Journal of Information Systems Engineering and Management
2025, 10(30s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

NLP-Driven Academic Research Management: A Catalyst for Organizational Research Information Retrieval

Dr.M. Madhu Bala[1], K. Akanksha[1*], Pooja Jain[1], K. Gayatri[1], L. Sai Prasad[1]
[1] *Department of CSE Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India*
**Corresponding Author: kavuriakanksha11@gmail.com*

ARTICLE INFO

Received: 30 Dec 2024

Revised: 05 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

The expeditious increase in research activities has resulted in the publication of a colossal amount of articles, papers, and journals authored by various researchers from diverse fields. The paper aims to build an efficient search engine that skilfully retrieves bibliographic information. The authorship recognition is critical in identifying the author for the given organization or institute and tracing out their contributions. This research acts as a catalyst to efficiently provide authors' and organization's contributions in the area of research in a systematic order. It involves the use of API scraping to extract data from websites and the application of data pre-processing techniques to facilitate data cleaning while a keyword-based matching method was employed to retrieve information based on context, the similarity is computed to know the relevance between the user query and the results given. The search engine is designed to aid in authorship recognition, enabling the identification of authors affiliated with specific organizations or institutes also while presenting their contributions to the field of research. The search engine successfully utilizes API scraping, pre-processing, and Natural Language Processing to efficiently retrieve bibliographic information enabling simplified access to relevant research information, becoming a valuable tool for institutes and organizations.

Keywords: API Scraping, Authorship recognition, Cosine Similarity, Keywords Extraction, Search Engine, Query matching

INTRODUCTION

The accelerated growth of research activities across various fields has led to the accumulation of articles, papers, and journals in repositories. This puts forward a challenge to individuals and organizations in identifying and accessing relevant and credible research articles, papers, or journals often a time-consuming process. At this juncture of the digital era, where the quantity of literature is overwhelming it is crucial to have a tool that can precisely retrieve and organize the research data to meet the specific needs of the user. Referring to this issue our paper emphasizes the development of an efficient search engine that skilfully retrieves pertinent bibliographic information based on the user query. The search engines use the scraped data to build and maintain structured datasets to present users with relevant results. Figure 1 illustrates the evolution of search engines over time.

Fig 1. Evolution of Search Engine

Authorship recognition serves a key role in this process of identifying authors and tracking and displaying their contributions to the respective fields. This not only traces out the author's research contribution but also highlights an organization's cohesive impact. The search engine employs NLP (Natural Language Processing) methods to process queries and generate relevant results while ensuring precision. The focal point of the search engine is the application of API scraping and data pre-processing techniques. Data is gathered from the sources Web of Science and Scopus, followed by integration of datasets, cleaning, and pre-processing to handle inconsistencies in data. Term

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Frequency-Inverse Document Frequency (TF-IDF) vectorization and cosine similarity metrics are employed to ensure accurate retrieval of results tailored to user queries.

The objective of our research is to simplify the process of accessing bibliographic information and research articles by presenting systematically ranked highly relevant results based on user preferences. It enables filter-based search to enhance search functionality. By employing advanced technologies, our research acts as a catalyst for individuals' and organizations' decision-making by presenting academic research information.

LITERATURE REVIEW

Data is crucial for business, organization, and research as it's the keystone for getting valuable insights, concluding, and assisting in decision-making and is present all over the internet [1]. Web scraping is the process of collecting large amounts of data in a brief duration of time with the least errors and can be employed through two kinds one being screen scraping and the other through API [3]. Web scraping enables the collection of diverse data from various sources leading to the collection of comprehensive datasets and traditional HTML Parsing; extracts data like text, images, and links through HTML code and, the Document object model (DOM) parsing extracts by navigating the Document object model tree and web scraping software are used to collect data [4]. UzunExt is a novel approach that uses a string-based method to search and calculate web pages and count HTML tags without DOM Tree and is 60 times faster than the DOM-based method [9]. Web scraping combined with Natural Language Processing (NLP) is a strong tool to extract data from unstructured data but web scraping also comes with legal and ethical implications due to breaches in data privacy a common problem faced by researchers [2]. Web scraping through screen scraping often results in unstructured data. API scraping facilitates the collection of semi-structured and rich datasets automatically and an API scraper is used in building a web data scraper [1]. A deep scraper was proposed to improve the crawling speed by suggesting a multiprocessing architecture where tweets are parallelly scraped through API and are seen to improve the crawling speed 2.37 times from using a normal standard API [5]. APIs enable easy access to the information on the websites but depend on owners' discretion and once permitted the issue of legal implications is eliminated [1].

A neural search engine is built by employing representation-focused neural models like a Deep Association Neural Network and a Deep Relevance Matching model for the ranking phase, but it resulted in having a higher cost from the latency and memory usage perspective, but they do offer better predictability [8]. Traditional NLP methods are also more cost-effective in comparison with neural models as they require less computational power and can be used to optimize the search. The NLP-based search engine named SciTechSE crawls academic websites and works on unstructured data employs Named Entity Recognition (NER) for identifying key entities, and organizes them in a hierarchical structure for intuitive filtering and search refinement [6]. The search tool can also be enhanced through the additional feature of automatic summarization of research articles. Summaries are generated through a Sentence rank algorithm and a pre-trained T-5 model while the search results are stored in CSV format with metadata as they help users get an outlook about the articles in offline mode paper aimed at refinement of the model in the future [7]. Personal Summarization involves extracting key features from user profiles using NLP methods like tokenization and parts-of-speech tagging. The Term Frequency-Inverse Document Frequency was applied to evaluate the relevance of the content and ranking algorithms were used to prioritize important information [13]. A novel ranking method named the Inverse Reinforcement ranking method was employed as the traditional ranking methods could not completely adapt to user preferences, in this novel approach a reward function represents the user's perceived utility and reinforces based on it and it has shown a remarkable improvement over the traditional ranking methods [10]. NLP-based search engines can be further enhanced by using semantic search methods to personalize search as the traditional search engines cannot completely understand the semantic meaning behind the user queries leading to incorrect search results. A semantic search engine using a Natural Language Process and Resource Description Framework was implemented to improve the relevance of search results through ontology mapping of user queries [12]. Semantic ontology-based analyzers like WordNet and advanced ranking algorithms like fuzzy membership values were employed to improve and achieve an accuracy of 97% and aim to refine the semantic relations model and handle issues of unstructured big data in the future [11]. The Intelligent search inclusive of a recommendation system enhances search by tailoring recommendations based on user queries. NLP and Machine learning are combined to develop a search engine and recommendation system using NLP techniques like tokenization, stop-word removal, and lemmatization. Latent Semantic Analysis and Word2Vec are applied to understand user intent and collaborative filtering and content-based filtering models are used for feature extraction and personalizing recommendations [14].

The traditional keyword-based document retrieval methods can be enhanced by apprehending the semantic relationship between the words which further increases relevancy [20]. The combination of Semantic-based search and keyword matching methods helps generate highly relevant search results and it outputs based on keywords and also the semantics of the words given by the user. Keywords and key phrases are crucial in analyzing large volumes of textual data and are highly used for content retrieval, query handling, summarization, etc [16]. Five statistical keyword extraction methods named Most Frequent Measure Based Keyword Extraction, Term Frequency-Inverse Sentence Frequency, Co-occurrence Statistical, Information Based Keyword Extraction, Eccentricity-Based Keyword Extraction, and, TextRank Algorithm were applied. Classification algorithms such as Naïve Bayes, Support Vector Machines, Logistic Regression, and Random Forest to classify the keywords obtained, and ensemble methods were employed to improve predictive performance [15]. These keyword extraction methods are used to identify key terms in user queries and ensemble methods help in enhancing the relevance ranking of search results. NLP methods are employed to extract relevant features including keywords and key phrases and these features are trained on the classification models like Support Vector Machine, Naïve Bayes, and Decision tree to classify the content and it is implemented to improve the accuracy of identifying relevant information [17]. An NLP-based keyword and key phrase extraction was implemented to improve the accuracy and relevance of extracting relevant words. Parts of speech tagging, Named entity recognition, and dependency parsing were used for keyword extraction [18]. A keyword recommendation system is used in a search engine to improve relevance and it employs term frequency analysis and semantic understanding for keyword extraction that are then ranked based on their relevance [19].

METHODOLOGY

The proposed work for developing an efficient and effective search engine that retrieves unerring and reliable bibliographic information and research papers based on user queries requires building search engine trails around five key steps: data collection, pre-processing, feature representation, query processing, filtering, and displaying query-based results. The system's process is illustrated through a structured visualization in Figure 2 that showcases its functionality.

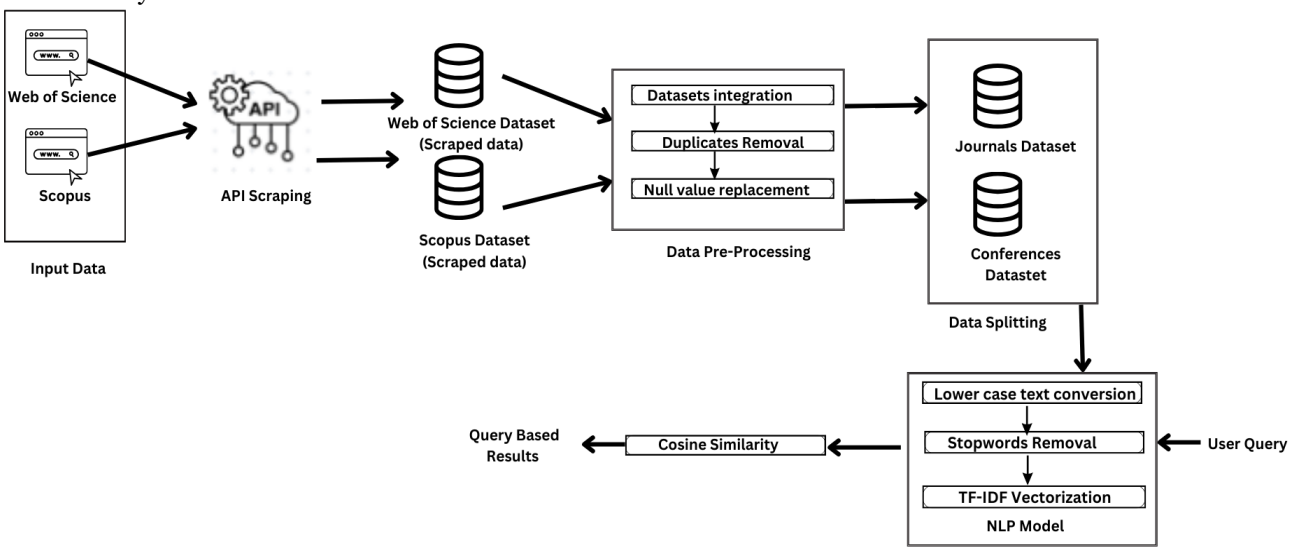

Fig 2. System Architecture for Search Engine

Every juncture of the procedure is meticulously planned to attain the functionality of being competent and effectual. The phases in development are elucidated below.

Data Description

The data collected for authorship recognition and tracking contributions of authors ought to have high-quality and comprehensive data on the crux of the matter, as they are the keystone for authorship recognition and tracking contributions across diverse research domains.

The dataset used in this study was constructed by unifying the bibliographic datasets collected from the Web of Science and Scopus; datasets were facilitated through API scraping. APIs provided by these websites were utilized to fetch metadata for research articles that generally include multiple attributes few of them include attributes such as

article title, authors, source title, publication year, and DOI link. The dataset collected is centric on the research articles from the institute "Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering &Technology" comprising 569 records from 1999 to 2024.The sample data is illustrated through Figure 3:

Authors	Source Title	Publication Year	DOI Link	Document Ty	Affiliations	Article Number	Times Cited (WosScore)
Sanduru, Bhanuteja;	COGENT ENGINEERING	2024	http://dx.doi.org/10.1080/2331	Article	Gandhi Institute of Technolog	2325028	1
Ramadevi, M.; Anura	PHYSICAL COMMUNICATION	2024	http://dx.doi.org/10.1016/j.phy	Article	National Institute of Technol	102489	0
Mirkute, Rushikesh;	INNOVATIVE INFRASTRUCTURE SOLU	2024	http://dx.doi.org/10.1007/s410	Article	Vallurupalli Nageswara Rao V	411	0
Ninama, Hitesh; Rail	SCIENTIFIC REPORTS	2024	http://dx.doi.org/10.1038/s415	Article	Devi Ahilya University; Intel	23019	0
Ramesh, M. R. Raja;	JOURNAL OF SOUTH AMERICAN EART	2024	http://dx.doi.org/10.1016/j.jsar	Article	Vishnu Institute of Technolog	105162	0

Fig 3. Sample Dataset

The datasets from two different sources are cleaned and merged to ensure the comprehensiveness of the data. The unique identifiers like the DOI link were harnessed to remove duplicate entries. The blend of Web of Science and Scopus datasets has uncloaked broad-gauged coverage of research publications. This phase resulted in a comprehensive dataset, encapsulating research publications across various domains.

Dataset Preparation and Pre-processing:

After merging, the pre-processing phase is crucial to handle the missing and inconsistent data. The missing values in columns such as Article Title and Authors were replaced with empty strings. Textual data in the article title and authors columns was then pre-processed, and it included:

- Lowercase conversion of textual data to standardize formatting.
- Stop word removal was employed to eliminate the commonly used and unnecessary English words called "stop words" using the NLTK library. The concatenation of the "Article Title" and "Authors" columns resulted in a new column and this concatenated textual representation accredited effective similarity-based retrieval.

NLP Model for Research Quest:

Feature Representation Using TF-IDF:

Converting textual data to numeric data for computational analysis was achieved by exercising the TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer from the Scikit library in Python. This vectorizer was then fitted to the concatenated column to produce a sparse matrix of TF-IDF features.

This approach is critical to ensuring the high relevancy of query matches during retrieval and the sparse matrix generated serves as the basis for similarity calculations because:

- Terms that are common and frequently occurring across the dataset, for example: "study", "research" or "paper" are assigned lower weights, as they are less informative and of the least importance.
- Rare, domain-specific terms for example: "neuro-informatics", "bioenergy" or "web scraping" are given higher weights, as they are meaningful and distinct, aiding in distinguishing the words.

Similarity Score:

User queries and filters like publication year were achieved through a Flask-based web interface. The workflow in query processing is discussed below:

- **Query Vectorization**: The user query is transformed into a vector using the pre-fitted TF-IDF vectorizer.
- **Computation of Cosine Similarity:** The cosine similarity between the previously generated TF-IDF matrix obtained from applying the vectorizer on the concatenated column and the user query vector is determined. Cosine similarity is a metric that calculates the cosine angle between the vectors. The similarity score and relevance between the retrieved data and user query are proportional, i.e., as the similarity score is higher, the relevance is also higher.
- The cosine similarity formula is illustrated below where 'A' and 'B' are vectors and ||A|| and ||B|| represent the magnitudes of the respective vectors.
- **Cosine Similarity formula:** $\frac{\vec{A}.\vec{B}}{||\vec{A}||||\vec{B}||}$

- The records are ranked in descending order of relevance based on their cosine similarity score, and the user is presented with the top five relevant records for their query.

Filtering and Result Presentation:

To enhance the search results an optional publication year filter is applied to customize the retrieval of the paper according to the user's interest. The metadata of the top-ranked articles are extracted and presented to the user. The metadata extracted for the top-ranked articles include:

- *Article Title:* Title of the article.
- *Authors:* The contributors to the article.
- *Source Title:* The journal or conference where the article was published.
- *Publication Year:* The year of publication.
- *DOI Link:* A distinct link to access the full article.
- *Similarity Percentage:* The similarity between the user query and retrieved data

The extracted results are returned in JSON format to simplify the integration with the front-end interface. The structured representation of the results retrieved ensures that users have an easy time accessing the relevant articles.

Implementation and Deployment:

The Flask framework was utilized to implement this application because it enables fast query processing and result generation. To input the user queries a home page interface was developed. This systematic procedure is worthwhile for academic research and helps organizations and individuals' access research articles with high relevance to their queries.

RESULTS

The proposed search engine effectively meets the objective of building a search engine for precise bibliographic information retrieval by employing a systematic approach involving data collection, pre-processing, and query processing and fetching top-ranked relevant results. The merged datasets offer a diverse coverage of research articles from diverse domains. The TF-IDF vectorizer efficiently differentiates between generic and domain-specific terms and cosine similarity aids in fetching top relevant bibliographic information of articles based on the similarity score enhancing the precision further. The results are organized and reduce the tedious work done by users while searching for research articles by presenting rank-based relevant information based on user queries. Overall it simplifies the access to relevant research information, becoming a valuable tool for institutes and organizations. The current implemented version of this project uses cosine similarity to find the similarity between the given input and the existing papers, and then sorts the similarity and displays the top 5 papers with the details about the paper and similarity percentage and this is shown in Figure 4, 5 and 6 below.

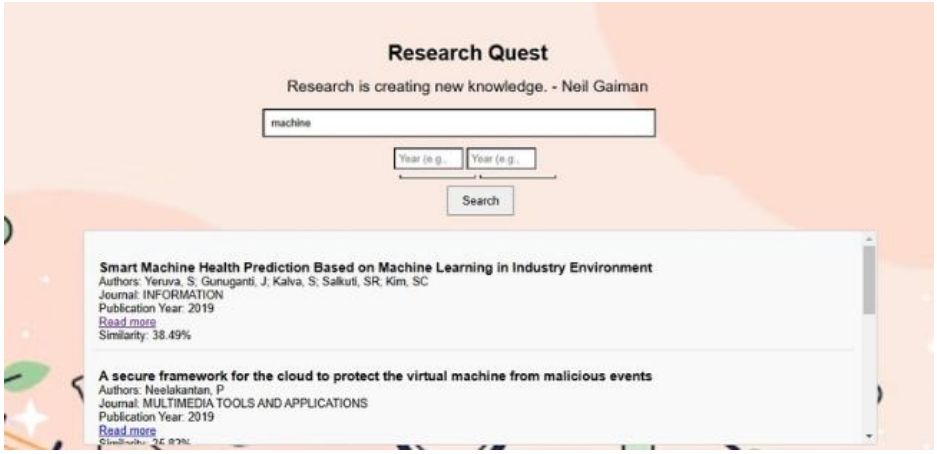

Fig 4. Top 2 results for the search 'machine' with similarity %

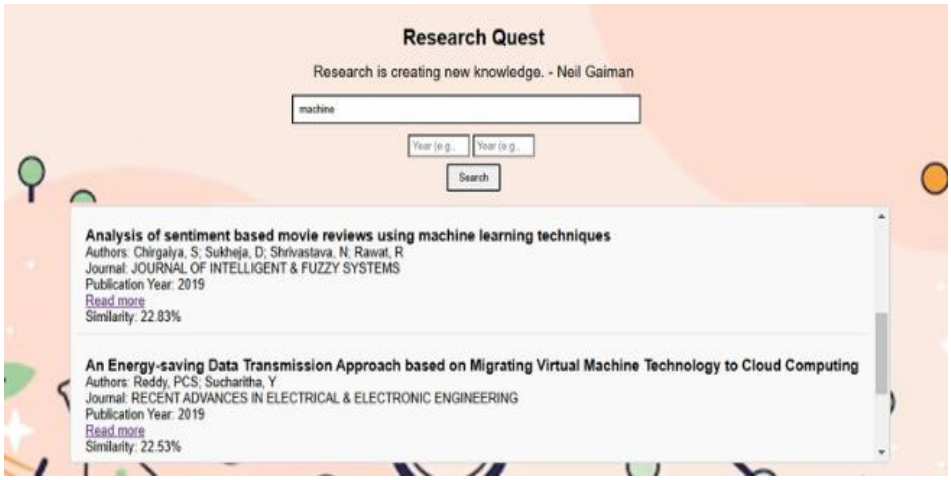

Fig 5. Top 3rd and Top 4th results for the search 'machine' with similarity %

Fig 6. Top 5th result for the search 'machine' with similarity %

DISCUSSION

Despite the potential of the search engine, it has certain setbacks reducing the effectiveness of the proposed system. The system depends on lexical keyword matching, lacking the contextual-based search that focuses on the context limiting the semantic search capabilities of the system. This hinders the ability to present relevant results and the absence of an intuitive dashboard to visualize the authors' bibliographic information and authors' contribution reduces its utility. The system does not support multiple languages restricting accessibility. The system does not offer a concise summary of the author when searched reducing the ability to give comprehensive insights.

The future scope of our research is to focus on improving the ability to visualize the author's and organization's contribution to research by providing a summary of the author and also through the creation of a comprehensive dashboard that traces the author's impact by visualizing key metrics like the number of citations, listing author's top cited paper, number of papers published in a year, etc. On the whole, providing key insights into understanding the author's performance and impact on an organization. This not only brings to the forefront the highly cited works but provides a clear view of the overall influence and trends observed in the authors' and organization's work in the field of research. We also aim at enhancing the search functionality by incorporating semantic search that not only generates results based on keywords but also the words matching with meaning and improving the accuracy of generating search results.

CONCLUSION

The proposed system efficiently retrieves bibliographic information by employing a systematic process involving data collection through API scraping, pre-processing, and query processing. Through TF-IDF vectorization and cosine

similarity computation, the system ranks the results based on relevance. It simplifies searching for articles becoming a valuable tool for researchers and institutes to access domain-related papers.

However, limitations like dependency on lexical-based matching, lacking semantic search ability, absence of interactive dashboards, and concise summary of the author are the areas of improvement of our system to enhance the utility of the search engine. Addressing these areas by implementing semantic search abilities, the interactive dashboard visualizes the bibliometric information of the author to provide comprehensive insights concerning the research trends, the contribution of the organization as well as an individual author in research. These enhancements to our proposed system help the academic and research community by providing expeditious and precise results simplifying the task of accessing relevant information.

REFRENCES

[1] Khder, M. A. (2021). Web scraping or web crawling: State of the art, techniques, approaches, and application. *International Journal of Advance Soft Computing and Applications, 13*(3). https://doi.org/10.15849/IJASCA.211128.11.

[2] Pichiyana, V., Muthulingam, S., Sathar, G., Nalajala, S., Ch, A., & Das, M. N. (2023). Web scraping using natural language processing: Exploiting unstructured text for data extraction and analysis. *3rd International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN 2023)*. https://doi.org/10.1016/j.procs.2023.12.074.

[3] Dogucu, M., & Çetinkaya-Rundel, M. (2021). Web scraping in the statistics and data science curriculum: Challenges and opportunities. *Journal of Statistics and Data Science Education*. https://doi.org/10.1080/10691898.2020.1787116.

[4] Sirisuriya, D. S. (2023). Importance of web scraping as a data source for machine learning algorithms - Review. *2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS)*. https://doi.org/10.1109/ICIIS58898.2023.10253502.

[5] You, J., Lee, K., & Kwon, H.-Y. (2024). DeepScraper: A complete and efficient tweet scraping method using authenticated multiprocessing. *Data & Knowledge Engineering, 149*. https://doi.org/10.1016/j.datak.2023.102260.

[6] Armentano, M. G., Godoy, D., Campo, M., & Amandi, A. (2014). NLP-based faceted search: Experience in the development of a science and technology search engine. *Expert Systems with Applications, 41*(6), 2886-2896. https://doi.org/10.1016/j.eswa.2013.10.023.

[7] Garg, S., Anand, P., Chanda, P. K., & Payyavula, S. R. (2024). An efficient summarization and search tool for research articles. *International Conference on Machine Learning and Data Engineering (ICMLDE 2023)*. https://doi.org/10.1016/j.procs.2024.04.210.

[8] Nakamura, T. A., Calais, P. H., Reis, D. de C., & Lemos, A. P. (2019). An anatomy for neural search engines. *Information Sciences, 480*, 339-353. https://doi.org/10.1016/j.ins.2018.12.041.

[9] Uzun, E. (2020). A novel web scraping approach using the additional information obtained from web pages. *IEEE Access*. https://doi.org/10.1109/ACCESS.2020.2984503.

[10] Karamiyan, F., Mahootchi, M., & Mohebi, A. (2024). Personalized ranking method based on inverse reinforcement learning in search engines. *Engineering Applications of Artificial Intelligence, 136*. https://doi.org/10.1016/j.engappai.2024.108915.

[11] El-Gayar, M. M., Mekk, N. E., Atwan, A., & Soliman, H. (2019). Enhanced search engine using proposed framework and ranking algorithm based on semantic relations. *IEEE Access*. https://doi.org/10.1109/ACCESS.2019.2941937.

[12] Yadav, U., & Duhan, N. (2021). Efficient retrieval of data using semantic search engine based on NLP and RDF. *Journal of Web Engineering*. https://doi.org/10.13052/jwe1540-9589.2084.

[13] Wang, Z., Li, S., & Zhou, G. (2017). Personal summarization from profile networks. *Frontiers of Computer Science*. https://doi.org/10.1007/s11704-016-5088-3.

[14] Balush, I., Vysotska, V., & Albota, S. (2021). Recommendation system development based on intelligent search, NLP, and machine learning methods. *CEUR Workshop Proceedings, 2917*. Retrieved from https://ceur-ws.org/Vol-2917/paper39.pdf.

[15] Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications, 57*, 232-247. https://doi.org/10.1016/j.eswa.2016.03.045.

[16] Siddiqi, S., & Sharan, A. (2015). Keyword and keyphrase extraction techniques: A literature review. *International Journal of Computer Applications*. https://doi.org/10.5120/19161-0607.

[17] Lalitha, T. B., & Sreeja, P. S. (2023). Potential web content identification and classification system using NLP and machine learning techniques. *International Journal of Engineering Trends and Technology, 71*(4), 403-415. https://doi.org/10.14445/22315381/IJETT-V71I4P235.

[18] Nesi, P., & Pantaleo, G. (2015). A distributed framework for NLP-based keyword and keyphrase extraction from web pages and documents. *21st International Conference on Distributed Multimedia Systems (DMS 2015)*. https://doi.org/10.18293/DMS2015-024.

[19] Wang, F., & Yu, L. (2024). The design of advertising text keyword recommendation for internet search engines. *Systems and Soft Computing*. https://doi.org/10.1016/j.sasc.2024.200109.

[20] Sharma, A., & Kumar, S. (2023). Ontology-based semantic retrieval of documents using Word2Vec model. *Data & Knowledge Engineering*. https://doi.org/10.1016/j.datak.2022.102110.