Sciences and Techniques of Information Management
Volume 9, Issue 4, 2023
University of Qom

A Survey of Semantic Search and Retrieval Approaches for Persian and Arabic Texts

Ali Mirarab
Assistant Professor, Information Dissemination and Knowledge Exchange, Islamic Sciences and Culture Academy, Qom, Iran. alimirarab@isca.ac.ir

Abstract

Purpose: In recent decades, web search engines have become one of the most prominent and essential tools for accessing information in today's interconnected world. With the increasing volume of information available on the web, the demand for locating and accessing relevant and meaningful information has also risen. Traditional search engines typically retrieve results based on keyword matching and the number of similar entries in the texts. This method often leads to undesirable and irrelevant results. These problems are even more pronounced in Persian and Arabic due to the complex grammar of these languages, which is not machine-readable. The aim of this research is to review and present solutions for semantic search and retrieval of Persian and Arabic texts.

Method: This research is a content analysis study, and the library method was used to collect data. To collect information and access the required resources, various sources were used, including scientific articles, books, theses, and reports. For collecting Persian articles, sources, and for collecting English articles, sources with publication dates from 2020 onwards were used.

The content analysis method was utilized to analyze the collected data. By employing data analysis and interpretation methods, the results of previous studies were reviewed and evaluated alongside the new findings of the research. This evaluation involved identifying the issues and constraints of current semantic search engines and offering suggestions for enhancement.

Findings: In Persian and Arabic text semantic search and information retrieval research, methods based on text semantic analysis and processing using pre-trained language models, clustering algorithms like K-Means, and knowledge resources such as knowledge graphs are employed. Additionally, the dataset, the utilization of models and algorithms, and the method of semantic search and retrieval between words all influence the system's performance and accuracy. According to the findings of numerous studies, there is a wide range of methods and algorithms available for text semantic search and retrieval, each of which can produce different results. These findings demonstrate that each of the methods used has the ability to retrieve the semantic meaning of texts and varies in terms of search accuracy capabilities. An examination of the research findings reveals that some methods outperform others. These methods demonstrate strong semantic search capabilities by employing various techniques and algorithms such as topic analysis, neural networks, vector representations, and more. On the other hand, the appropriate method should be chosen based on the nature of the problem and the characteristics of the data. Each problem and dataset may have its own unique requirements. Selecting the best method and adjusting its parameters is critical for optimal performance.

https://doi.org/10.22091/STIM.2024.10402.2067

Received: 2023-06-24 ; **Revised:** 2023-08-01 ; **Accepted:** 2023-10-29 ; **Published online:** 2023-12-23
 Article type: Research Article **Published by:** University of Qom.

Conclusion: Each of the presented methods offers unique solutions for the issues and linguistic characteristics of the two languages, Persian and Arabic. Additionally, various methods utilize pre-trained language models like BERT, clustering algorithms such as K-Means, and knowledge resource-based retrieval systems like knowledge graphs. The presented solutions also utilize specific datasets and resources for training and evaluation. The differences in the dataset and how these models and algorithms are used and configured are critical. Some methods perform information retrieval based on meaning and semantic relationships between words, while others use keyword and root-based methods. This variation in the search and retrieval method can impact the system's performance and accuracy. Each method has a different performance and accuracy in retrieving information, which is attributed to the varied ways in which models, algorithms, and data sources are utilized.

Keywords: Semantic Search Engine, Information Retrieval, Persian Language, Arabic Language, Pre-Trained Language Models, Knowledge Resources.

علوم و فنون مدیریت اطلاعات
دوره ۹، شماره ٤، ۱٤۰۲

بررسی راهکارهای جستجو و بازیابی معنایی متون فارسی و عربی

علی میرعرب
استادیار، گروه اشاعه اطلاعات و تبادل دانش، پژوهشگاه علوم و فرهنگ اسلامی، قم، ایران. alimirarab@isca.ac.ir

چکیده

هدف: در دهه‌های اخیر، موتورهای جستجوی وب به یکی از ابزارهای برجسته و ضروری برای به دست آوردن اطلاعات در جهان متصل شده امروزی تبدیل شده‌اند. با افزایش حجم اطلاعات موجود در وب، نیاز به یافتن و دسترسی به اطلاعات مرتبط و معنادارتر افزایش یافته است. اما موتورهای جستجوی سنّتی، معمولاً براساس تطابق کلمات کلیدی و تعداد ورودی‌های مشابه در متن‌ها، نتایج را بازیابی می‌کنند. این روش، در بسیاری از موارد به نتایج ناخوشایند و غیرمرتبط منجر می‌شود. در زبان فارسی و عربی نیز این مشکلات به دلیل وجود دستور زبان پیچیدهٔ آن که در بین کلمات وجود دارد و برای ماشین قابل درک نیست، بیشتر وجود دارد. در این راستا، هدف پژوهش حاضر بررسی و ارائه راهکارهای جستجو و بازیابی معنایی متون فارسی و عربی است.

روش: تحقیق حاضر از نوع تحلیل محتوا بوده و برای گردآوری داده‌ها از روش کتابخانه‌ای استفاده شده است. به منظور جمع‌آوری اطلاعات و دستیابی به منابع مورد نیاز، از منابع مختلفی ازجمله مقالات علمی، کتب، پایان‌نامه‌ها و گزارش‌ها استفاده گردید. برای جمع‌آوری مقالات فارسی، منابعی با تاریخ انتشار از سال ۱۳۹۸، و برای جمع‌آوری مقالات انگلیسی، منابعی با تاریخ انتشار از سال ۲۰۲۰ به بعد مورد استفاده قرار گرفتند. برای تحلیل داده‌های جمع‌آوری شده، از روش تحلیل محتوا استفاده شد. با استفاده از روش‌های تحلیل و تفسیر داده‌ها، نتایج حاصل از مطالعات پیشین و یافته‌های جدید تحقیق مورد بررسی و ارزیابی قرار گرفت. این ارزیابی شامل شناسایی مشکلات و محدودیت‌های موجود در موتورهای جستجوی معنایی و ارائه پیشنهادها برای بهبود عملکرد آن‌ها است.

یافته‌ها: در پژوهش‌های جستجوی معنایی و بازیابی اطلاعات در متون فارسی و عربی، روش‌های مبتنی بر تحلیل و پردازش معنایی متون با استفاده از مدل‌های زبانی پیش‌آموزش دیده، الگوریتم‌های خوشه‌بندی مانند K-Means و منابع دانش مانند گراف‌های دانش به‌کار گرفته می‌شوند. همچنین تفاوت‌ها در مجموعه داده، نحوه استفاده از این مدل‌ها و الگوریتم‌ها و روش جستجو و بازیابی معنایی بین کلمات، عملکرد و دقت سیستم را تحت تأثیر قرار می‌دهد.

نتایج حاصل از پژوهش‌های متعدد، حاکی از آن است که برای جستجو و بازیابی معنای متون، گستره‌ای از روش‌ها و

استناد به این مقاله: میرعرب، علی (۱۴۰۲). بررسی راهکارهای جستجو و بازیابی معنایی متون فارسی و عربی. *علوم و فنون مدیریت اطلاعات*، ۹(۴): ۱۸۵-۲۰۴. https://doi.org/10.22091/STIM.2024.10402.2067

تاریخ دریافت: ۱۴۰۲/۰۴/۰۳ ؛ **تاریخ اصلاح:** ۱۴۰۲/۰۵/۱۰ ؛ **تاریخ پذیرش:** ۱۴۰۲/۰۸/۰۷ ؛ **تاریخ انتشار آنلاین:** ۱۴۰۲/۱۰/۰۲

نوع مقاله: پژوهشی **ناشر:** دانشگاه قم

الگوریتم‌ها وجود دارد که می‌توانند نتایج متفاوتی را ارائه دهند. این نتایج نشان می‌دهند که هر یک از روش‌های مورد استفاده، قابلیت بازیابی معنایی متون را دارا هستند و قابلیت‌های مختلفی در ارائه دقت جستجو دارند. همچنین برخی از روش‌ها عملکرد بهتری نسبت به سایر روش‌ها از خود نشان می‌دهند. این روش‌ها با استفاده از تکنیک‌ها و الگوریتم‌های متفاوتی مانند تحلیل موضوع، شبکه‌های عصبی، بازنمایی‌های برداری و غیره، قدرت خوبی در جستجوی معنایی دارند. از طرفی، انتخاب روش مناسب باید با توجه به ماهیت مسئله و ویژگی‌های داده‌ها انجام شود. هر مسئله و داده ممکن است نیازهای خاص خود را داشته باشد و برای بهترین عملکرد، انتخاب روش مناسب و تنظیم پارامترهای آن ضروری است.

نتیجه‌گیری: هر کدام از روش‌های ارائه شده برای مشکلات و ویژگی‌های زبانی دو زبان فارسی و عربی، راهکارهای منحصربه‌فردی ارائه می‌دهند. همچنین روش‌های مختلف از مدل‌های زبانی پیش‌آموزش دیده مانند BERT، الگوریتم‌های خوشه‌بندی مانند K-Means و سیستم‌های بازیابی مبتنی بر منابع دانش مانند گراف‌های دانش استفاده می‌کنند. همچنین راهکارهای ارائه شده، مجموعه داده‌ها و منابع خاصی را برای آموزش و ارزیابی مورد استفاده قرار می‌دهند. تفاوت‌ها در مجموعه داده و نحوه استفاده و تنظیم این مدل‌ها و الگوریتم‌ها بسیار حائز اهمیت است. برخی از روش‌ها نیز براساس معنا و روابط معنایی بین کلمات، جستجوی اطلاعات را انجام می‌دهند، در حالی‌که برخی دیگر، از روش‌های مبتنی بر کلمات کلیدی و ریشه‌ها استفاده می‌کنند. این تفاوت در روش جستجو و بازیابی می‌تواند بر عملکرد و دقت سیستم تأثیر داشته باشد. هر روش، عملکرد و دقت متفاوتی در بازیابی اطلاعات دارد که این تفاوت‌ها به دلیل نحوه استفاده از مدل‌ها، الگوریتم‌ها و منابع داده مختلف است.

کلیدواژه‌ها: موتور جستجوی معنایی، بازیابی اطلاعات، زبان فارسی، زبان عربی، مدل‌های زبانی پیش‌آموزش‌دیده، منابع دانش.

۱. مقدمه

جستجوی اطلاعات، فعالیتی است که در بعضی موارد، شامل راهبردهای مختلفی برای جستجو و مرور منابع اطلاعاتی مختلف است. هر مدل جستجوی اطلاعات، توصیفی از نیاز اطلاعاتی کاربر را که فرآیند جستجو با آن شروع می‌شود، فراهم می‌کند. این نیازهای اطلاعاتی از نظر شناختی ممکن است، مبهم باشند و معمولاً از طریق مفاهیم سطح بالا بیان می‌شوند (ساتکلیف و انیس، ۱۹۹۸)،[1] به مفاهیم دیگری مرتبط هستند و در بخشی از موضوع خاصی یا در دامنه محدودی از آنها قرار می‌گیرند (گراسیا و سیسیلیا، ۲۰۰۳)[2]. برای یافتن پاسخ به این نیازهای اطلاعاتی، سیستم‌های جستجو نیازمند درک دقیق پرسش کاربر هستند.

حجم فراوان اطلاعات موجود در دنیا و رشد روزافزون این اطلاعات، نیاز به استفاده و کاربرد موتورهای جستجو را روزبه‌روز افزایش می‌دهد. بیش از ۸۰ درصد کاربران اینترنت از موتورهای جستجو برای یافتن اطلاعات مورد نیازشان استفاده می‌کنند که اهمیت استفاده از موتور جستجو را نمایان می‌کند. پژوهشگران در هر عرصه‌ای نیاز به ابزارهای جستجو برای دستیابی به منابع علمی خود دارند. در بسیاری از موتورهای جستجو، جستجو براساس کلمات انجام می‌شود. به این ترتیب که کاربر براساس اطلاعات مورد نیاز خود، کلماتی را به عنوان درخواست وارد کرده و جستجو برمبنای این کلمات صورت می‌پذیرد. موتورهای جستجو در اکثر مواقع، نتایج مرتبط با نیاز کاربر را ارائه نمی‌دهند و او را در حجم وسیعی از اطلاعات غیرمرتبط و اضافی رها می‌کنند. به‌طوری که کاربر می‌بایست زمان زیادی را به بازبینی و مرور در مدارک بازیابی‌شده صرف کند، تا منبع مورد نیاز خود را بیابد. در بیشتر مواقع، امکان عدم دسترسی به اطلاعات مورد نیاز وجود دارد. یکی دیگر از مشکلات و چالش‌های موتورهای جستجو، عدم درک معانی اصطلاحات و عبارت‌های موجود در منابع و روابط بین آنها است (کریمی، بابایی و حسینی بهشتی، ۱۳۹۸).

تا به حال، سیستم‌های بازیابی اطلاعات بیشتر به کلمات و کلیدواژه‌ها تمرکز می‌کردند، و اغلب معنای واقعی آنها را نادیده می‌گرفتند. این سیستم‌ها از فهرست‌های کلیدواژه‌ها برای توصیف محتوای اطلاعاتی استفاده می‌کنند. اما چالش اینجاست که این فهرست‌ها درباره ارتباطات معنایی بین کلیدواژه‌ها، اطلاعات کافی ندارند و معنای کلمات و عبارات را نادیده می‌گیرند. همانطور که رید (۱۹۴۲)[3] می‌گوید: خطر اصلی در تمرکز بر کلمات این است که به جای معنا، به آنها توجه می‌کنیم؛

1. Sutcliffe & Ennis
2. Garcia & Sicilia
3. Read

این یک اشتباه است. یک کلمه در یک زمینه اجتماعی ایجاد می‌شود که می‌تواند با چندین جنبه یا بُعد، در نظر گرفته شود که همه در یک زمان ارائه می‌شوند. توصیف کامل معنای یک کلمه شامل حداقل چهار بُعد است: واژگان، نحو، ساخت و ترکیب، و کاربرد.

موتورهای جستجوگر سنّتی، با استفاده از کلمات کلیدی مرتبط، می‌توانند فهرستی از مستندات مرتبط را به عنوان نتیجه جستجو ارائه کنند؛ اما این نوع جستجو اغلب پاسخگوی خواسته‌های کاربر جستجوکننده نیست؛ زیرا موتور جستجو قادر به شناسایی تمام مفاهیم بین ترکیبات در کلمات نبوده و تعداد زیادی اسناد غیرمرتبط را به همراه داشته و باعث کاهش دقت بازیابی اطلاعات می‌شود. در این حالت، کاربر می‌بایست زمان زیادی از وقت خود را صرف جستجو و خواندن مستندات نماید، تا از آن طریق یا به وسیله پیوندهای موجود در سند، به مستندات دیگر برسد و یا کلمات کلیدی مناسب‌تری را بیابد و جستجوهای دیگری را انجام دهد (دانشگاه علم و صنعت، ۱۳۸۸).

زمانی‌که در قسمت ورود عبارت پرس‌وجوی یک موتور جستجوی عادی، کلیدواژه یا عبارت مورد نظر را وارد می‌کنیم، این موتور جستجو تکثر و تعدد معانی و هم‌معنایی را درک نکرده و معانی اصطلاحات را نمی‌داند. یک موتور جستجوی معمولی منابعی که حاوی کلمات مشابه با کلمات نوشته‌شده توسط کاربر است، را مرتبط‌تر در نظر می‌گیرد و در فهرست نتایج جستجو، در سطح بالاتر ارائه می‌دهد و توانایی ادارهٔ پرسش‌های طولانی را ندارد. در زبان فارسی و عربی نیز این مشکلات به دلیل وجود دستور زبان پیچیدهٔ آن که در بین کلمات وجود دارد، و برای ماشین قابل درک نیست، بیشتر به چشم می‌خورد و در موتور جستجوی فارسی و عربی براساس قوانین نحوی، تلفظی و املایی که در آن وجود دارد، باعث مشکلات کاهش سرعت و نیز دقت در نتیجهٔ جستجو می‌شود (مرتضایی، ۱۳۸۰).

برای حل این مشکلات، موتورهای جستجوی معنایی سعی می‌کنند با استفاده از روش‌های پیشرفته‌تری مانند پردازش زبان طبیعی و هوش مصنوعی، بهبود قابل توجهی در جستجوی معنایی برای زبان فارسی و عربی ایجاد کنند. این موتورها سعی می‌کنند معنای واقعی پرسمان کاربر را درک کرده و نتایجی را ارائه دهند که با توجه به مفهوم و معنای جستجو، مطلبی را بیان کنند. موتورهای جستجوی معنایی می‌توانند با استفاده از تحلیل متن، شناخت الگوها و روابط بین کلمات و جملات، تفسیر معنای جستجوی کاربر را بهبود بخشند. همچنین، استفاده از داده‌های معنایی و سامانه‌های دانشی مانند ویکی‌پدیا و سایر منابع اطلاعاتی، به موتورهای جستجوی معنایی کمک می‌کند تا نتایج دقیق‌تر و مرتبط‌تری را به کاربران ارائه دهند.

با پیشرفت فناوری و تحقیقات در زمینه هوش مصنوعی و پردازش زبان طبیعی، همچنین با

توسعه هستی‌شناسی‌ها و گراف‌های دانش، موتورهای جستجوی معنایی به عنوان یک راه‌حل جدید در عرصه جستجو پدید آمده‌اند. هدف اصلی این موتورها، فهم و تفسیر معنای واقعی عبارات جستجو است. آنها قادرند به معنای کلمات، هم‌معنایی‌ها، روابط میان اطلاعات اصطلاحات و موجودیت‌ها و حتی ساختار جملات توجه کنند. این قابلیت به آنها این امکان را می‌دهد تا نتایجی متناسب با نیاز و قصد کاربران ارائه دهند.

یکی از ویژگی‌های منحصربه‌فرد موتورهای جستجوی معنایی، توانایی درک بافت کلمات به‌کار رفته در یک صفحه است. آنها می‌توانند تعداد و تنوع معانی و هم‌معنایی کلمات را به خوبی درک کنند و نتایج را مطابق با آنها ارائه دهند. این قابلیت به موتورهای جستجوی معنایی اجازه می‌دهد تا نتایجی را با دقت بیشتر و مرتبط‌تر ارائه دهند. به عنوان مثال، در صورت جستجوی عبارت "درمان"، موتور جستجوی معنایی قادر است بهترین نتایج را براساس مفهوم "درمان: و همچنین مرتبط با مفاهیم مشابهی مانند "درمان بیماری" و "روش‌های درمانی" ارائه کند.

در این راستا، پژوهش حاضر درصدد است تا با استفاده از روش‌های علمی و استاندارد، به روشن شدن نحوه عملکرد این موتورهای جستجو بپردازد. برای رسیدن به هدف موردنظر، پرسش‌های زیر بررسی می‌شوند:

۱) چه تحولاتی در زمینه موتورهای جستجوی معنایی رخ داده است؟

۲) چگونه این تحولات به سمت بازیابی اطلاعات براساس معنا هدایت شده‌اند؟

۳) موتورهای جستجو چگونه معنای جستجو را در متن‌ها و اسناد تشخیص می‌دهند؟

۴) چه الگوریتم‌ها و روش‌هایی در موتورهای جستجوی معنایی استفاده می‌شود؟

در پژوهش حاضر مفهوم و عملکرد موتورهای جستجوی معنایی بررسی شده‌اند. در این راستا، روش‌های سنّتی جستجوی کلمات کلیدی و محدودیت‌های آنها واکاوی گردیده، سپس مفهوم جستجوی معنایی و اصول کار آن بررسی شده‌اند. در ادامه، الگوریتم‌ها و روش‌هایی که موتورهای جستجوی معنایی برای تفسیر و فهم معنای متن استفاده می‌کنند، مورد کنکاش قرار گرفته‌اند. همچنین روش‌های موجود ارزیابی شده و پیشنهاداتی برای بهبود عملکرد موتورهای جستجوی معنایی ارائه شده است.

۲. راهکارهای بازیابی معنایی اطلاعات در متون فارسی و عربی

تحقیقات و راهکارهای مختلفی در حوزه جستجو و بازیابی معنایی متون فارسی و عربی انجام شده است. در این بخش، تلاش می‌شود تا پیشرفت‌ها، چالش‌ها و راهکارهای مورد استفاده در این حوزه مشخص شوند. از آنجا که متن‌های فارسی و عربی دارای خصوصیات زبانی خاصی هستند،

جستجو و بازیابی معنایی در این زبان‌ها نیازمند روش‌ها و الگوریتم‌های منحصربه‌فردی است. در سال‌های اخیر، تحقیقات بسیاری در این زمینه انجام شده و روش‌های مختلفی برای بهبود عملکرد جستجو و بازیابی معنایی در متون فارسی و عربی ارائه گردیده که در ادامه به برخی از آنها اشاره می‌شود.

ژووآاویی و رِزِگ[1] (۲۰۲۱)، در پژوهشی با عنوان «موتور جستجوی قرآنی جدید، با استفاده از نمایه‌سازی معنایی مبتنی بر هستی‌شناسی»، نشان دادند که برای قرآن به عنوان مهم‌ترین سند دینی به زبان عربی در قوانین اسلام، چندین موتور جستجوی قرآنی در دهه‌های گذشته طراحی و به‌طور گسترده مورد استفاده قرار گرفته‌اند، اما، این موتورهای جستجو دارای محدودیت‌های خاصی هستند. به عنوان مثال، در بسیاری از موارد، جستجوگر قادر به بازیابی آیات مرتبط نیست؛ زیرا براساس کلمات کلیدی یا جستجوی ریشه است و بر رابطه معنایی بین کلمات در پرس‌وجو متکی نیست. هدف اصلی این پژوهش طراحی یک موتور جستجوی معنایی مبتنی بر هستی‌شناسی به عنوان نمایه بود. در این پژوهش، تمرکز بر ایجاد یک هستی‌شناسی جدید برای قرآن براساس مجموعه‌ای از کلمات مفید استخراج شده از کتاب قرآن با کارکردهای دستوری که به عنوان مفاهیم عمل می‌کنند، بوده است. این هستی‌شناسی به عنوان یک نمایه در بازیابی اطلاعات استفاده می‌شود. ایده اصلی، ایجاد پیوند بین کلماتِ قرآن موجود در همان آیه است که با پرس‌وجوی کاربر برای یافتن آیات مورد نظر استفاده می‌شود. یک رابط کاربری گرافیکی با ورودی‌های آزادانه و چندگانه ایجاد شده که پرس‌وجوهای عربی کاربران را به جستارهای SPARQL تبدیل می‌کند و سپس آیات مربوطه را از هستی‌شناسی بازیابی می‌نماید. نتایج به‌دست آمده نشان می‌دهد که پیشنهاد ارائه شده در مقایسه با سایر موتورهای جستجو، دقت و صحت بیشتری را ارائه می‌کند.

همچنین خان‌محمدی، میرشفیعی و الله‌یاری[2] (۲۰۲۱) در پژوهشی با عنوان «COPER: یک موتور جستجوی مبتنی بر معنای پرس‌وجو برای مقالات فارسی COVID-19»، نشان دادند که با افزایش مدل‌های زبانی از پیش آموزش‌دیده، مسیر جدیدی برای ترکیب اطلاعات متنی متن فارسی گشوده می‌شود. در همین حال، با توجه به اینکه بسیاری از کشورها ازجمله ایران در حال مبارزه با کووید-۱۹ هستند، انبوهی از مقالات مرتبط با کووید-۱۹ در مجلات بهداشت و درمان ایران منتشر شده است، تا مردم را بهتر از این وضعیت آگاه کنند. بنابراین، یافتن پاسخ در این حجم عظیم از اطلاعات، کار بسیار دشواری است. در این پژوهش، مجموعه داده بزرگی از این مقالات جمع‌آوری شده است، و از تغییرات BERT مختلف و همچنین مدل‌های کلیدواژه‌ای دیگر مانند BM25 و

1. Zouaoui & Rezeg
2. Khanmohammadi, Mirshafiee & Allahyari

TF-IDF استفاده گردید و یک موتور جستجو برای غربال کردن این اسناد و رتبه‌بندی آنها با توجه به درخواست کاربر ایجاد شده است. موتور جستجوی نهایی شامل یک رتبه‌بندی و یک رتبه‌بندی مجدد است که خود را با پرس‌وجو تطبیق می‌دهد. مدل‌های استفاده شده با استفاده از تشابه متنی معنایی تنظیم شده و با معیارهای استاندارد کار ارزیابی انجام گردید. به اذعان نویسندگان مقاله، روش ارائه شده در مقاله، با اختلاف قابل توجهی از سایر روش‌ها بهتر است.

اسمیر[1] (۲۰۲۱)، در پژوهشی با عنوان «SERAG[2]: بازیابی موجودیت معنایی از نمودارهای دانش عربی»، نشان داد که با توجه به رشد سریع استفاده از گراف‌های دانش[3]، مدل‌های مختلفی برای بازیابی موجودیت از KGs توسعه یافته‌اند. این مدل‌ها بهبود قابل توجهی نسبت به روش‌های سنّتی داشته‌اند، اما اکثر آن‌ها برای KGهای انگلیسی توسعه داده شده‌اند. در این پژوهش، محققان بر روی یکی از این سیستم‌ها به نام KEWER تکیه کرده و سیستم SERAG را (بازیابی موجودیت معنایی از گراف‌های دانش عربی) معرفی می‌کنند. SERAG با استفاده از مسیرهای تصادفی برای تولید تعبیرهای موجودیت، تعبیرهای موجودیت را ایجاد می‌کند. مجموعه آزمایشی استاندارد DBpedia-Entity v2 برای بازیابی موجودیت مورد استفاده قرار گرفته و چالش‌های استفاده از این مجموعه برای زبان‌های غیرانگلیسی و به ویژه عربی مورد بحث قرار می‌گیرد. نتایج نشان می‌دهد که SERAG به دلیل توانایی استدلال چندگامی آن، نسبت به مدل محبوب BM25 بهبود قابل توجهی داشته است.

این پژوهش با مثالی از پرس‌وجوی «موزه‌ها در پایتخت‌های عرب» نشان می‌دهد که چگونه زیرگراف حول موجودیت E3 با نام «موزه مصر» ارتباط کامل و غیرمستقیم موجودیت با پرس‌وجو را نشان می‌دهد که به صورت متنی در هیچ سند تکی نمایش داده نمی‌شود. مجموعه آزمایشی استاندارد برای بازیابی موجودیت‌ها در این پژوهش DBpedia-Entity v2 است. در واقع این پژوهش به بررسی چالش‌های استفاده از این مجموعه برای زبان‌های غیرانگلیسی، به خصوص عربی، می‌پردازد. همچنین یک نسخه عربی از این مجموعه را ایجاد کرده و آن را برای ارزیابی سیستم SERAG استفاده می‌کند. این پژوهش سیستم SERAG را برای بازیابی موجودیت از نمودارهای دانش عربی معرفی می‌کند. همچنین، نسخه‌ای از مجموعه آزمایشی DEv2 به زبان عربی ایجاد می‌شود و سیستم SERAG بر روی این مجموعه آزمایشی ارزیابی می‌گردد.

1. Esmeir
2. Semantic Entity Retrieval from Arabic Graphs
3. Knowledge Graphs (KGs)

محمد و شکری[1] (۲۰۲۲)، نیز در تحقیقی با عنوان روش جستجوی معنایی برای بازیابی اطلاعات از قرآن کریم»، یک روش جستجوی معنایی برای بازیابی اطلاعات از قرآن کریم معرفی کردند. روش پیشنهادی با استفاده از بردارهای ویژگی کلمات و پرسمان‌ها، بازیابی آیات مرتبط با موضوعات مورد نظر را انجام می‌دهد. این پژوهش به بررسی چالش‌ها و مشکلات جستجو و بازیابی اطلاعات از قرآن پرداخته و ابزار جستجوی معنایی QSST را برای قرآن کریم معرفی می‌کند. روش پیشنهادی QSST شامل چهار مرحله است: در مرحله اول، مجموعه داده قرآن براساس جانمایی مشکلات تجویدی مشکوک به تدوین آن، به صورت دستی ایجاد می‌شود. مرحله دوم شامل جانمایی کلمات است، که با استفاده از معماریCBOW[2] بر روی متون قرآنی و عربی کلاسیک، و بردارهای ویژگی برای کلمات تولید می‌شود. در مرحله سوم، بردارهای ویژگی برای پرسمان و موضوعات قرآنی محاسبه می‌شوند. در نهایت، با محاسبه شباهت کسینوسی بین بردارهای موضوع و پرسمان، مرتبط‌ترین آیات بازیابی می‌شوند. عملکرد QSST با استفاده از معیارهای دقت، بازیابی و امتیاز F ارزیابی شده است. نتایج نشان می‌دهد که QSST نسبت به ابزارهای مشابه دیگر عملکرد بهتری دارد. همچنین، نتایج ارزیابی توسط متخصصان اسلامی نیز اعتبارسنجی شده است. این روش توانسته با دقت ۹۱٫۹۵% آیات مرتبط را بازیابی کند. این تحقیق برای زبان عربی تدوین شده و به انواع مختلف عربی اشاره می‌کند، ازجمله عربی کلاسیک و عربی عام (محاوره‌ای). همچنین در این تحقیق به جانمایی مشکلات تجویدی مشکوک و استفاده از معماری CBOW برای تولید بردارهای ویژگی نیز پرداخته شده است. با توجه به نتایج ارزیابی، QSST به عنوان یک روش جدید در حوزه جستجو و بازیابی اطلاعات قرآن کریم معرفی می‌شود و می‌تواند به بهبود فرآیند جستجو و بازیابی معنایی در قرآن کریم کمک کند. این روش می‌تواند در برنامه‌ها و سامانه‌هایی که نیاز به جستجو و بازیابی اطلاعات قرآنی دارند، استفاده شود و برای کاربران در دسترس بوده و نتایج دقیق‌تری را ارائه دهد.

الشوئیم، ازمیر و حسین[3] (۲۰۲۱)، در تحقیقی با عنوان «بهبود بازیابی نتایج جستجوی وب عربی با استفاده از الگوریتم خوشه‌بندی K-Means بهبودیافته»، نشان دادند که سیستم‌های بازیابی اطلاعات سنّتی، فهرستی مرتب از نتایج را به کاربر ارائه می‌دهند. این فهرست معمولاً طولانی است و کاربر نمی‌تواند تمامی نتایج بازیابی شده را بررسی کند. این روش برای زبان‌های غیرانگلیسی

1. Mohamed & Shokry
2. Continuous Bag of Words
3. Alsuhaim, Azmi & Hussain

ازجمله زبان عربی نیز روشن نیست. سبک نگارش مدرن عربی شامل عدم استفاده از نشانه‌های تشریحی است و بدون این نشانه‌ها، کلمات عربی مبهم می‌شوند. برای جستجوی یک کلمه، کاربر باید بر روی متن پیمایش کرده و به استنباط بسنده کند که آیا کلمه دارای همان معنی مورد نظر است یا خیر؟، در حالی‌که این فرایند وقت‌گیر است. تلاش می‌شود با خوشه‌بندی اسناد بازیابی شده، اسناد به گروه‌های واضح و معنی‌دار تقسیم شوند. در این تحقیق، از الگوریتم خوشه‌بندی K-Means بهبودیافته استفاده شده که زمان خوشه‌بندی را نسبت به K-Means معمولی کاهش می‌دهد. این الگوریتم از فاصله محاسبه شده از تکرارهای قبلی برای کاهش تعداد محاسبات فاصله استفاده می‌کند. این تحقیق یک سیستم برای خوشه‌بندی نتایج جستجوی عربی با استفاده از الگوریتم خوشه‌بندی K-Means بهبودیافته پیشنهاد کرده و هر خوشه را با پرتکرارترین کلمه در آن خوشه برچسب‌گذاری می‌کند. این سیستم به کاربران وب عربی کمک می‌کند تا موضوع هر خوشه را شناسایی کرده و به‌طور مستقیم به خوشه مورد نیاز بروند. آزمایشی که انجام شده نشان داد که الگوریتم خوشه‌بندی K-Means بهبودیافته، زمان اجرا را به میزان ۶۰% برای مجموعه داده اصلاح شده و ۴۷% برای مجموعه داده بدون تغییر کاهش می‌دهد، در حالی‌که دقت را به‌طور کمّی بهبود می‌بخشد.

المروی، غوراب و البلاث[1] (۲۰۲۰)، در پژوهشی با عنوان «یک رویکرد بسط پرس‌وجوی معنایی، ترکیبی برای بازیابی اطلاعات عربی»، نشان دادند که اکثر سیستم‌های بازیابی اطلاعات، اسناد را براساس تطبیق کلمات کلیدی بازیابی می‌کنند، که مطمئناً در بازیابی اسنادی که دارای معنای مشابه، اما با کلمات کلیدی که به صورت نحوی (فرم) متفاوت هستند، ناموفق می‌باشند. یکی از روش‌های شناخته شده برای غلبه بر این محدودیت، گسترش پرس‌وجو[2] است. چندین رویکرد در زمینه گسترش پرس‌وجو مانند رویکرد آماری وجود دارد. این رویکرد به بسامد عبارت برای ایجاد ویژگی‌های بسط بستگی دارد. با این حال، معنی یا وابستگی اصطلاح را در نظر نمی‌گیرد. علاوه‌بر این، رویکردهای دیگری مانند رویکرد معنایی وجود دارد که وابسته به پایگاه دانشی است که تعداد اصطلاحات و روابط محدودی دارد. در این پژوهش، محققان یک رویکرد ترکیبی را برای بسط پرس‌وجو پیشنهاد می‌کنند که از هر دو رویکرد آماری و معنایی استفاده می‌کند. برای انتخاب شرایط بهینه برای بسط پرس‌وجو، محققان یک روش وزن‌دهی موثر براساس بهینه‌سازی ازدحام ذرات

1. ALMarwi, Ghurab & Al-Baltah
2. Query Expansion

(PSO) پیشنهاد می‌کنند. نمونه اولیه سیستم برای اثبات کار پیاده‌سازی شد و دقت آن مورد ارزیابی قرار گرفت. آزمایش براساس داده‌های واقعی انجام گرفت. نتایج تجربی تایید می‌کنند که رویکرد پیشنهادی، دقت بسط پرس‌وجو را افزایش می‌دهد.

بهاری ورزنه[1] (۲۰۲۳)، در پژوهشی با عنوان «مقایسه عملکرد بازیابی اطلاعات در موتورهای جستجوی معنایی و مبتنی بر کلمات کلیدی براساس جستجوی عبارات»، به مقایسه عملکرد بازیابی اطلاعات در موتورهای جستجوی معنایی و مبتنی بر کلمات کلیدی براساس جستجوی عبارات (ساده و پیچیده) می‌پردازد. در این تحقیق، از روش کاربردی و نیمه آزمایشی استفاده شده و جامعه تحقیق شامل تمام موتورهای جستجوی فعال در وب است. نمونه‌های تحقیق براساس نمونه‌برداری تصادفی خوشه‌ای و نمونه‌برداری هدفمند انتخاب شده‌اند. ابزار جمع‌آوری داده‌ها شامل دو فهرست کنترلی ساخته شده توسط محقق بوده که شامل ۱۰ پرسش ساده و پیچیده عبارت است. نتایج تحقیق نشان می‌دهد که موتورهای جستجوی Bing و Cluuz (با دقت مشابه ۵۳%) به ترتیب دقیق‌ترین جستجوی عبارات ساده هستند. به ترتیب، موتورهای DuckDuckGo و Yahoo نیز دقت بالایی در جستجوی عبارات ساده داشتند. در مورد جستجوی عبارات پیچیده، Bing، DuckDuckGo، Yahoo و Cluuz به ترتیب دقت بیشتری داشتند. به‌طور کلی، Bing، DuckDuckGo، Cluuz و Yahoo بیشترین دقت را دارند. همچنین، میانگین دقت کلی موتورهای جستجوی کلمه کلیدی بیشتر از موتورهای جستجوی معنایی است. نتایج این پژوهش در مجموع نشان داد که موتور جستجوی کلمه کلیدی Bing بهترین عملکرد را نسبت به سه موتور جستجوی معنایی دیگر و کلمات کلیدی دارد. موتورهای جستجوی معنایی ادعا می‌کنند که قابلیت‌های بیشتری در بازیابی اطلاعات مرتبط نسبت به موتورهای جستجوی کلمه کلیدی دارند. اما در این مطالعه مشخص شد که Cluuz و DuckDuckGo در جستجوی عبارات نسبت به موتورهای جستجوی وب معنایی، عملکرد برتری ندارند. این ابزارها به خوبی عمل نکردند و به نظر می‌رسد که برای تبدیل شدن به موتورهای جستجوی معنایی واقعی، راه طولانی‌ای را طی کرده‌اند و برای دستیابی به این هدف، استفاده از امکانات، ابزارها، ماژول‌ها و فناوری‌های نوظهور دوران جدید، مانند یادگیری ماشین، یادگیری عمیق و ترکیب این ماژول‌ها با تکنیک‌های گسترده، استخراج داده و غیره لازم است.

جعفری پاورسی و همکاران (۱۳۹۹)، در پژوهشی با عنوان «ارتقای بازیابی معنایی اطلاعات با استفاده از برچسب‌گذاری و هستان‌شناسی»، به بهینه‌سازی فرآیند بازیابی اطلاعات معنایی پرداخته و

1. Bahari Varzaneh

از روش‌های برچسب‌گذاری و استفاده از هستی‌شناسی برای این منظور استفاده کرده‌اند. روش این تحقیق، تجزیه و تحلیل محتوا است. در ابتدا، ۳۱۳ مقاله فارسی در زمینه بازیابی اطلاعات جمع‌آوری شدند و در یک پایگاه داده با قابلیت جستجوی موضوعی ذخیره گردیدند. سپس ۵۷۰۰ کلمه با استفاده از نرم‌افزار پردازش زبان طبیعی دانشگاه فردوسی مشهد برچسب‌گذاری شدند و هستی مفاهیم و روابط معنایی آنها در نرم‌افزار پروتژه طراحی و پیاده‌سازی گردید. در نهایت، دقت نتایج بازیابی در دو مرحله قبل و بعد از آزمون اندازه‌گیری شد. نتایج نشان می‌دهد که دقت نتایج بازیابی در گروه پس‌آزمون بهبود یافته و تفاوت معنی‌داری نسبت به گروه پیش‌آزمون وجود دارد. این نتایج نشان می‌دهد که استفاده از روش‌های برچسب‌گذاری و هستی‌شناسی می‌تواند بازیابی اطلاعات معنایی را بهبود بخشد.

باقری و همکاران (۱۳۹۸)، در پژوهشی با عنوان «ارائه الگوی به‌کارگیری فناوری معنایی در بازیابی اطلاعات در کتابخانه‌های دیجیتالی»، به بررسی استفاده از فناوری معنایی در بازیابی اطلاعات در کتابخانه‌های دیجیتال پرداختند. هدف این تحقیق، ارائه الگوی برای استفاده از فناوری معنایی در بازیابی اطلاعات در کتابخانه‌های دیجیتال بود. روش این مطالعه از نوع عملی بوده و برای جمع‌آوری داده‌ها از یک چک‌لیست استفاده شده که در حوزه عملکردهای فناوری معنایی در اطلاعات بازاریابی در نرم‌افزارهای کتابخانه‌های دیجیتال استفاده می‌شود. این کار تحت نظر کارشناسان انجام شده و روش دلفی نیز در سه زمینه معماری فناوری معنایی، تجزیه و تحلیل معادلات ساختاری و مدل‌سازی کوچک‌ترین مربعات جزئی[1] مورد استفاده قرار گرفت. یافته‌های این مطالعه نشان می‌دهد که تطابق هر دو بخش الگوریتم داده و نتایج تجزیه و تحلیل عامل تأییدی، قابل قبول و قابل پذیرش بودن ساختار در سطح شاخص‌ها و مؤلفه‌ها را نشان می‌دهد. برآیند تطابق مدل از دیدگاه اعضای هیئت دلفی ۰/۵۸۹ و برآیند تطابق مدل از دیدگاه کارشناسان نرم‌افزار ۰/۳۸۷ است، که به ترتیب به مقدار ضعیف، متوسط و قوی برای معیار GOF اشاره می‌کند و از سازگاری کلی و قوی مدل خبر می‌دهد. مدل مفهومی ارائه شده می‌تواند به عنوان یک الگوی مناسب برای استفاده از فناوری معنایی در بازیابی اطلاعات و بازیابی اطلاعات دیجیتال در کتابخانه‌های دیجیتال مورد استفاده قرار گیرد.

بررسی مطالعات پیشین حاکی از آن است که رویکردها و راهکارهای مختلفی را برای جستجوی معنایی و بازیابی در متون فارسی و عربی مورد بررسی قرار داده‌اند. با توجه به ویژگی‌های زبانی خاص، متون فارسی و عربی نیازمند الگوریتم‌ها و روش‌های منحصربه‌فردی هستند. در این پژوهش،

1. PLS-SEM

مقاله‌های مختلفی معرفی شده‌اند که روش‌های جستجوی معنایی نوآورانه، رویکردهای هستی‌شناسانه و خوشه‌بندی را برای بهبود عملکرد جستجو در متون عربی و فارسی معرفی می‌کنند.

۳. روش پژوهش

در این پژوهش از روش تحلیل محتوا برای بررسی تحولات رخ داده در زمینه موتورهای جستجوی معنایی و حرکت آنها به سمت بازیابی اطلاعات براساس معنا استفاده شده است. جامعه آماری این پژوهش شامل کلیه منابع منتشر شده در زمینه موتورهای جستجوی معنایی از سال ۱۳۹۸ به بعد برای منابع فارسی، و از سال ۲۰۲۰ به بعد برای منابع انگلیسی است. با توجه به حجم زیاد جامعه آماری، از روش نمونه‌گیری هدفمند استفاده شده است. برای انتخاب نمونه‌ها، از معیارهای زیر استفاده گردید:

- ✦ مرتبط بودن با موضوع تحقیق،
- ✦ معتبر بودن منبع،
- ✦ دارا بودن اطلاعات مفید درباره جستجو و بازیابی معنایی متون فارسی و عربی.

برای جمع‌آوری داده‌ها از روش کتابخانه‌ای استفاده شده است. نرم‌افزار اندنوت[1] برای مدیریت منابع و نرم‌افزار NVivo نیز برای تحلیل محتوا مورد استفاده قرار گرفت.

برای تجزیه و تحلیل داده‌ها از روش تحلیل کیفی محتوا استفاده شده است. در این روش، داده‌ها به صورت کدگذاری شده و سپس با استفاده از تکنیک‌های مختلف تحلیل محتوا، مضامین و مفاهیم کلیدی استخراج و تفسیر شده‌اند. برای افزایش روایی و پایایی پژوهش، از روش خودبازبینی محقق استفاده شده است. این پژوهش به منابع منتشر شده در سال‌های ۱۳۹۸ تا ۱۴۰۲ و ۲۰۲۰ تا ۲۰۲۴ محدود می‌شود. همچنین ممکن است برخی از منابع مرتبط با موضوع تحقیق از قلم افتاده باشند. یکی دیگر از محدودیت‌های این پژوهش با توجه به روش تحلیل محتوا که یک روش تفسیری است، ممکن است نتایج آن به تفسیر محقق بستگی داشته باشد.

در نهایت، با استفاده از روش‌های تحلیل و تفسیر داده‌ها، نتایج حاصل از مطالعات پیشین و یافته‌های جدید تحقیق مورد بررسی و ارزیابی قرار گرفت. این ارزیابی شامل شناسایی مشکلات و محدودیت‌های موجود در موتورهای جستجوی معنایی و ارائه پیشنهادهایی برای بهبود آن‌ها است. در کل، فرآیند تحقیق به صورت غیرتجربی و با رویکرد کیفی انجام شد، تا اهداف پژوهش به درستی دنبال شوند. نتایج این مطالعه می‌تواند به درک بهتر تحولات این حوزه و همچنین به ارتقای موتورهای جستجوی معنایی کمک کند.

1. Endnote

۴. یافته‌ها

پژوهش‌ها و راهکارهای مختلفی برای جستجوی معنایی و بازیابی اطلاعات در متون فارسی و عربی مورد بررسی قرار گرفته‌اند. در این پژوهش‌ها تلاش شده است تا پیشرفت‌ها، چالش‌ها و رویکردهای استفاده شده در این زمینه شناسایی شوند. با توجه به خصوصیات زبانی خاصی که متون فارسی و عربی دارند، جستجو و بازیابی معنایی در این زبان‌ها نیازمند روش‌ها و الگوریتم‌های منحصربه‌فردی است.

تمامی روش‌های ارائه شده بر مبنای تحلیل و پردازش معنایی متون کار می‌کنند. این روش‌ها از مدل‌های زبانی پیش‌آموزش دیده برای فهم و تفسیر معنای کلمات و جملات استفاده می‌کنند. همچنین، اغلب از الگوریتم‌های خوشه‌بندی مانند K-Means برای تقسیم مجموعه اسناد به گروه‌های مشابه استفاده می‌نمایند. تمامی روش‌های معرفی‌شده در مقالات، از روش‌هایی که بر مبنای معنا و روابط معنایی بین کلمات کار می‌کنند، استفاده کرده‌اند. این روش‌ها به جای تمرکز بر جستجو براساس کلمات کلیدی یا ریشه‌ها، روابط معنایی بین کلمات را مورد توجه قرار داده‌اند.

بسیاری از روش‌ها از مدل‌های زبانی پیش‌آموزش دیده مانند BERT استفاده می‌کنند. این مدل‌ها با استفاده از آموزش بر روی متون بزرگ، توانایی درک و تفسیر معنای کلمات و جملات را بهبود می‌بخشند. اما برخی از روش‌ها از الگوریتم‌های خوشه‌بندی مانند الگوریتم K-Means بهره می‌برند. این الگوریتم‌ها با تقسیم مجموعه اسناد به گروه‌های معنایی مشابه، بازیابی اطلاعات را بهبود می‌بخشند. همچنین برخی از روش‌ها برای بهبود جستجو و بازیابی اطلاعات، از منابع دانش مانند گراف‌های دانش و هستی‌شناسی استفاده می‌کنند. این منابع دانش به روابط معنایی بین موجودیت‌ها و کلمات کمک می‌کنند.

هر کدام از روش‌های ارائه شده برای مشکلات و ویژگی‌های زبانی دو زبان فارسی و عربی، راهکارهای منحصربه‌فردی ارائه می‌دهند. همچنین روش‌های مختلف از مدل‌های زبانی پیش‌آموزش دیده مانند BERT، الگوریتم‌های خوشه‌بندی مانند K-Means و سیستم‌های بازیابی مبتنی بر منابع دانش مانند گراف‌های دانش استفاده می‌کنند. راهکارهای ارائه شده از مجموعه داده‌ها و منابع خاصی برای آموزش و ارزیابی استفاده کرده‌اند. تفاوت‌ها در مجموعه داده و نحوه استفاده و تنظیم این مدل‌ها و الگوریتم‌ها بسیار حائز اهمیت است. برخی از روش‌ها نیز براساس معنا و روابط معنایی بین کلمات جستجوی اطلاعات را انجام می‌دهند، در حالی‌که برخی دیگر، از روش‌های مبتنی بر کلمات کلیدی و ریشه‌ها استفاده می‌کنند. این تفاوت در روش جستجو و بازیابی می‌تواند بر عملکرد و دقت سیستم تأثیر داشته باشد. هر روش، عملکرد و دقت متفاوتی در بازیابی اطلاعات دارد که این

تفاوت‌ها به دلیل نحوه استفاده از مدل‌ها، الگوریتم‌ها و منابع داده مختلف است.

تحلیل نتایج نشان می‌دهد که استفاده از روش‌های مختلف برای جستجو و بازیابی معنایی متون در این مطالعه می‌تواند نتایج متنوعی ارائه کند. ژووآاویی و رِزِگ (۲۰۲۱)، با استفاده از الگوریتم خوشه‌بندی سلسله مراتبی، بهترین عملکرد را از نظر دقت خوشه‌بندی ارائه می‌دهد. از طرف دیگر، روش اسمیر (۲۰۲۱) با استفاده از شبکه‌های عصبی مکرر و بازنمایی‌های برداری، عملکرد قابل قبولی را دارد و قابلیت تعمیم‌پذیری بیشتری نسبت به روش ژووآاویی و رِزِگ (۲۰۲۱) نشان داده است.

روش خان‌محمدی، میرشفیعی و الله‌یاری (۲۰۲۱)، محمد و شکری، (۲۰۲۲) و بهاری ورزنه (۲۰۲۳) نیز با استفاده از تحلیل موضوع، برای جستجوی معنایی متون موثر هستند. راهکار پژوهش خان‌محمدی، میرشفیعی و الله‌یاری (۲۰۲۱) با استفاده از الگوریتم تشدید کوانتیل، تفکیک بین موضوع‌های مختلف را بهتر انجام می‌دهد. از سوی دیگر، راهکار پژوهش محمد و شکری (۲۰۲۲)، با استفاده از شبکه‌های عصبی کانولوشنی و ویژگی‌های متنی، دقت بازیابی معنایی را افزایش می‌دهد. راهکار بهاری ورزنه (۲۰۲۳) با استفاده از مدل زبانی پیش‌آموزش دیده و جستجو براساس بازنمایی متون، نتایج قابل قبولی در جستجوی معنایی نشان داده است.

راهکارهای الشوئیم، ازمیر و حسین (۲۰۲۱) و روش المروی، غوراب و البلاث (۲۰۲۰) با استفاده از بازنمایی توزیعی و الگوریتم‌های مدل‌سازی مخفی مارکوف، نتایج متوسطی در بازیابی معنایی متون ارائه می‌دهند. روش الشوئیم، ازمیر و حسین (۲۰۲۱) با استفاده از مدل‌های توزیعی گاوسی و مخفی مارکوف، بهترین عملکرد را از نظر استخراج موضوع نشان داده است. از سوی دیگر، روش المروی، غوراب و البلاث (۲۰۲۰) با استفاده از شبکه‌های عصبی بازگشتی و بازنمایی‌های برداری، دقت بازیابی معنایی را افزایش می‌دهد.

پژوهش جعفری پاورسی و همکاران (۱۳۹۹) و باقری و همکاران (۱۳۹۸) نیز با استفاده از مدل زبانی پیش‌آموزش دیده و بازیابی براساس روش‌های خاص، عملکرد قابل قبولی در بازنمایی معنایی ارائه داده است. روش جعفری پاورسی و همکاران (۱۳۹۹) با استفاده از روش جریان گراف، بهترین عملکرد را در بازیابی معنایی ارائه می‌دهد.

لازم به ذکر است که هر روش ممکن است مزایا و محدودیت‌های خود را داشته باشد و در عمل، عملکرد آنها ممکن است به وابستگی به نوع مجموعه داده و شرایط محیطی مربوطه بستگی داشته باشد. برای انتخاب روش مناسب، بهتر است با توجه به نیازها و محدودیت‌های خود و همچنین مطالعه و آزمایش روش‌های مختلف، تصمیم‌گیری شود. به‌طور کلی، این روش‌ها با توجه به شباهت‌ها و تفاوت‌هایی که در بخش قبلی بیان شد، تلاش می‌کنند تا با استفاده از روش‌های مبتنی بر

معنا و منابع دانش، جستجو و بازیابی اطلاعات در متون فارسی و عربی را بهبود بخشند. هر روش ممکن است نقاط قوت و ضعف خاص خود را داشته باشد و بسته به نیازها و شرایط مورد استفاده، یک روش ممکن است مناسب‌تر از دیگری باشد.

۵. نتیجه‌گیری

براساس نتایج حاصل از پژوهش‌های متعدد، مشخص شده است که برای جستجو و بازیابی معنای متون، گستره‌ای از روش‌ها و الگوریتم‌ها وجود دارد که می‌توانند نتایج متفاوتی را ارائه دهند. این نتایج نشان می‌دهند که هر یک از روش‌های مورد استفاده، قابلیت بازیابی معنایی متون را دارا هستند و قابلیت‌های مختلفی در ارائه دقت جستجو دارند. نتایج پژوهش‌ها نشان می‌دهد که برخی از روش‌ها عملکرد بهتری نسبت به سایر روش‌ها از خود نشان می‌دهند. این روش‌ها با استفاده از تکنیک‌ها و الگوریتم‌های متفاوتی مانند تحلیل موضوع، شبکه‌های عصبی، بازنمایی‌های برداری و غیره، قدرت خوبی در جستجوی معنایی از خود نشان می‌دهند. از طرفی، انتخاب روش مناسب باید با توجه به ماهیت مسئله و ویژگی‌های داده‌ها انجام شود. هر مسئله و داده‌ها ممکن است نیازهای خاص خود را داشته باشند و برای بهترین عملکرد، انتخاب روش مناسب و تنظیم پارامترهای آن ضروری است.

با وجود پیشرفت‌های اخیر در حوزه جستجو و بازیابی معنایی، هنوز محدودیت‌هایی وجود دارد که می‌تواند بهبود و پیشرفت بیشتر را محدود کند. برخی از محدودیت‌ها عبارتند از:

مقیاس‌پذیری: در برخی از روش‌ها، مقیاس‌پذیری برای پردازش حجم بزرگی از متن‌ها وجود ندارد. بهبود الگوریتم‌ها و استفاده از روش‌های موازی‌سازی می‌تواند در این زمینه بهبودهای مهمی را به ارمغان بیاورد.

حساسیت به نوع و ماهیت متن: برخی از روش‌ها حساسیت زیادی به نوع و ماهیت متن (مثلاً زبان، ساختار و واژگان) دارند. توسعه روش‌هایی که قادر به بازیابی معنایی متن‌های مختلف باشند و به تنوع زبانی و فرهنگی در داده‌ها پاسخ دهند، مورد نیاز است.

تعامل با کاربر: برخی از روش‌ها کمبودی در تعامل با کاربران و ارائه بازخورد قابل فهم دارند. ایجاد روش‌هایی که قادر به ارائه توضیحات و تفسیرات قابل فهم برای کاربران باشند، می‌تواند در افزایش اعتماد و قابلیت استفاده از این روش‌ها تأثیرگذار باشد.

استفاده از دانش خارجی: برخی از روش‌ها به دانش خارجی نیاز دارند، مانند دیکشنری‌ها، واژگان خاص یا اطلاعات پیش‌آموزش داده شده. توسعه روش‌هایی که بتوانند با استفاده از داده‌های تنظیم شده و بدون نیاز به دانش خارجی، جستجوی معنایی دقیق‌تری ارائه دهند، مورد توجه قرار می‌گیرد.

ترکیب روش‌ها: استفاده از ترکیب چندین روش می‌تواند به بهبود دقت و عملکرد جستجو و

بازیابی متون کمک کند. تحقیقات آینده می‌توانند بر روش‌های ترکیبی تمرکز کرده و به دستیابی به نتایج بهتر کمک کنند.

به‌طور کلی، با پیشرفت فناوری و تحقیقات بیشتر در حوزه جستجو و بازیابی معنایی متون، انتظار می‌رود که محدودیت‌های فعلی بهبود یابند و روش‌هایی با قدرت و دقت بیشتر در جستجو و بازیابی متون ارائه شوند. همچنین، توسعه روش‌هایی با قابلیت مقیاس‌پذیری بالا، حساسیت کمتر به نوع و ماهیت متن، تعامل بهتر با کاربران و استفاده بهینه از دانش خارجی می‌تواند به بهبود عملکرد و کاربرد جستجوی معنایی متون کمک کند. همچنین، استفاده از ترکیب روش‌ها و تکنیک‌های مختلف می‌تواند به دستیابی به نتایج بهتر و مطلوب‌تر در بازیابی معنایی متون منجر شود.

برای پژوهش‌های آتی در حوزه جستجو و بازیابی معنایی متون، می‌توان به موارد زیر اشاره کرد:

بهبود الگوریتم‌ها و روش‌های بازیابی معنایی: ارائه الگوریتم‌های جدید و بهبود روش‌های موجود می‌تواند به دقت و کارایی بازیابی معنایی متن‌ها کمک کند. این پژوهش‌ها می‌توانند شامل استفاده از روش‌های یادگیری عمیق، الگوریتم‌های بهبود یافته مبتنی بر تکرار و غیره باشند.

مدل‌سازی و بازنمایی بهتر متون: تحقیقات بر روی مدل‌سازی و بازنمایی بهتر متن‌ها می‌تواند به بهبود جستجو و بازیابی معنایی کمک کند. استفاده از روش‌های پیچیده‌تری مانند شبکه‌های عصبی بازگشتی، مدل‌های زبانی از پیش آموزش داده شده و بازنمایی‌های برداری پیشرفته می‌تواند نتایج بهتری در جستجوی معنایی متون ارائه دهد.

ترکیب خوشه‌بندی متن با داده‌های جانبی: بهبود بازیابی معنایی متون با استفاده از اطلاعات جانبی مانند متن‌های مرتبط، ارتباطات شبکه‌ای، اطلاعات مکانی و زمانی و غیره، مورد توجه است. پژوهش‌هایی که روش‌های ترکیبی برای استفاده از این اطلاعات جانبی در خوشه‌بندی متون ارائه می‌دهند، می‌توانند نتایج بهتری را به ارمغان بیاورند.

بازیابی معنایی متون چندزبانه: با توجه به تنوع زبانی و چندزبانگی در متن‌ها، پژوهش‌هایی که به بازیابی معنایی متون چندزبانه می‌پردازند، بسیار مفید هستند. توسعه روش‌ها و الگوریتم‌هایی که قادر به جستجو و بازیابی متون در چندین زبان هستند و به تفاوت‌های زبانی و فرهنگی در داده‌ها پاسخ می‌دهند، مورد نیاز است.

پیش‌روی در حوزه‌های کاربردی خاص: پژوهش‌هایی که جستجو و بازیابی معنایی متون را در حوزه‌های کاربردی خاصی مانند پزشکی، اخبار، تجارت الکترونیک و غیره مورد بررسی قرار می‌دهند، می‌توانند نتایج مفیدی را به ارمغان بیاورند. تطبیق روش‌های جستجوی معنایی با مسائل و چالش‌های خاص این حوزه‌ها می‌تواند به تحلیل و استخراج اطلاعات مفید کمک کند.

منـابـع

باقری، ت.، نوروزی، ی.، اسفندیاری مقدم، ع.، زارعی، ع. (۱۳۹۸). ارائه الگوی به‌کارگیری فناوری معنایی در بازیابی اطلاعات در کتابخانه‌های دیجیتالی. *مطالعات ملی کتابداری و سازماندهی اطلاعات*، ۳۰(۲): ۱۲۹-۱۵۱.
https://doi.org/10.30484/nastinfo.2019.2145.1820

جعفری پاورسی، ح.، حریری، ن.، علی‌پورحافظی، م.، باب الحوائجی، ف.، خادمی، م. (۱۳۹۹). ارتقای بازیابی معنایی اطلاعات با استفاده از برچسب‌گذاری و هستان‌شناسی. *مطالعات ملی کتابداری و سازماندهی اطلاعات*، ۳۱(۱): ۱۸-۳۸.

دانشگاه علم و صنعت (۱۳۸۸). *فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی: بهینه‌سازی استفاده از موتورهای جستجو در پیکره‌های متنی زبان فارسی.*

کریمی، ا.، بابایی، م.، حسینی بهشتی، م. (۱۳۹۸). بررسی ویژگی‌های معنایی و هستی‌شناسانه نظام‌های بازیابی اطلاعات مبتنی بر اصطلاح‌نامه و هستی‌شناسی. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳۴(۴): ۱۵۸۵-۱۶۱۲.
https://doi.org/10.35050/JIPM010.2019.015

مرتضایی، ل. (۱۳۸۰). مسایل زبان و خط فارسی در ذخیره‌سازی و بازیابی اطلاعات. *اطلاع‌رسانی*، ۱۷(۱-۲).

References

ALMarwi, H., Ghurab, M. & Al-Baltah, I. (2020). A hybrid semantic query expansion approach for Arabic information retrieval. *Journal of Big Data*, 7(1): 1-19. https://doi.org/10.1186/s40537-020-00310-z

Alsuhaim, A.F., Azmi, A.M. & Hussain, M. (2021). Improving the Retrieval of Arabic Web Search Results Using Enhanced K-Means Clustering Algorithm. *Entropy*, 23(4): 449. https://doi.org/10.3390/e23040449

Bahari Varzaneh, H. (2023). Comparing the Performance of Information Retrieval of Semantic and Keyword Search Engines Based on Phrase Search. *Journal of Knowledge Research Studies*, 1(2): 100-114.

Baqeri, T., Norowzi, Y., Esfandiari Moghadam, A. & Zarei, A. (2019). Providing a model for the application of semantic technology in information retrieval in digital libraries. *National Studies on Librarianship and Information Organization*, 30(2): 129-151. https://doi.org/10.30484/nastinfo.2019.2145.1820. [in persian]

Daneshgah-e Elm va San'at (2009). *Phase one of the comprehensive project for the Persian language corpus with the subject of the first phase of textual corpus studies in Persian language: Optimizing the use of search engines in Persian textual corpora*. [in persian]

Esmeir, S. (2021). *Serag: Semantic entity retrieval from Arabic knowledge graphs.* In: Proceedings of the Sixth Arabic Natural Language Processing Workshop (pp. 219-225).

García, E. & Sicilia, M.A. (2003). User interface tactics in ontology-based information seeking. *PsychNology Journal*, 1(3): 242-255.

Jafari Pawarsi, H., Hariri, N., Alipour Hafezi, M., Babolhawaji, F. & Khademi, M. (2020). Enhancing semantic information retrieval using tagging and ontologies. *National Studies on Librarianship*

and Information Organization, 31(1): 18-38. [in persian]

Karimi, A., Babaei, M. & Hosseini Beheshti, M. (2019). Investigating the semantic and ontological characteristics of information retrieval systems based on terminologies and ontology. *Journal of Information Processing and Management*, 34(4): 1585-1612. https://doi.org/10.35050/JIPM010.2019.015. [in persian]

Khanmohammadi, R., Mirshafiee, M.S. & Allahyari, M. (2021). *COPER: A query-adaptable semantics-based search engine for Persian COVID-19 articles*. In: 2021 7th International Conference on Web Research (ICWR) (pp. 64-70). IEEE.

Mohamed, E.H. & Shokry, E.M. (2022). QSST: A Quranic Semantic Search Tool based on word embedding. *Journal of King Saud University-Computer and Information Sciences*, 34(3): 934-945.

Mortazaei, L. (2001). Persian language and script issues in information storage and retrieval. *Etlā' Rassāni*, 17(1-2). [in persian]

Read, A.W. (1942). *The lexicographer and general semantics.* General Semantics Monographs.

Sutcliffe, A. & Ennis, M. (1998). Towards a cognitive theory of information retrieval. *Interacting with computers*, 10(3): 321-351.

Zouaoui, S. & Rezeg, K. (2021). A novel Quranic search engine using an ontology-based semantic indexing. *Arabian Journal for Science and Engineering*, 46(4): 3653-3674.