ٪۲۳/۶۳ (mikro: ٪۷۹/۳۱)
کی‌ان‌ان	٪۸۵/۰۰ +/- ٪۱۸/۹۳ (mikro: ٪۸۶/۲۱)

۵. بحث و نتیجه‌گیری

در این پژوهش با بررسی متن کامل تعداد ۳۱۳ مقاله در حوزهٔ بازیابی اطلاعات و وزن‌دهی آن‌ها، مقالات حوزهٔ پژوهش در ۱۶ دسته بر اساس موضوعات پربسامد و پرتکرار دسته‌بندی شد. این ۱۶ دستهٔ پربسامد نشان‌دهندهٔ تخصص موضوعی نویسندگان این مقالات نیز است.

با توجه به محدودیت‌های این پژوهش در مورد استفاده از نرم‌افزار «رپیدماینر» در پردازش و متن‌کاوی تعداد بیش از ۱۰۰ متن کامل فارسی و ویژگی‌هایش و عدم دسترسی پژوهشگر به یک ابرکامپیوتر برای پردازش یکبارهٔ متن کامل ۳۱۳ مقاله و تحلیل ماتریس‌های متقاطع بسیار پیچیده، اقدام به نمونه‌گیری و محاسبهٔ دقت ۲ نمونه از پرکاربردترین مدل‌های کلاسیک سنجش دقت دسته‌بندی‌ها، یعنی «نایو بیز» و «کی‌ان‌ان» شد. لازم به ذکر است که این نرم‌افزار در متن‌کاوی متون انگلیسی در نمونه‌های گسترده به‌خوبی عمل می‌کند. به همین دلیل، میزان دقت ۵ دستهٔ به‌دست‌آمده به‌عنوان نمونهٔ تصادفی مورد اندازه‌گیری قرار گرفت. این ۵ دسته نمونه عبارت‌اند از: کتابخانهٔ دیجیتال، ربط، رابط کاربر، سواد اطلاعاتی، و رفتار اطلاع‌یابی. پس از اجرای مدل‌های «نایو بیز» و «کی‌ان‌ان» به‌ترتیب، با دقت ۷۸ و ۸۵ درصدی، اعتبارسنجی دسته‌بندی مقالات این پژوهش حاصل شد. با مشاهدهٔ میزان دقت آن‌ها می‌توان نتیجه گرفت: از آنجا که دقت مدل «کی‌ان‌ان» با مقدار ۸۵ درصد اعتبار بیشتری دارد، لذا عملکرد بهتر، صحت و اعتبار

فرایند دسته‌بندی و تعیین گرایش‌های موضوعی مقالات و تخصص موضوعی نویسندگان هر مقاله در نمونهٔ آماری این پژوهش را تأیید می‌کند. با توجه به نتایج این پژوهش به نظر می‌رسد که بتوان نرم‌افزار «رپیدماینر» را به‌عنوان یکی از ابزارهای مناسب برای نمایه‌سازی خودکار قلمداد کرده و به‌کار گرفت.

مشاهدهٔ نتایج محاسبهٔ دقت‌های به‌دست‌آمده از مدل‌ها، گواه کارایی قابل قبول نرم‌افزار «رپیدماینر» در نمایه‌سازی ماشینی است. انتظار می‌رود که در پژوهش‌های آتی با استفاده از نرم‌افزارهای پیشرفته‌تر در حوزهٔ متن‌کاوی و افزایش مهارت برنامه‌نویسی بتوان به اهداف متن‌کاوی متون فارسی در ابعاد بزرگ‌تر و به‌تبع آن، نمایه‌سازی معنایی و بهبود نتایج بازیابی اطلاعات در موتورهای جست‌وجو دست یافت.

با توجه به این‌که همه‌ساله در سراسر دنیا، سرمایه‌های بسیاری صرف پژوهش‌های علمی و کاربردی می‌شود که گاهی تکراری و یا غیرکاربردی هستند، بنابراین، اقداماتی در راستای نمایه‌سازی، به‌ویژه از نوع معنایی می‌تواند در تعیین گرایش‌های حال و آیندهٔ حوزه‌های علمی کمک شایانی کند. به‌عبارت دیگر، آینده‌پژوهی و روشن نمودن مسیری سازنده برای سیاست‌های علمی دنیا و در نتیجه رشد و شکوفایی رویکردهای نوین پژوهشی و پرکردن خلأهای علمی به شکلی مؤثر با استفاده از طراحی و به‌کارگیری ابزارهای نوین نیمه‌خودکار و خودکار می‌تواند به‌عنوان رویکردی جدید از فعالیت‌های مفید و تخصصی در حوزهٔ علم اطلاعات و دانش‌شناسی باشد؛ چرا که با استفاده از همین رویکرد می‌توان به اشتغال‌زایی پرداخت و مهارت‌های متخصصان این حوزه را نیز از ارزشی افزوده برخوردار ساخت.

این پژوهش در راستای پیشینه‌های «ایرانپور و مینایی‌بیدگلی» (۱۳۸۷)، «خون سیاوش (۱۳۸۹) و موسوی‌زاده» (۱۳۸۹)، «تکلی، چیبر و ترینا» (۲۰۱۸) انجام شده و پژوهشگر با به‌کارگیری رویکرد جدید متن‌کاوی برای یادگیری نیمه‌نظارتی سیستم و استفاده از مدل‌های پنهان «مارکوف» و به‌کارگیری ابزارهای «ان-گرام» به‌عنوان عملگرهای مربوط به هم‌رخدادی واژگان و (به‌طور خاص عملگر رایج دوکلمه‌ای) در نرم‌افزار «رپیدماینر» که مورد استفادهٔ هیچ‌یک از پیشینه‌های مورد اشاره در این پژوهش نبوده است، برای نزدیک‌سازی اهداف نمایه‌سازی با نتایج مورد انتظار کاربران در بازیابی اطلاعات از پایگاه‌های اطلاعاتی تلاش نمود تا کاربران در جست‌وجوی اطلاعات با ریزش کاذب و نتایج غیرمرتبط کمتری مواجه شوند. به‌عبارت دیگر، نمایه‌سازی ماشینی در تکنیک‌های

متن‌کاوی جلوه‌گر شده است. متن‌کاوی به‌عنوان تکنیکی کاربردی برای کتابداران به استخراج مفاهیم اصلی متون علمی مورد استفاده می‌پردازد. استفاده از نتایج این پژوهش می‌تواند به کارایی روش‌های جدید در نمایه‌سازی ماشینی منتهی شود. انگیزهٔ اصلی پژوهشگر با استخراج موضوعات هستهٔ هر مقاله و گرایش و تخصص موضوعی نویسندهٔ هر مقاله با تعریف ابزارهای هم‌رخدادی واژگان (ان-گرام‌ها در رپیدماینر) می‌تواند در آینده بسترساز ارائهٔ راهکارهای جدید در پردازش زبان فارسی و طراحی سیستم‌های اطلاعاتی معنامحور به‌صورت کاربردی شود.

همان‌گونه که در پیشینهٔ پژوهش اشاره رفت، هیچ‌یک از پژوهش‌ها از دیدگاه رشتهٔ علم اطلاعات و دانش‌شناسی و کاربردهای استخراج واژگان کلیدی با الگوریتم‌های ویژه برای نمایه‌سازی و به‌طور خاص با استفاده از نرم‌افزار «رپیدماینر» نپرداخته‌اند. بنابراین، پژوهش حاضر تلاش نموده تا با کمک ابزارهای متن‌کاوی و الگوریتم‌های ویژه، مدل‌هایی را استخراج کند که کمترین میزان خطا را دربرداشته و بتواند نمایه‌سازی ماشینی را با استفاده از استخراج مفاهیم کلیدی و موضوع هستهٔ هر مقاله، به نمایه‌سازی معنایی نیمه‌نظارتی نزدیک‌تر سازد. امروزه، پایگاه‌های اطلاعاتی تخصصی متون زیادی را در حوزه‌های مختلف دربرمی‌گیرند که در آن‌ها از ابزارهای نمایه‌سازی ماشینی برای تسهیل در بازیابی اطلاعات موتورهای جست‌وجو استفاده شده است، و جای خالی نمایه‌سازی معنایی و استفاده از مدل‌های مخفی «مارکوف» و تکنیک‌های «ان-گرام» و ابزارهایی از این دست و کاربرد آن‌ها در بهینه‌سازی نتایج بازیابی اطلاعات همواره از سوی پژوهشگر احساس می‌شد. این پژوهش تلاش نموده تا گامی هرچند کوچک در راستای به‌کارگیری این ابزارها در جهت بهینه‌سازی نتایج و بازیابی اطلاعات شبه‌معنایی در موتورهای جست‌وجوی پایگاه‌های اطلاعاتی و همچنین نمایه‌سازی معنایی بردارد.

فهرست منابع

ارسطوپور، شعله. ۱۳۸۸. دسته‌بندی نتایج جست‌وجو بر مبنای ویژگی‌های مدارک و امکان‌سنجی استفاده از الگوریتم‌های خوش بندی مختلف در سطح وب. *کتابداری و اطلاع‌رسانی* ۴۶: ۱۴۵-۱۷۴.

اسماعیلی، مهدی. ۱۳۹۵. *آموزش گام‌به‌گام داده‌کاوی با RapidMiner*. تهران: آتی‌نگر، وینا.

ایرانپور، مجید، و بهروز مینایی بیدگلی. ۱۳۸۷. یک روش جدید برای استخراج کلمات و عبارات کلیدی تک‌سنده فارسی با استفاده از تعیین حدود جمله. *دومین کنگره مشترک سیستم‌های فازی و سیستم‌های*

هوشمند. مشهد، دانشگاه فردوسی مشهد.

حسینی بهشتی، ملوک. ۱۳۸۲. کاربرد اصطلاح‌شناسی و واژه‌گزینی در نمایه‌سازی ماشینی و بازیابی اطلاعات. *علوم اطلاع‌رسانی* ۱۸ (۳-۴): ۳۱-۴۴.

خون سیاوش، احسان. ۱۳۸۹. *ارائه یک روش نمایه‌سازی معنایی بر پایه هستی‌شناسی برای نمایه‌سازی متون و اسناد علمی*. پایان‌نامه کارشناسی ارشد. دانشگاه اصفهان. دانشکده فنی و مهندسی.

دانش، علی. ۱۳۹۱. *بهبود طبقه‌بندی متن با استفاده از روش‌های ترکیب*. پایان‌نامه کارشناسی ارشد. دانشگاه کردستان. دانشکده مهندسی.

رمضانی اومالی، نرجس. ۱۳۹۲. *دسته‌بندی اسناد وب با استفاده از گراف نمایه‌سازی اسناد*. پایان‌نامه کارشناسی ارشد. دانشگاه صنعتی شاهرود، دانشکده کامپیوتر و فناوری اطلاعات.

سلگی، غلامرضا. ۱۳۸۲. نقش نمایه‌سازی در بازیابی اطلاعات. *کتاب ماه کلیات* ۶۵: ۸۱-۸۴.

سهرابی، بابک، بابک رئیسی‌وانانی، و مرضیه طالبیان. ۱۳۹۵. ارائه الگویی برای تحلیل رفتار کاربران شبکه‌های اجتماعی با استفاده از روش‌های داده‌کاوی: یک شبکه اجتماعی در ایران». *پژوهش‌های مدیریت منابع انسانی* ۶ (۴): ۸۳-۱۰۶.

سهیلی، فرامرز، علی‌اکبر خاصه، و پریوش کراتیان (زودآیند). ترسیم ساختار فکری حوزه علم اطلاعات و دانش‌شناسی ایران بر اساس تحلیل هم رخدادی واژگان. *پژوهش‌نامه پردازش و مدیریت اطلاعات*.

شیروانی، پریسا، مهرداد وطن‌خواه خوزانی، و خشایار یغمایی. ۱۳۹۳. بازشناسی متون فارسی با استفاده از مدل زبانی n-gram و پالایش گرامری. *دو فصلنامهٔ پردازش علائم و داده‌ها* ۱ (۲۱): ۱۰۷-۱۱۵.

صلواتی، سامان. ۱۳۹۵. *اندازه‌گیری شباهت معنایی بین صفحات وب بر اسا شباهت بین رخداد کلمات*. پایان‌نامه کارشناسی ارشد. دانشگاه دیلمان لاهیجان. دانشکده فنی و مهندسی.

عابدین، احسان. ۱۳۹۶. *داده‌کاوی و متن‌کاوی در نرم‌افزار رپیدماینر*. تهران: اندیشه رشد.

عربی‌نرئی، سمیه، مجتبی وحیدی‌اصل، و بهروز مینایی‌بیدگلی. ۱۳۸۶. «استخراج کلمات کلیدی جهت طبقه‌بندی متون فارسی. *اولین کنفرانس داده‌کاوی ایران*، تهران، دانشگاه صنعتی امیرکبیر، مؤسسه پژوهشی داده‌پردازان گیتا.

علیپور حافظی، مهدی، هادی رمضانی، و عصمت مؤمنی. ۱۳۹۶. *ترسیم نقشهٔ دانش حوزهٔ کتابخانه‌های دیجیتالی در ایران: تحلیل هم‌رخدادی واژگان. پژوهشنامه پردازش و مدیریت اطلاعات* ۳۳(۲): ۴۵۳-۴۸۸.

علیمرادی، مصطفی. ۱۳۹۶. روش‌های استخراج خودکار دانش از متون حدیثی؛ مروری بر پژوهش‌های صورت‌گرفته». *ره‌آورد نور* ۶۰: ۳۶-۴۴.

غضنفری، مهدی، سمیه علیزاده، و بابک تیمورپور. ۱۳۹۳. *داده‌کاوی و کشف دانش*. تهران: دانشگاه علم و صنعت ایران.

کربلاآقایی کامران، معصومه، منصوره باقری، و مریم موسوی‌زاده. ۱۳۹۳. *مصورسازی حوزه سازماندهی*

اطلاعات: بررسی ساختار گرایش‌های موضوعی مقالات فارسی حوزه سازماندهی اطلاعات. پژوهشنامه کتابداری و اطلاع‌رسانی ۸: ۱۹۰-۲۱۱.

کلائی دارابی، خلیل. ۱۳۹۶. *نظرکاوی خوانندگان اخبار فارسی در شبکه‌های اجتماعی بر اساس الگوریتم LDA*. پایان‌نامه کارشناسی ارشد. دانشگاه آزاد اسلامی واحد الکترونیکی. دانشکده فنی و مهندسی.

گزنی، علی. ۱۳۸۵. استخراج خودکار عبارت‌های کلیدی از متون مقاله‌های فارسی. *کتابداری و اطلاع‌رسانی* ۳۵: ۶۶-۷۴.

لنکستر، فردریک. ۱۳۸۸. *نمایه‌سازی و چکیده‌نویسی، مبانی نظری و عملی*. ترجمه عباس گیلوری. تهران: چاپار.

موسوی‌زاده، مریم. ۱۳۸۹. *بررسی ساختار گرایش‌های موضوعی مقالات تألیفی فارسی و انگلیسی حوزه سازماندهی اطلاعات از طریق تحلیل عنوان، چکیده و کلیدواژه‌ها؛ وزن‌دهی و تحلیل هم‌رخدادی اصطلاحات*. پایان‌نامه کارشناسی ارشد. دانشگاه الزهرا (س). دانشکده علوم تربیتی و روان‌شناسی.

نصیری، مهدی، و سارا اسماعیلی. ؟۱۳۹۴. *روش انجام پروژه داده‌کاوی (روش کریسپ)*.

نوروزی، علیرضا و خالید ولایتی. ۱۳۸۹. *نمایه‌سازی موضوعی: نمایه‌سازی مفهومی*. تهران: چاپار.

References

Fox, Robert. 2010. Digital Libraries: the Systems Analysis Perspective Mining the Digital Library. *OCLC Systems & Services: International digital library Perspectives* 26 (4): 232-238. 10.1108/10650751011087585 (accessed March 24, 2019).

Jiang, Dongyang & Wei Zheng. 2012. Research On the Selection of Feature Transfer Relations in Latent Semantic Indexing. *AASRI Procedia* : 680-685. 10.1016/j.aasri.2012.11.108 (accessed March 24, 2019).

Okerson, Ann. 2013. Text & Data Mining - A Librarian Overview. *IFLA WLIC 2013*. Singapore. http://library.ifla.org/252/1/165-okerson-en.pdf (accessed March 24, 2019).

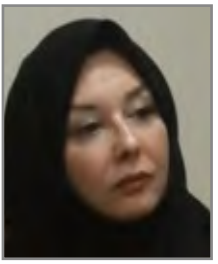
Tekli, J., Richard Chbeir, & Agma Traina. 2018. Full-fledged semantic indexing and querying model designed for seamless integration in legacy RDBMS. *Data & Knowledge Engineering* [Accepted Manuscript], 10.1016/j.datak.2018.07.007 (accessed March 24, 2019).

Rzhetsky, Andrey, Michael Seringhaus, and Mark Gerstein 2008. Seeking a New Biology through Text Mining. *Cell*. 134: 9-13. 10.1016/j.cell.2008.06.029 (accessed March 24, 2019).

حمیده جعفری پاورسی

متولد ۱۳۶۳، دارای مدرک دکتری در رشتهٔ علم اطلاعات و دانش‌شناسی از دانشگاه آزاد واحد علوم تحقیقات است. ایشان هم‌اکنون مسئولیت منابع دیجیتالی کتابخانهٔ تخصصی فرهنگستان هنر را بر عهده دارد.

هستان‌شناسی، داده‌کاوی، کتابخانه‌های دیجیتالی و بازیابی اطلاعات از جمله علایق پژوهشی وی است.



نجلا حریری

دارای مدرک دکتری در رشتهٔ علم اطلاعات و دانش‌شناسی است. ایشان هم‌اکنون استاد دانشگاه آزاد علوم و تحقیقات تهران است. نظام‌های بازیابی اطلاعات، رفتار اطلاع‌یابی، سازماندهی اطلاعات، مدیریت اطلاعات و پایگاه‌های اطلاعاتی از جمله علایق پژوهشی ایشان است.

مهدی علیپور حافظی

متولد ۱۳۵۲، دارای مدرک دکتری در رشتهٔ علم اطلاعات و دانش‌شناسی است. ایشان هم‌اکنون استادیار دانشگاه علامه طباطبائی است. محتوای دیجیتال، کتابخانه‌های دیجیتال، و بازیابی اطلاعات از جمله علایق پژوهشی وی است.

فهیمه باب‌الحوائجی

دارای مدرک دکتری در رشتهٔ علم اطلاعات و دانش‌شناسی است. ایشان هم‌اکنون دانشیار دانشگاه آزاد علوم و تحقیقات تهران است. معماری اطلاعات، ذخیره و بازیابی اطلاعات، اقتصاد اطلاعات، سواد اطلاعاتی و مدیریت دانش از جمله علایق پژوهشی وی است.

مریم خادمی

متولد ۱۳۴۶، دارای مدرک دکتری در رشتهٔ ریاضی کاربردی است. ایشان هم‌اکنون دانشیار گروه ریاضی کاربردی دانشکده فنی و مهندسی دانشگاه آزاد اسلامی واحد تهران جنوب است. داده‌کاوی، متن‌کاوی، حل مسایل بهینه‌سازی، کاربرد شبکهٔ عصبی مصنوعی از جمله علایق پژوهشی وی است.