Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 46 (2025), No. 5, pp. 1587–1603
DOI : 10.47974/JIOS-1839

Enhancing information retrieval through advanced query expansion techniques and semantic analysis

Hemendra Shanker Sharma *
Ashish Sharma †
Department of Computer Engineering and Applications
GLA University
Mathura 281004
Uttar Pradesh
India

Abstract

Optimizing the speed and precision with which useful information could be retrieved is the primary goal of Information Retrieval Enhancement. The authors discuss their motivation for creating better data retrieval methods in this work. The study investigates the use of semantic analysis and complex Query Expansion (QE) techniques to enhance the quality of search engine results. This research aims to enhance current techniques for retrieving data by using cutting-edge algorithms and semantic analysis tools. This study proposed the novel Bidirectional Encoder Representations from Transformers (BERT) and conditional random fields (CRF) to extract information from the Yahoo Dataset. Contextualized word embeddings with BERT and sequence labelling with CRF improve data analysis and extraction. The experimental results presented in the research demonstrate the effectiveness of these enhanced tactics in producing more precise and relevant search results by intelligently extending query terms and understanding the context of user searches. The results indicate that the proposed method exhibits a high degree of accuracy (99.7%), a substantial rate of recall (98.0%), a considerable rate of precision (99.8%), and a noteworthy value of F1-score (98%). Future research should focus on intelligent search engines that make it easy to access large volumes of data from several sources. These tools would use state-of-the-art algorithms for machine learning (ML) and methods for natural language processing.

Subject Classification: *Primary 68T50 Natural language processing, Secondary 68P20 Information storage and retrieval of data.*

Keywords: *Information retrieval enhancement, Semantic analysis, Query expansion, Natural language processing, Data retrieval.*

* *E-mail:* **hss.agra@gmail.com (Corresponding Author)**
† *E-mail:* **ashish.sharma@gla.ac.in**

1. Introduction

Daily activities all around the world, combined with the rapid development and appearance of breakthrough technology, have resulted in a massive flow of data. The massive amount of generated data has actionable insights powerful enough to enhance the service. Information retrieval is used to unearth the valuable insights buried behind a mountain of data [1]. An emerging method, information retrieval is commonly used to analyze data from big datasets that might be completely unstructured, somewhat organized, or completely structured. From the simplest Boolean model to more complex frameworks like probabilistic, vector space, and natural language modelling, a wide variety of models are used to execute information retrieval [2]. Figure 1 depicts an elementary process in the data retrieval system.

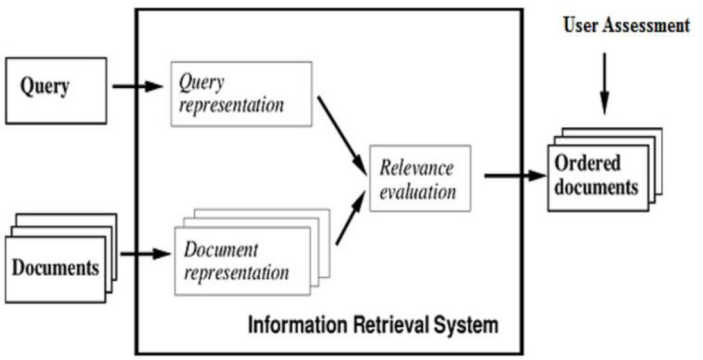

Figure 1

Information Retrieval system [3]

Due to the immensity of the Internet, it is increasingly difficult to find specific information using only a keyword or two in a search query. When it comes to improving online searches, QE is important [4]. In this case, the user's original inquiry is recast by adding related phrases. The field of information retrieval (IR) [5] has been engaged with QE for quite some time. It has made significant contributions to the study of cross-lingual IR, information filtering, multimedia IR, and personalized social documents. Conferences devoted solely to IR [6][7], such as the Text Information Retrieval Conference (TREC) [8] and Conference and Labs of the Evaluation Forum (CLEF) [9], have helped elevate the profile of QE research [10]. The IR system's primary focus is on documents that contain user-supplied

phrases, as this is where the closest match to the user's demands could be found. It is not uncommon for this basic retrieval model to make mistakes due to the imprecision of spoken language [11], especially when dealing with brief user inquiries [12, 13]

The rapid development of ontology-based techniques offers possibilities and efficient mechanisms for evaluating information at the semantic level, extracting, and refining the relations among individual concepts from large, unstructured documents [14-17] that could then be stored and organized in design ontology-based systems [18-20] for knowledge representation [21, 22, 53].

This study investigates commercial search engines that employ Boolean logic and keyword indexing, highlighting a significant limitation: keyword-based information retrieval relies on string matching and often overlooks semantic and statistical relationships between terms. Concept-based retrieval, in contrast, searches for words and phrases to better match user queries. To address query processing and information retrieval challenges, the author presents a general cognitive Query Expansion (QE) system. The study utilizes Named Entity Recognition (NER) and semantic addition to resolve document retrieval issues related to QE.

The novelty of the research lies in integrating advanced QE methods with semantic analysis, aiming to enhance precision and relevance in information retrieval. By leveraging machine learning and sophisticated algorithms, the study improves query understanding and incorporates diverse data sources, leading to more comprehensive search results. Additionally, the research proposes innovative methodologies for evaluating system performance and demonstrates practical applications across various domains, advancing state-of-the-art information retrieval technologies.

2. Related Work

A review of the literature analyzing the relevant work by different authors.

Jagerman et al. [23] proposed a method for query expansion, leveraging LLMs' generative capabilities. Experimentally demonstrated on MS-MARCO and BEIR, it was found that LLM-generated query expansions had the potential to outperform more conventional approaches. Russell-Rose et al. [24] explored a range of QE techniques using Boolean search algorithms, optimizing recall and precision using linguistic signals. Context-free distributional language models were found to be useful.

Gao et al. [25] introduced a retrieval model incorporating semantic matching signals from a neural embedding matching model to enhance standard lexical models like BM25. Their results outperformed state-of-the-art retrieval models and greatly improved reranking pipeline accuracy. AL Marwi et al. [26] proposed a QE method combining statistical and semantic techniques, utilizing particle swarm optimization (PSO) to select optimal phrases. Their results achieved 0.53062 recall and 0.4728 precision. Devi et al. [27] presented an ontologically based, scalable IR system that integrated semantic web and IR system benefits. Using M-KNN and the syntactic correlation coefficient (SCC), their system demonstrated high performance, with a 0.95 precision, 0.98 recall, and 0.97 F-measure. Zamani et al. [28] introduced a reinforcement learning model to generate and evaluate questions autonomously, with candidate answers achieving 87.3% good results on secondary result pages.

Qu et al. [29] refined a training method for dense passage retrieval, and experiments on MSMARCO and Natural Questions showed that RocketQA surpassed prior state-of-the-art models. Karpukhin et al. [30] demonstrated the effectiveness of dense representations in retrieval. Their dual-encoder framework outperformed previous models on four of five datasets, with a 1% to 12% improvement in exact match accuracy.

Safder et al. [31] developed a deep learning-based feature engineering method, using bi-directional LSTM on 37,000 algorithm-specific metadata lines. Their model achieved a 0.81 F1 score, surpassing the support vector machine by 9.46%. Liu et al. [32] explored neural network-based QE in code search, finding that Network Quality Evaluation + National Career Service (NQE + NCS) outperformed using NCS alone. Azad et al. [33] implemented a novel method for QE utilizing WordNet and other sources of data. In particular, the phrase-specific expansion words provided by Wikipedia and the individual-term rich expansion terms provided by WordNet were found to be complementary. Based on the results, the suggested strategy outperformed unexpanded searches on the FIRE dataset by 24% on the MAP score and 48% on the GMAP score. The experimental results surpassed related state-of-the-art techniques and individual expansion by a large margin. Khalifi et al. [34] introduced a search method combining machine learning and natural language pre-processing, achieving 99.2% accuracy and 96.2% recall.

3. Research Methodology

Design architecture is discussed in the framework of research techniques.

3.1 *Technique Used*

This research methodology incorporates a variety of techniques. These methods are described in detail below:

A. Learning-to-rank Algorithm

Learning-to-rank algorithms use supervised machine learning to solve ranking problems, initially designed for information retrieval tasks like document tracking. These algorithms are now widely used in Internet search and recommendation systems [35]. They enhance relevance ranking, personalize search results, and address query ambiguities through advanced QE and semantic analysis. By continuously adapting to user behaviour and content changes, these algorithms help search engines better understand user intent and preferences, leading to more accurate and personalized results [36, 37].

B. Analytic Hierarchy Process (AHP)

widely used decision-making machine learning state of the ar method that employs comparisons and expert opinions to construct priority scales. It is popular for its simplicity and effectiveness, making it a valuable tool for both researchers and decision-makers [38]. AHP is particularly useful in information retrieval, where it systematically evaluates and prioritizes criteria, compares alternative techniques, and facilitates complex decisions. By breaking decisions into pairwise comparisons and incorporating subjective judgments, AHP ensures transparency, consistency, and stakeholder involvement, leading to more informed decisions and improved retrieval outcomes [39].

C. Semantic Addition

Word sense disambiguation, sentiment classification, and topic sentiment analysis heavily depend on accurate semantic similarity measurements between words, making this a critical issue in natural language processing. Applications extend to text retrieval [40], Gene Ontology identification, medical language measurements, and biological entity recognition [41]. Semantic resources like ontologies and dictionaries are commonly used for calculating semantic similarity [42]. Key methods for adding meaning include Word WSD, WordNet, and QE.

- Word Sense Disambiguation (WSD)

Word sense disambiguation, sentiment classification, and topic sentiment analysis rely heavily on precise semantic similarity measurements between words, positioning this as a crucial challenge in natural language processing. Applications span across text retrieval [40], Gene Ontology identification, medical terminology evaluation, and biological entity recognition [41]. Ontologies and dictionaries, as semantic resources, are frequently used to compute semantic similarity [43, 44]. Prominent methods for enhancing meaning include WSD, WordNet, and QE.

- Query expansion (QE)

QE enhances a user's search by adding related phrases or terms [45]. It is employed to refine queries by incorporating synonyms, related terms, or conceptually similar words, improving the relevance of search results. By addressing vocabulary mismatches, ambiguity, and polysemy, QE aims to boost recall, enhance precision, and deliver more accurate, comprehensive search outcomes. This approach increases the effectiveness of information retrieval systems and enhances user satisfaction by providing more relevant and thorough results.

To improve the efficiency of the information-gathering process, QE restructures the user's initial query. Let Q = {t_1, t_2, ..., t_i, t_{i+1}, ..., t_n} represent an n-term user query. There are two parts to a reformulated query: Elimination of stop words T″= {ti+1, ti+2…, tn} the inclusion of new terms T′ = {t'_1, t'_2, ..., t'_m} from the data source (D). It is possible to express the revised query as [45]:

$$Q_{\exp} = (Q - T'')UT' = \{t_1, t_2, \ldots, t_i, t'_1, t'_2, \ldots, t'_m\} \quad (1)$$

The key element of QE is the set T′, which involves adding a meaningful collection of phrases to a user's search query to enhance content retrieval and reduce query ambiguity. As reported by Karovetz and Croft, generating T′ based on term similarity, without altering the core concept, can significantly boost recall in query results. Thus, two critical aspects of QE research are the calculation of set T′ and the selection of appropriate data sources D [45].

- Word net

WordNet is employed to enhance the effectiveness of information retrieval systems by leveraging its extensive lexical database and semantic

relationships. The study aims to improve QE techniques by incorporating synonyms and related terms, thereby capturing a broader range of relevant documents. Additionally, WordNet supports semantic analysis, helping to disambiguate ambiguous terms and providing insights into word meanings and relationships. This structured representation of word knowledge improves retrieval precision by aligning search queries more closely with intended concepts, ultimately enhancing overall retrieval performance [46, 47].

WordNet has been utilized in various applications, including information retrieval, automatic text categorization, automatic text summarization, machine translation, and the development of computerized crossword puzzles.

D. Named Entity Recognition (NER)

Named Entity Recognition (NER) enhances precision and semantic understanding in information retrieval by accurately identifying named entities, such as people, organizations, and locations, within queries and documents. This capability improves the interpretation of user intent and context, facilitating better QE with semantically related entities for more relevant search results. Additionally, NER supports entity linking to knowledge bases, enriching search outcomes with contextual information. Overall, integrating NER significantly boosts the effectiveness and relevance of information retrieval systems [48].

- BERT+CRF classifier

An innovative BERT+CRF hybrid model has been developed to extract entities and relations concurrently, thereby enhancing the healthcare data chain [49]. This model combines BERT's contextual understanding with CRF's ability to capture structured dependencies within sequences. The integration facilitates improved QE and semantic analysis, as BERT effectively comprehends the context and semantics of queries or documents, while CRF aids in identifying relevant terms or entities and extracting structured information. As a result, the retrieval system can better understand user queries and documents, leading to increased relevance and effectiveness in search results [50].

3.2 *Proposed Methodology*

A comprehensive procedure for query optimization and data retrieval involves several key steps. Initially, a query is used to access the database,

followed by pre-processing, which includes normalizing, tokenizing, and part-of-speech (POS) labelling the data. The algorithm then ranks and weights query phrases based on their importance using the AHP. A learning-to-rank algorithm assesses the relative importance of each term to identify the most suitable terms for QEs. The highest-ranked keywords are selected for further QE procedures. Semantic techniques such as WSD, QE, and WordNet enhance the query's context, with WSD aiding in information extraction and QE reducing query-document mismatches. WordNet is utilized for various applications, including automatic text classification and summarization. NER with a BERT+CRF classifier is employed to classify entities in texts. The process concludes with the system retrieving relevant documents based on the revised queries. Figure 2 illustrates the steps involved in queries.

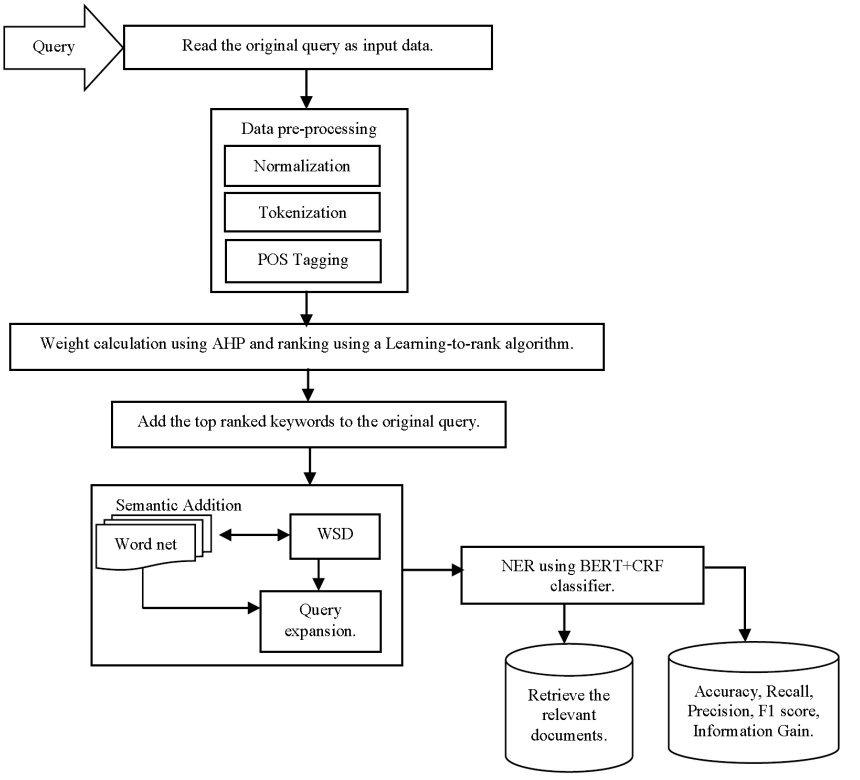

Figure 2

Proposed methodology flowchart

3.3 *Proposed Algorithm*

Start

1. **Preprocessing:**
   - Tokenize Q into individual terms: Q_{tokens}
   - Apply Part-of-Speech Tagging to Q_{tokens} **:** Q_{POS}
2. **Ranking and Weight Calculation:**
   - Calculate the relevance weight W of each term in Q_{POS} Using the AHP.
   - Assign relevance rank R to each term using a learning-to-rank algorithm based on their respective relevance weights.
3. **Keyword Selection:**
   - Select top-ranked keywords from Q_{POS} Based on their ranks.
4. **Semantic Addition:**
   - Apply WSD to Q_{POS} to resolve word meanings.
   - Perform QE on selected keywords to refine the query.
   - Utilize WordNet for additional semantic enrichment.
5. **Named Entity Recognition (NER):**
   - Apply the BERT+CRF classifier to Q_{POS} for NER as a subtask of IE.
6. **Information Retrieval:**
   - Retrieve relevant documents based on the refined query.

End procedure

4. Results and Discussions

4.1 *Dataset description*

To evaluate the effectiveness of the proposed strategy, the Yahoo dataset was selected. This dataset comprises 22,938 unlabeled documents, with a total of 236,205 tokens and marks extracted. After preprocessing, the significant tokens were reduced to less than 15,000, allowing for the potential exclusion of additional terms by raising the threshold, though this may cause overfitting due to mismatches in new queries. The second dataset includes over 93,000 news articles from LinkedIn, Google+, and Facebook, spanning eight months and covering four topics: the economy, Microsoft, Obama, and Palestine. The texts are more technical and varied

in length compared to the Yahoo dataset. As there are no known prior associations, the focus during the weighing stage is on these technical terms. The datasets were split, creating a training set of successful texts and a test set containing about 20% unsuccessful texts. This corpus includes 4,483,032 questions and corresponding answers, with a topic categorization dataset created using the ten most prominent categories, yielding large training samples of 140,000 and testing samples of 5,000 for each class [51, 52].

4.2 *Results*

Table 1, model architecture is related to the model's construction and it illustrates the hyperparameter values for the BERT+CRF model. The pre-trained model shows the pre-trained BERT variant used, the number of labels shows the number of output labels, the CRF layer suggests whether a conditional random field layer is included in BERT for token categorization, batch size is the number of sample examples used in one iteration, and number of epochs is the number of data points used.

Table 1

Hyperparameter table for the BERT + CRF model

Pretrained Model	**Bert-base-uncased**
Hyperparameter	Value
Model architecture	BERT
Number of Labels	3
CRF Layer	Yes (on top of BERT)
Batch Size	16
Number of Epochs	100
Learning Rate	$2e^{-5}$
Optimizer	AdamW
Loss Function	CrossEntropyLoss

Table 2 displays the results of applying the Yahoo dataset. BERT+CRF performed exceptionally well, with an accuracy of 99.7%, recall of 98%, precision of 99.8% and recall of 98%. These findings suggest that BERT+CRF achieved a high degree of accuracy in its classification of the data, with most of the predicted positives being true positives. Hence,

BERT+CRF proved to be a dependable and precise approach to data classification.

Table 2
Parameter Values of BERT+CRF on Yahoo Dataset

Parameters	Accuracy	Recall	Precision	F1 score
BERT+CRF on Yahoo Dataset	99.7%	98%	99.8%	98%

Table 3 presents a comparison of the different models' performance of four different analysed classification methods, namely SCC and M-KNN, bi-directional LSTM, SVM, and BERT+CRF. Devi et al. [27] introduced the SCC and M-KNN methods, which accomplished a recall of 98%, with similarly precision of 95%, and also achieved F1-Score of 97%. Safder et al. [31] proposed the bi-directional LSTM, which attained a recall of 82.9%, precision of 80.3%, and F1-Score of 81%. Khalifi et al. [34] introduced SVM and reached a recall of 96.2% and a good recognized precision of 99.2%. The BERT+CRF model proposed in this study demonstrated outstanding performance, with an accuracy of nearly 99.7%, recall of almost 98%, precision of 99.8%, and an F1-Score of almost 98%. These results suggest that BERT+CRF is the most effective method for classifying text samples in the Yahoo dataset, significantly surpassing other methods such as SCC and M-KNN, bi-directional LSTM, and SVM.

Table 3
Comparative Analysis

Parameters	Technique	Accuracy	Recall	Precision	F1 Score
Devi et al., (2020) [27]	SCC and M-KNN	-	98%	95%	97%
Safder et al., (2019) [31]	bi-directional LSTM	-	82.9%	80.3%	81%
Khalifi et al., (2019) [34]	SVM	-	96.2%	99.2%	-
Proposed Method	BERT+CRF	99.7%	98%	99.8%	98%

Figure 3 depicts the data comparison analysis graph. Devi et al. [27] introduced SCC and M-KNN with notable performance metrics. Safder et al. [31] proposed a Bi-directional LSTM, while Khalifi et al. [34] introduced SVM. However, the BERT+CRF model proposed in this study demonstrates

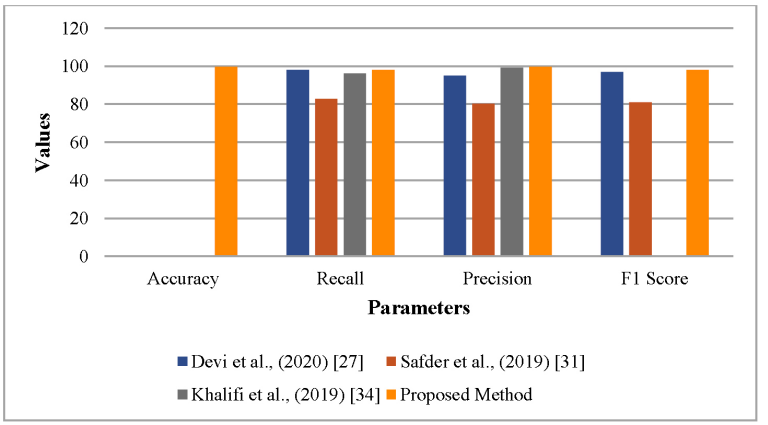

Figure 3

Comparison Analysis Graph

superior performance, with an accuracy of nearly 99.7%, recall of almost 98%, precision of 99.8%, and an F1-Score of nearly 98%. These findings indicate that BERT+CRF significantly outperforms other methods, including SCC and M-KNN, Bi-directional LSTM, and SVM, in classifying text samples within the Yahoo dataset.

5. Conclusion

Investigating cutting-edge QE methods and semantic analysis, the study aspires to improve search results and usher in a new era of information retrieval. In conclusion, the research highlights the importance of novel approaches in enhancing the performance and precision of data retrieval systems. The results demonstrate that the proposed method exhibits a high degree of accuracy (99.7%), along with a high rate of recall (98.0%), precision (99.8%), and F1-score (98%). The results show the promise of using sophisticated algorithms and semantic comprehension to hone search queries and return more pertinent and meaningful results for users. The future of information retrieval would be shaped in large part by the adoption of these cutting-edge methods, which would help make searches more streamlined and user-friendly for people in all kinds of fields. Search engines' future rests on their capacity to use sophisticated ML algorithms and NLP techniques to make large amounts of data from a variety of sources easily available.

References

[1] T. Guo, R. Zhou, and C. Tian, "On the information leakage in private information retrieval systems," *IEEE Trans. Inf. Forensics Security,* vol. 15, pp. 2999-3012 (2020).

[2] T. Barik and V. Singh, "AQtpUIR: Adaptive query term proximity based user information retrieval," *J. Inf. Optim. Sci.,* vol. 41, no. 6, pp. 1479–1497 (2020), doi: 10.1080/02522667.2020.1820190.

[3] V. Suma, "A novel information retrieval system for the distributed cloud using a hybrid deep fuzzy hashing algorithm," *J. Inf. Technol. Digit. World,* vol. 2, no. 3, pp. 151-160 (2020).

[4] L. Afuan, A. Ashari, and Y. Suyanto, "A study: query expansion methods in information retrieval," in *J. Phys.: Conf. Ser.,* vol. 1367, no. 1, p. 012001 (2019).

[5] V. Suma, "A novel information retrieval system for the distributed cloud using a hybrid deep fuzzy hashing algorithm," *J. Inf. Technol. Digit. World,* vol. 2, no. 3, pp. 151-160 (2020).

[6] S. Shaukat, A. Shaukat, K. Shahzad, and A. Daud, "Using TREC for developing semantic information retrieval benchmark for Urdu," *Inf. Process. Manag.,* vol. 59, no. 3, p. 102939 (2022).

[7] G. Pasi, G. J. F. Jones, L. Goeuriot, L. Kelly, S. Marrara, and C. Sanvitto, "Overview of the CLEF 2019 personalized information retrieval lab (PIR-CLEF 2019)," in Exp. IR Meets Multilinguality, Multimodality, Interaction: 10th Int. Conf. CLEF Assoc., Lugano, Switzerland, pp. 417-424 (2019).

[8] I. Soboroff, "Overview of TREC 2021," in 30th Text Retrieval Conf., Gaithersburg, Maryland (2021).

[9] L. Goeuriot, P. Ruch, J. J. McAuley, D. G. Oppenheim, M. Y. Tse, M. C. G. Scherp, J. Garcia, and B. P. D. Rojas, "Overview of the CLEF eHealth evaluation lab 2020," in Int. Conf. Cross-Language Eval. Forum Eur. Lang., Cham, Switzerland, pp. 255-271 (2020).

[10] H. K. Azad and A. Deepak, "Query expansion techniques for information retrieval: a survey," *Inf. Process. Manag.,* vol. 56, no. 5, pp. 1698-1735 (2019).

[11] V. Raskin and J. M. Taylor, "Fuzziness, uncertainty, vagueness, possibility, and probability in natural language," in 2014 IEEE Conf. Norbert Wiener 21st Century (21CW), pp. 1-6 (2014).

[12] L. Zhao and J. Callan, "Automatic term mismatch diagnosis for selective query expansion," in Proc. 35th Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, pp. 515-524 (2012).

[13] P. Gupta, K. Bali, R. E. Banchs, M. Choudhury, and P. Rosso, "Query expansion for mixed-script information retrieval," in Proc. 37th Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, pp. 677-686 (2014).

[14] M. Fernández, J. M. Albornoz, and M. O. M. Gutiérrez, "Semantically enhanced information retrieval: An ontology-based approach," *J. Web Semantics,* vol. 9, no. 4, pp. 434-452 (2011).

[15] M. W. Glier, D. A. McAdams, and J. S. Linsey, "Exploring automated text classification to improve keyword corpus search results for bio-inspired design," *J. Mech. Des.,* vol. 136, no. 11, p. 111103 (2014).

[16] S. Lim and C. S. Tucker, "A Bayesian sampling method for product feature extraction from large-scale textual data," *J. Mech. Des.,* vol. 138, no. 6, p. 061403 (2016).

[17] L. Lan, Y. Liu, and W. F. Lu, "Discovering a hierarchical design process model using text mining," in Int. Design Eng. Tech. Conf. Comput. Inf. Eng. Conf., vol. 50084 (2016), p. V01BT02A024.

[18] Y. Rezgui, S. Boddy, M. Wetherill, and G. Cooper, "Past, present and future of information and knowledge sharing in the construction industry: Towards semantic service-based e-construction?" *Comput.-Aided Des.,* vol. 43, no. 5, pp. 502-515 (2011).

[19] X. Chang, R. Rai, and J. Terpenny, "Development and utilization of ontologies in design for manufacturing," p. 021009 (2010).

[20] Y. Liu, S. C. J. Lim, and W. B. Lee, "Product family design through ontology-based faceted component analysis, selection, and optimization," *J. Mech. Des.*, vol. 135, no. 8, p. 081007 (2013).

[21] F. Shi, L. Chen, J. Han, and P. Childs, "A data-driven text mining and semantic network analysis for design information retrieval," *J. Mech. Des.,* vol. 139, no. 11, p. 111402 (2017).

[22] Y. Ding and S. Foo, "Ontology research and development. Part 1-a review of ontology generation," *J. Inf. Sci.,* vol. 28, no. 2, pp. 123-136 (2002).

[23] R. Jagerman, H. Zhuang, Z. Qin, X. Wang, and M. Bendersky, "Query expansion by prompting large language models," arXiv preprint arXiv:2305.03653 (2023).

[24] T. Russell-Rose, P. Gooch, and U. Kruschwitz, "Interactive query expansion for professional search applications," *Bus. Inf. Rev.*, vol. 38, no. 3, pp. 127-137 (2021).

[25] L. Gao, F. Qi, X. Qian, Z. Yang, and J. Zhang, "Complement lexical retrieval model with semantic residual embeddings," in Adv. Inf. Retrieval: 43rd Eur. Conf. IR Res., ECIR 2021, Virtual Event, Mar. 28–Apr. 1, 2021, Proc., Part I 43, Springer Int. Publishing, pp. 146-160 (2021).

[26] H. ALMarwi, M. Ghurab, and I. Al-Baltah, "A hybrid semantic query expansion approach for Arabic information retrieval," *J. Big Data,* vol. 7, no. 1, pp. 1-19 (2020).

[27] M. U. Devi and G. M. Gandhi, "Scalable information retrieval system in the semantic web by query expansion and ontological-based LSA ranking similarity measurement," *Int. J. Adv. Intell. Paradigms*, vol. 17, no. 1-2, pp. 44-66 (2020).

[28] H. Zamani, S. Dehghani, and W. B. Croft, "Generating clarifying questions for information retrieval," in Proc. Web Conf. 2020, pp. 418-428 (2020).

[29] Y. Qu, L. Chen, and X. Gao, "RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering," arXiv preprint arXiv:2010.08191 (2020).

[30] V. Karpukhin, L. MacAvaney, D. M. Radev, J. Miller, and R. Lewis, "Dense passage retrieval for open-domain question answering," arXiv preprint arXiv:2004.04906 (2020).

[31] I. Safder and S. Hassan, "Bibliometric-enhanced information retrieval: a novel deep feature engineering approach for algorithm searching from full-text publications," Scientometrics, vol. 119, pp. 257-277 (2019).

[32] J. Liu, P. Wang, and Y. Zhou, "Neural query expansion for code search," in Proc. 3rd ACM Singplan Int. Workshop Mach. Learn. Program. Lang., pp. 29-37 (2019).

[33] H. K. Azad and A. Deepak, "A new approach for query expansion using Wikipedia and WordNet," *Inf. Sci.,* vol. 492, pp. 147-163 (2019).

[34] H. Khalifi, H. Zare, and M. M. Sadjadi, "Query expansion based on clustering and personalized information retrieval," Prog. Artif. Intell., vol. 8, pp. 241-251 (2019).

[35] M. Morik, M. Kim, and P. M. M. Patel, "Controlling fairness and bias in dynamic learning-to-rank," in Proc. 43rd Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, pp. 429-438 (2020).

[36] Q. Song, J. Zhang, and R. M. Raj, "Stock portfolio selection using learning-to-rank algorithms with news sentiment," Neurocomputing, vol. 264, pp. 20-28 (2017).

[37] Z. Hu, Z. Wang, and D. M. Martinez, "Unbiased lambdamart: an unbiased pairwise learning-to-rank algorithm," in World Wide Web Conf., pp. 2830-2836 (2019).

[38] R. de FSM Russo and R. Camanho, "Criteria in AHP: A systematic review of literature," *Procedia Comput. Sci.,* vol. 55, pp. 1123-1132 (2015).

[39] A. Darko, M. Osei, and D. D. Nyarku, "Review of the application of analytic hierarchy process (AHP) in construction," *Int. J. Constr. Manag.,* vol. 19, no. 5, pp. 436-452 (2019).

[40] A. Pereira, M. J. Silva, and S. Pinto, "Querying semantic catalogues of biomedical databases," *J. Biomed. Inf.,* vol. 137, p. 104272 (2023).

[41] J.-P. Calbimonte, D. R. Mercier, and F. Ciravegna, "Decentralized semantic provision of personal health streams," *J. Web Semantics,* vol. 76, p. 100774 (2023).

[42] F. Zhao, Z. Zhu, and P. Han, "A novel model for semantic similarity measurement based on WordNet and word embedding," *J. Intell. Fuzzy Syst.,* vol. 40, no. 5, pp. 9831-9842 (2021).

[43] X. Pu, L. Yuan, J. Leng, T. Wu, and X. Gao, "Lexical knowledge enhanced text matching via distilled word sense disambiguation," Knowl.-Based Syst., vol. 263, p. 110282 (2023).

[44] M. Bevilacqua, T. Pasini, A. Raganato, and R. Navigli, "Recent trends in word sense disambiguation: A survey," in Proc. Int. Joint Conf. Artif. Intell., pp. 4330-4338 (2021).

[45] H. Husni, Y. Kustiyahningsih, F. H. Rachman, E. M. S. Rochman, and H. Yulian, "Query expansion using pseudo relevance feedback based on the Bahasa version of the Wikipedia dataset," in AIP Conf. Proc., vol. 2679, no. 1, AIP Publishing (2023).

[46] M. J. Hussain, H. Bai, S. H. Wasti, G. Huang, and Y. Jiang, "Evaluating semantic similarity and relatedness between concepts by combining taxonomic and non-taxonomic semantic features of WordNet and Wikipedia," Inf. Sci., vol. 625, pp. 673-699 (2023).

[47] G. A. Miller and C. Fellbaum, "WordNet then and now," Lang. Resour. Eval., vol. 41, pp. 209-214 (2007).

[48] G. Puccetti, V. Giordano, I. Spada, F. Chiarello, and G. Fantoni, "Technology identification from patent texts: A novel named entity recognition method," Technol. Forecast. Soc. Change, vol. 186, p. 122160, 2023.J. Mao and W. Liu, "Hadoken: a BERT-CRF Model for Medical Document Anonymization," in IberLEF@ SEPLN, pp. 720-726 (2019).

[49] M. Liu, Z. Tu, Z. Wang, and X. Xu, "LTP: A new active learning strategy for BERT-CRF based named entity recognition," arXiv preprint arXiv:2001.02524 (2020).

[50] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," Advances in Neural Information Processing Systems, vol. 28 (2015).

[51] N. Moniz and L. Torgo, "Multi-source social feedback of online news feeds," arXiv preprint arXiv:1801.07055 (2018).

[52] T. Barik and V. Singh, "AQtpUIR: Adaptive query term proximity based user information retrieval," *Journal of Information & Optimization Sciences*, vol. 41, no. 6, pp. 1479–1497 (2020), doi: 10.1080/02522667.2020.1820190.

[53] A. Pawar and V. Mago, "Calculating the similarity between words and sentences using a lexical database and corpus statistics," (2018). arXiv preprint arXiv:1802.05667.

Received April, 2024