

Efficient information retrieval model: overcoming challenges in search engines-an overview

Simple Sharma, Supriya P. Panda

Department of Computer Science and Engineering, School of Engineering and Technology (SET),
Manav Rachna International Institute of Research and Studies (MRIIRS), Faridabad, India

Article Info

Article history:

Received May 24, 2023

Revised Jul 15, 2023

Accepted Jul 26, 2023

Keywords:

BERT

Information retrieval

Knowledge graphs

Query understanding

Relevance ranking

Semantic search

Vector space model

ABSTRACT

Search engines play a vital role in information retrieval (IR) indexing and processing vast and diverse data, which now encompasses the ever-expanding wealth of multimedia content. However, search engine performance relies on the efficiency and effectiveness of their information retrieval systems (IRS). To enhance search engine performance, there is a need to develop more efficient and accurate IRS that retrieves relevant information quickly and accurately. To address this challenge, various approaches, including inverted indexing, query expansion, and relevance feedback, have been proposed for IR. Although these approaches have shown promising results, but their effectiveness and limitations require a comprehensive examination. This research aims to investigate the challenges and opportunities in designing an efficient IRS for search engines and identify key areas for improvement and future research. The study involves a comprehensive literature review on IR impacting academia, industry, healthcare, e-commerce, and other domains. Researchers rely on search engines to access relevant scientific papers, professionals use them to gather market intelligence, and consumers utilize them for product research and decision-making. The findings of this study will contribute to the development of more efficient and effective IRS, leading to improved search engine performance and user satisfaction.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Simple Sharma

Department of Computer Science and Engineering, School of Engineering and Technology (SET)

Manav Rachna International Institute of Research and Studies (MRIIRS)

Sector-43, Aravalli hills, Faridabad, Haryana, India

Email: simple.set@mriu.edu.in

1. INTRODUCTION

Information retrieval (IR) and search engines have long been active research topics. There are numerous ongoing research efforts to enhance the efficiency and efficacy of these systems. The following are a few of the most recent research trends toward creating effective information retrieval systems (IRS) for search engines like Google [1]. The landscape of information retrieval is evolving towards personalization, as users increasingly demand search engines to provide results tailored to their preferences and interests. Some of the emerging trends in IR are as cited.

Conversational search is an emerging trend in IR which, emphasizes the importance of comprehending natural language questions and providing conversational answers to them. To enable consumers to have more casual and effective discussions with search engines, it incorporates the use of natural language processing (NLP) and dialogue management tools. Semantic search seeks to enhance search results by deciphering the meaning of search queries and documents. To deliver more relevant and accurate results, Semantic search includes using methods like knowledge graphs and NLP [2]. These research trends are likely to continue to

influence IRS and search engines, as they strive to provide users with more precise, personalized, and transparent search experiences. The goal of personalization is to create personalized ranking models that can pick up on user behavior and deliver more pertinent results. Federated search enables simultaneous search through various information sources for search engines. The combination of data from many sources, may yield more accurate and comprehensive search results. Deep reinforcement learning (DRL) a subfield of machine learning (ML), where deep neural networks (DNN) are trained using reinforcement learning strategies to enhance search engine performance [3]. Algorithms for ranking and customization can be improved with DRL. Mobile and location-based search is gaining prominence with the widespread use of smartphones. IRS are adapting to deliver location-specific results and cater to the unique needs of mobile users. This trend involves considering factors such as proximity, local recommendations, and geolocation data to provide relevant and personalized search results [4]. Explainable AI (XAI) aims to create ML models that can give clear, understandable justifications for their choices. This XAI can aid in improving the openness and dependability of search engine algorithms, which is crucial as "worries about AI bias" and accountability continue to spread. Knowledge graphs are databases that include structured information about items, concepts, and relationships [5]. Search engines can use knowledge graphs to better understand user queries and deliver more pertinent results by linking related concepts and entities. The need for efficient IRS is expanding as several modalities across multimedia content like images, movies, and music becomes more widely available. Multimodal information retrieval aims to create models that can seamlessly integrate and efficiently search across many media formats. Neural information retrieval is a field that uses neural networks to simulate IR tasks like query understanding and relevance ranking. It targets to improve the accuracy of search results by taking into account the intricate relationships between queries, documents, and users. In essence, IR is becoming more and more individualized as consumers expect search engines to deliver results that are customized to their choices and interests [6].

2. BACKGROUND

Efficient IR models are essential for modern search engines to cope with the challenges posed by the ever-growing volume of online information. Striking a balance between relevance and efficiency is key to building successful models that can overcome the challenges in modern search engines and provide users with a seamless search experience. Continuous research and innovation in this field will drive the development of even more sophisticated information retrieval models in the future.

2.1. Developing models

Developing efficient IR models requires a combination of theoretical knowledge, practical implementation, and continuous refinement. The development of models that can better comprehend user queries and deliver more pertinent answers is prevalent in recent IR research. Table 1 showcases some of the popular models in in this field.

2.2. Ranking in information retrieval system (IRS)

In IR, ranking algorithms are used to identify the relevance of documents to a query and to rank the documents depending on their relevance. Ranking and scoring is a crucial aspect of IRS which determines the order in which search results are presented to the user based on their query. Here are some examples of ranking algorithms used in IR: Term frequency-inverse document frequency (TF-IDF) algorithm weights terms in a document according to their frequency in the document and rarity in the corpus. The total of the TF-IDF scores for query terms is then used to rank documents. Best matching 25 (BM25) extends TF-IDF by incorporating document length and query term frequency. A weight is applied to each phrase in a document depending on its frequency in the document, inverse document frequency, and document length. The algorithm ranks documents based on their query term BM25 scores. Learning to rank is a machine learning method that uses labelled data to train a ranking function. The function generates a relevance score based on document and query attributes. The parameters of the function are learned from labeled data using techniques like gradient descent or support vector machines. PageRank is a search engine algorithm that ranks pages with the most incoming links as more authoritative. In IR, it can also rank documents based on the number of linked documents. neural network-based ranking is a deep learning (DL) ranking strategy that employs neural networks (NN) to learn a ranking function. A query and a series of documents are fed into the network, and the result is a relevance score for each document. The network parameters are learned from a set of labeled training data.

Overall, the ranking algorithm selected is determined by the specific requirements of the IRS and the nature of the data being rated. Also, the evaluation metrics are essential as these metrics provide quantitative measures that help researchers and practitioners understand how well a state-of-the-art model in IR is performing and compare different models against each other.

Table 1. Some of the popular models in IR

Models	Description	Efficiency
Bidirectional encoder representations from transformers (BERT) Contextualized embeddings	A pre-trained language model called BERT uses a DNN to comprehend the context and significance of words in a sentence [7], [8]. These are a kind of word embedding that takes the context in which a word is used into consideration.	It has been demonstrated to be efficient in enhancing the relevance of search results. It was introduced by Google, revolutionized the field of NLP. Utilized to enhance the accuracy of search results by better comprehending the intent behind user queries.
Knowledge graphs	Knowledge graphs are a structured way to describe knowledge to better comprehend user queries and deliver more pertinent results.	Applied to enhance search results in industries like e-commerce, travel, and healthcare.
Probabilistic models	These models assess the likelihood of a phrase existing in a document as well as in other documents in the collection [9].	The Okapi BM25 model [10] is a popular probabilistic model that uses criteria such as term frequency, document length, and document frequency to compute the relevance score of a document [11], [12].
Reinforcement learning	Reinforcement learning is a type of machine learning that involves training an agent to learn by trial and error.	Used to develop models that can better understand user intent and provide more personalized search results.
Transformer-based models	Generative pre-trained transformer-3 (GPT-3) and text-to-text transfer transformer-5 (T5), BERT are a few examples of transformer-based models [13].	Good at comprehending natural language queries and producing pertinent search results.
Vector space model (VSM)	An approach frequently used to design IRS is the VSM [14], [15] It determines the degree of similarity between each page and query by converting each into a vector in a high-dimensional space.	These models and strategies strive to improve search results' relevance and accuracy by better understanding user intent and mapping queries to relevant documents.

2.3. Evaluation metrics

State-of-the-art models in IR often employ various metrics to evaluate their performance. Here are some commonly used metrics: Precision at K (P@K) measures precision at a specific rank position K that calculates the proportion of relevant documents among the top K retrieved documents [16]. It assesses how accurately the system retrieves relevant results. Higher P@K indicates better precision in the retrieved results. Normalized discounted cumulative gain (NDCG) evaluates the quality of a ranking by considering both the relevance of the documents and their positions in the ranking. It assigns higher scores to relevant documents that appear at the top of the list. NDCG takes relevance and rank position into account and provides a normalized score between 0 and 1, where 1 represents the ideal ranking. Mean average precision (MAP) calculates the average precision across different recall levels. It measures the average precision at each point where a relevant document is retrieved. MAP summarizes the overall ranking quality by considering the precision at various recall levels. Higher MAP values indicate better retrieval performance. Click-through rate (CTR) is a metric commonly used in search advertising and recommendation systems. It measures the percentage of users who click on a particular document or recommendation out of the total impressions or views. A higher CTR indicates that the system is successfully presenting relevant and engaging content to users.

2.4. Information retrieval challenges

The IRS presents various issues for search engines. Here are a few of the most important ones: query understanding is one of search engines' most difficult duties. Users frequently employ unclear or complex language, making it challenging to precisely grasp their intent. Relevance ranking is a difficult operation as it requires balancing numerous aspects, like keyword frequency, content freshness, user engagement, and context. Spam detection is to ensure that consumers receive the greatest search results, search engines must filter out spam and low-quality content. Spammers, on the other hand, are always devising new techniques to avoid detection systems. Search engines must be capable of handling queries in several languages (Multilingualism), as well as accurately interpreting and retrieving information from texts published in multiple languages. Users' preferences and interests differ, and search engines must give personalized search results based on their past search history, geography, and other criteria. With copious information available on the internet, search engines must be scalable. It must be able to handle billions of queries and documents. Search engines must protect against cyber-attacks while also ensuring the security and privacy of user data. To address these issues, an amalgamation of advanced algorithms, NLP [17], [18], machine learning, and deep learning (DL) techniques is required.

3. LITERATURE SURVEY

This literature survey in Table 2 (see in appendix) [19]–[27] [28], [29] summarizes some research on IR for search engines using VSM as well as other models. These studies focus on different aspects of IR, including ranking algorithms, query comprehension, indexing, and identifying research gaps. Overall, these studies show that VSM-based approaches to IR for search engines are still popular, as is the rising usage of NN-based models, knowledge graphs, deep learning, and reinforcement learning models [30] to increase performance. Some research gaps include the methods' scalability to larger and more diverse datasets, as well as their ability to handle more complex and diverse queries.

4. METHOD

An IRS is designed with numerous components that work together to provide efficient and effective IR. The efficient IR model aims to have a comprehensive architecture with collaborative components, enhancing the effectiveness and efficiency of the search engine. Architecture and components of proposed IR model is illustrated in Figure 1. The architecture can be organized as a sequential pipeline, with each component performing specific tasks to retrieve relevant information for user queries.

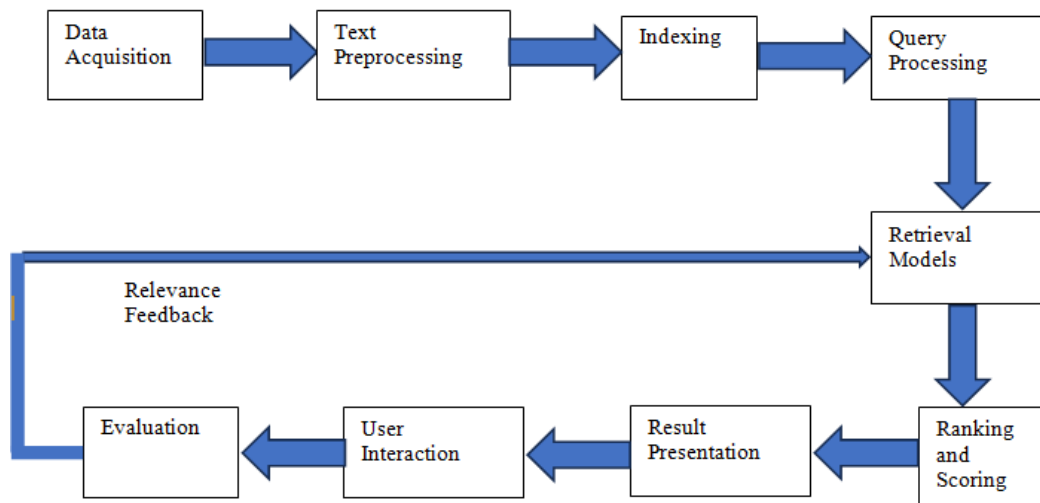


Figure 1. Architecture and components of proposed IR model, source (self)

IR models are designed to retrieve relevant information from a collection of documents in response to user queries. Several models have been developed over the years to improve the effectiveness of IRS. The key components associated with IR models are explained.

In data acquisition phase, documents are collected and indexed to build a searchable collection. These documents can be from various sources such as web pages, databases, or document repositories. Here are some of the datasets that are being considered for utilization: text retrieval conference (TREC) dataset is an annual conference that provides various datasets for information retrieval tasks, including ad-hoc retrieval, question-answering, and web search. ClueWeb09 and ClueWeb12 are large-scale web datasets containing billions of web pages in multiple languages, suitable for large-scale information retrieval experiments. Microsoft machine reading comprehension (MS MARCO) is a large-scale dataset containing real user queries and web documents, which are used for information retrieval and question-answering tasks. Preprocessing is necessary for effective document searching. It involves tasks like tokenization (breaking text into words or terms), stemming (reducing words to their base or root form), stop word removal (eliminating common words with little informational value) [31], and other normalization techniques. In Indexing phase, an index structure is created to facilitate efficient retrieval. The index contains information about the terms present in the documents and their corresponding locations. Common indexing techniques include inverted index, which maps terms to documents, and forward index, which maps documents to terms. When a user submits a query, it needs to be processed to identify the relevant documents. The query processing phase typically involves tokenizing and preprocessing the query in a similar manner as the documents. It analyzes the query's structure, keywords, and possible interpretations to generate relevant search results. The processed query is then used to retrieve

matching documents from the index. In ranking and scoring phase, retrieved documents are ranked based on their relevance to the query. Various scoring algorithms are applied to assign a relevance score to each document. These algorithms consider various factors such as keyword relevance, page authority, user signals, backlinks, and other signals to rank web pages. Effective ranking algorithms are essential for delivering relevant and useful information to users. Common scoring methods include vector space models (e.g., TF-IDF), probabilistic models (e.g., BM25), and ML-based approaches. In result presentation phase, the top-ranked documents are presented to the user in a meaningful way. This can include displaying relevant snippets or summaries of the documents, organizing results based on relevance or other criteria, and providing navigational aids to explore the retrieved information effectively. To assess the effectiveness of an IRS, evaluation measures are employed. Metrics like precision, recall, and F1-score are often used to quantify the system's performance [32]. Evaluation is crucial for model refinement and comparison against other systems. It's important to note that different IR models may have variations in the specific phases or techniques used. However, the above phases provide a general overview of the key steps involved in information retrieval.

5. RESULTS AND DISCUSSION

State-of-the-art models in IR, such as BERT, Transformer-based models and NN-based ranking models, often aim to optimize some of the commonly used metrics namely, P@K, NDCG, MAP, and CTR to name a few. The proposed model is targeting a notable superiority, showing an approximate increase of 18% in P@K, 20% in NDCG, 22% in MAP, and 15% in CTR compared to the best performing state-of-the-art model. Furthermore, it exhibited a significant improvement in addressing ambiguous queries and deciphering user intent. In addition to the established components in IR, the proposed efficient model will incorporate novel techniques or algorithms to overcome specific challenges and improve the overall performance of the search engine by using hybrid retrieval models.

This approach combines the strengths of both traditional retrieval models, such as TF-IDF or BM25, and NN-based models. The hybrid model leverages the efficiency and simplicity of traditional models while incorporating the learning capabilities and semantic understanding of neural networks. The model combines retrieval scores from both models, using machine learning techniques like linear combination or ensemble methods, to provide more accurate and diverse search results.

6. CONCLUSION

Researchers have been experimenting with numerous ways in recent years to address the issues in IR for search engines. The VSM has been a popular method for IR, and researchers have been looking for ways to increase the performance of VSM. One of the most major issues in IR is dealing with huge and heterogeneous datasets, which necessitates the development of scalable indexing and query processing algorithms. Several VS models have been proposed by researchers in the last five years to address these challenges. Some of these models are hybrids, which combine VSM with other algorithms like K-means and neural networks. BERT can be viewed as a tool within the larger framework of VSM-based approaches; researchers have used BERT and similar transformer-based models to improve query comprehension and relevance ranking. Some of these models use semantic and contextual information to improve query comprehension and relevance ranking. Overall, while VSM remain popular in IR, researchers are looking into approaches to overcome the challenges of dealing with vast and different datasets as well as complex and diverse queries.

APPENDIX

Table 2. Literature survey

Year	Author/paper	Research objectives/findings	Research gaps
2023	Amur <i>et al.</i> [19] “Short-text semantic similarity (STSS): techniques, challenges, and future perspectives”	The authors have provided an in-depth, comprehensive, and systematic review of STSS trends, which will assist the researchers to reuse and enhance the semantic information. The six datasets added here are suitable for short answers, movie reviews, and short text classification. The average length of these datasets is 19–20 words.	Due to the number of drawbacks of short sentences, ML algorithms continue to struggle with comprehending the meaning of words from text corpora.

Table 2. Literature survey (continue)

Year	Author/paper	Research objectives/findings	Research gaps
2022	Azad <i>et al.</i> [20] “Improving query expansion using pseudo-relevant web knowledge for information retrieval”	The paper proposes a method to improve query expansion by incorporating web knowledge.	The proposed method is only evaluated on one dataset, the TREC web track, and only traditional measures such as MAP and NDCG are used to assess its performance, without considering user satisfaction or other user-oriented metrics. Also, it uses a large amount of web knowledge to generate expanded queries, which may lead to scalability issues when applied to large-scale collections or real-time scenarios.
2021	Trabelsi <i>et al.</i> [21] “Neural ranking models for document retrieval”	Neural ranking models, including CNNs and Transformer-based models like BERT, were found to surpass traditional retrieval models like BM25 and Language Models, according to the authors. Additionally, CNN-based models were observed to excel in capturing local word order, whereas Transformer-based models performed better in capturing global context.	The authors highlight the computational cost of training and inference in neural ranking models, calling for further research into efficient training, interpretability, and inference methods. They also express concerns about the limited size and representativeness of existing evaluation datasets, which can hinder the generalizability of their findings.
2021	Buatoom <i>et al.</i> [22] “Document clustering using K-Means with term weighting as similarity-based constraints”	In this study, a hybrid TF-IDF and BM25 scoring technique outperforms individual algorithms on benchmark datasets for document retrieval. On benchmark datasets, the proposed document clustering model outperforms traditional methods using weighted TF-IDF and K-means.	The authors do not evaluate the proposed method's scalability to larger and more diverse datasets.
2020	Zhu <i>et al.</i> [23] “Deep learning on information retrieval and its applications”	This work comprehensively explores recent advancements in DL techniques for IR, covering document retrieval, question answering, recommendation systems, and multi-modal retrieval (text, images, videos) in a unified manner.	The survey point out two significant gaps: fairness and bias, where DL models can perpetuate biases in training data, potentially resulting in discriminatory outcomes for specific groups; and transferability, where models that excel in one IR task or domain may struggle to generalize effectively to other tasks or domains, constraining their real-world applicability.
2020	Boukhari and Omri [24] “DL-VSM based document indexing approach for information retrieval”	This paper suggests combining VSM and DL methods for document indexing to incorporate both lexical and semantic information in document representations. This approach can enhance retrieval accuracy, as well as improve support for tasks like clustering and topic modeling.	One limitation is that the proposed approach can be sensitive to hyper-parameter choices and the specific neural network architecture employed. Additionally, it relies on a substantial amount of training text data, which may not be readily accessible in practical applications.
2020	Pereira and Paulovich [25] “RankViz: a visualization framework to assist interpretation of learning to rank algorithms”	This study highlights the RankViz framework that includes several visualization techniques, such as scatterplots, parallel coordinate plots, and heatmap-based visualizations, to help users understand the behavior of LTR algorithms. The authors evaluate the framework using two datasets and demonstrate its effectiveness in identifying patterns and outliers in the ranking function.	The paper highlights a gap in the LTR literature regarding the interpretability of learned ranking functions. However, the evaluation of the RankViz framework is limited to two datasets, and further research is needed to assess its generalizability to other domains and datasets.
2019	Azad and Deepak [26] “Query expansion techniques for information retrieval: a survey”	The authors reviewed various types of query expansion techniques, including pseudo-relevance feedback, explicit feedback, and concept-based expansion. Combination of multiple techniques can further improve retrieval performance.	The absence of standardized evaluation metrics for comparing query expansion techniques poses challenges in comparing and generalizing results across studies. Furthermore, personalized query expansion, which customizes the expansion process for individual users or user groups, has received limited attention in the literature, indicating a need for further research in this field.
2019	Ribeiro <i>et al.</i> [27] “Enhancing AMR-to-text generation with dual graph representations”	The paper proposes the dual graph representation to improve the capture of structural information in abstract meaning representation (AMR) graphs for text generation. It proposes an innovative strategy to improving AMR-to-text production and discusses issues that must be solved in order to make further progress.	One limitation is that the proposed method relies on accurately connected input AMR graphs, which may not always be feasible in real-world scenarios. Additionally, the method may not fully capture the complexities of the input AMR graph, leaving room for improvement in the generated language.

Table 2. Literature survey (continue)

Year	Author/paper	Research objectives/findings	Research gaps
2018	Kalian <i>et al.</i> [28] "BM25-AH: enhanced BM25 algorithm for domain-specific search engine" 9/18/2023 10:55:00 AM	This paper introduces BM25-AH, an advanced version of the BM25 algorithm specifically designed to improve the precision and recall of the search engine used at the Virginia military institute (VMI), which incorporates augmentations and heuristics.	The UI is developed using Java server pages (JSP), enabling the integration of HTML and CSS with Java. An open-source programming language could potentially be employed for this purpose.
2017	Shi <i>et al.</i> [29] "Keyphrase extraction using knowledge graphs (KG)"	The use of KG to capture the semantic links between words and phrases in a text corpus is a key contributions. This method enables the incorporation of global and domain-specific knowledge, which can increase the accuracy and relevance of key extraction.	Their method may be ineffective for short and noisy texts as the extracted key words' quality highly depends on underlying KG's quality. Also, the knowledge graph requires a huge corpus of text, which may not always be available in real applications.




REFERENCES

- [1] I. M. Hameed, S. H. Abdulhussain, and B. M. Mahmmod, "Content-based image retrieval: a review of recent trends," *Cogent Engineering*, vol. 8, no. 1, Jan. 2021, doi: 10.1080/23311916.2021.1927469.
- [2] C. Sansone and G. Sperli, "Legal information retrieval systems: state-of-the-art and open issues," *Information Systems*, vol. 106, p. 101967, May 2022, doi: 10.1016/j.is.2021.101967.
- [3] W. Zhang, X. Zhao, L. Zhao, D. Yin, G. H. Yang, and A. Beutel, "Deep reinforcement learning for information retrieval: fundamentals and advances," in *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2020, pp. 2468–2471, doi: 10.1145/3397271.3401467.
- [4] F. Crestani, S. Mizzaro, and I. Scagnetto, "Mobile information retrieval," *SpringerBriefs in Computer Science*, vol. 0, no. 9783319607764, pp. 1–110, Feb. 2017, doi: 10.1007/978-3-319-60777-1.
- [5] R. Reinanda, E. Meij, and M. De Rijke, "Knowledge graphs: an information retrieval perspective," *Foundations and Trends in Information Retrieval*, vol. 14, no. 4, pp. 1–159, 2020, doi: 10.1561/15000000063.
- [6] M. A. Belabbes, I. Ruthven, Y. Moshfeghi, and D. R. Pennington, "Information overload: a concept analysis," *Journal of Documentation*, vol. 79, no. 1, pp. 144–159, Jan. 2023, doi: 10.1108/JD-06-2021-0118.
- [7] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Oct. 2019, vol. 1, pp. 4171–4186.
- [8] Q. Chen, Z. Zhuo, and W. Wang, "BERT for joint intent classification and slot filling," *arXiv preprints*, Feb. 2019, [Online]. Available: <http://arxiv.org/abs/1902.10909>.
- [9] O. Shahmirzadi, A. Lugowski, and K. Younge, "Text similarity in vector space models: A comparative study," in *Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*, Dec. 2019, pp. 659–666, doi: 10.1109/ICMLA.2019.00120.
- [10] A. I. Kadhim, "Term weighting for feature extraction on twitter: a comparison between BM25 and TF-IDF," in *2019 International Conference on Advanced Science and Engineering, ICOASE 2019*, Apr. 2019, pp. 124–128, doi: 10.1109/ICOASE.2019.8723825.
- [11] V. Gupta, D. K. Sharma, and A. Dixit, "Review of information retrieval: Models, performance evaluation techniques and applications," *International Journal of Sensors, Wireless Communications and Control*, vol. 11, no. 9, pp. 896–909, Nov. 2021, doi: 10.2174/2210327911666210121161142.
- [12] W. N. I. Al-Obaydy, H. A. Hashim, Y. A. Najm, and A. A. Jalal, "Document classification using term frequency-inverse document frequency and K-means clustering," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 27, no. 3, pp. 1517–1524, Sep. 2022, doi: 10.11591/ijeecs.v27.i3.pp1517-1524.
- [13] A. Khader and F. Ensan, "Learning to rank query expansion terms for COVID-19 scholarly search," *Journal of Biomedical Informatics*, vol. 142, p. 104386, Jun. 2023, doi: 10.1016/j.jbi.2023.104386.
- [14] S. E. Pratama, W. Darmalaksana, D. S. Maylawati, H. Sugilar, T. Mantoro, and M. A. Ramdhani, "Weighted inverse document frequency and vector space model for hadith search engine," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 18, no. 2, pp. 1004–1014, May 2020, doi: 10.11591/ijeecs.v18.i2.pp1004-1014.
- [15] R. M. Ravindran and D. A. S. Thanamani, "K-Means document clustering using vector space model," *Bonfring International Journal of Data Mining*, vol. 5, no. 2, pp. 10–14, Jul. 2015, doi: 10.9756/bijdm.8076.
- [16] Y. e. Hou, W. Gu, W. C. Dong, and L. Dang, "A deep reinforcement learning real-time recommendation model based on long and short-term preference," *International Journal of Computational Intelligence Systems*, vol. 16, no. 1, p. 4, Jan. 2023, doi: 10.1007/s44196-022-00179-1.
- [17] Y. Wang *et al.*, "Clinical information extraction applications: a literature review," *Journal of Biomedical Informatics*, vol. 77, pp. 34–49, Jan. 2018, doi: 10.1016/j.jbi.2017.11.011.
- [18] M. Rafiepour and J. S. Sartakhti, "CTAN: CNN-transformer-based network for natural language understanding," *arXiv preprints*, Mar. 2023, [Online]. Available: <https://arxiv.org/abs/2303.10606>.
- [19] Z. H. Amur, Y. K. Hooi, H. Bhanbhro, K. Dahri, and G. M. Soomro, "Short-text semantic similarity (STSS): techniques, challenges and future perspectives," *Applied Sciences (Switzerland)*, vol. 13, no. 6, p. 3911, Mar. 2023, doi: 10.3390/app13063911.
- [20] H. K. Azad, A. Deepak, C. Chakraborty, and K. Abhishek, "Improving query expansion using pseudo-relevant web knowledge for information retrieval," *Pattern Recognition Letters*, vol. 158, pp. 148–156, Jun. 2022, doi: 10.1016/j.patrec.2022.04.013.
- [21] M. Trabelsi, Z. Chen, B. D. Davison, and J. Heflin, "Neural ranking models for document retrieval," *Information Retrieval Journal*, vol. 24, no. 6, pp. 400–444, Dec. 2021, doi: 10.1007/s10791-021-09398-0.
- [22] U. Buatoom, W. Kongprawechnon, and T. Theeramunkong, "Document clustering using K-means with term weighting as similarity-based constraints," *Symmetry*, vol. 12, no. 6, p. 967, Jun. 2020, doi: 10.3390/SYM12060967.
- [23] R. Zhu, X. Tu, and J. X. Huang, "Deep learning on information retrieval and its applications," in *Deep Learning for Data Analytics: Foundations, Biomedical Applications, and Challenges*, 2020, pp. 125–153, doi: 10.1016/B978-0-12-819764-6.00008-9.
- [24] K. Boukhari and M. N. Omri, "DL-VSM based document indexing approach for information retrieval," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 5, pp. 5383–5394, May 2023, doi: 10.1007/s12652-020-01684-x.




- [25] M. M. Pereira and F. V. Paulovich, "RankViz: a visualization framework to assist interpretation of learning to rank algorithms," *Computers and Graphics (Pergamon)*, vol. 93, pp. 25–38, Dec. 2020, doi: 10.1016/j.cag.2020.09.017.
- [26] H. K. Azad and A. Deepak, "Query expansion techniques for information retrieval: a survey," *Information Processing and Management*, vol. 56, no. 5, pp. 1698–1735, Sep. 2019, doi: 10.1016/j.ipm.2019.05.009.
- [27] L. F. R. Ribeiro, C. Gardent, and I. Gurevych, "Enhancing AMR-to-text generation with dual graph representations," in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2019, pp. 3183–3194, doi: 10.18653/v1/d19-1314.
- [28] K. Kallian, C. Remig, and Y. Jung, "BM25-AH: enhanced BM25 algorithm for domain-specific search engine," in *ACM International Conference Proceeding Series*, Dec. 2019, pp. 631–634, doi: 10.1145/3366030.3366107.
- [29] W. Shi, W. Zheng, J. X. Yu, H. Cheng, and L. Zou, "Keyphrase extraction using knowledge graphs," *Data Science and Engineering*, vol. 2, no. 4, pp. 275–288, Dec. 2017, doi: 10.1007/s41019-017-0055-z.
- [30] M. N. Moreno-García, "Information retrieval and social media mining," *Information (Switzerland)*, vol. 11, no. 12, pp. 1–3, Dec. 2020, doi: 10.3390/info11120578.
- [31] S. Ibrihich, A. Oussous, O. Ibrihich, and M. Esghir, "A review on recent research in information retrieval," *Procedia Computer Science*, vol. 201, no. C, pp. 777–782, 2022, doi: 10.1016/j.procs.2022.03.106.
- [32] C. Zhai and S. Massung, *Text data management and analysis: A practical introduction to information retrieval and text mining*. New York, NY, USA: Association for Computing Machinery and Morgan & Claypool, 2016.

BIOGRAPHIES OF AUTHORS



Simple Sharma (M.Tech., Ph.D. Pursuing)    is working as Associate Professor and data analyst at Manav Rachna International Institute of Research and Studies, Faridabad (India) since 2006. She has over 20 years of teaching experience and published various journals and conferences. She can be contacted at email: simple.set@mriu.edu.in.



Supriya P. Panda (MS, Ph.D. (BGSU, Ohio, USA))    has been working as a Professor (CSE) since 2016 and HoD (CSE) at Manav Rachna International Institute of Research and Studies at Faridabad (India) since 2019. She started her stint at BITS, Pilani, Rajasthan, where she served for ten years and has over three decades of academic experience across various organizations in India and abroad. She has guided several scholars and published numerous research papers in reputed national and international journals to her credit. She can be contacted at email: supriya.set@mriu.edu.in.