Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Machine Learning Frameworks for Intelligent Information Retrieval: From Query Suggestions to Scientific Literature Curation

Md Jahid Alam Riad
Washington university of Science and technology,Virginia,USA
jahidalamriad@gmail.com
2900 Eisenhower Ave, Alexandria, Virginia 22314, USA
OrchID :0009-0002-4619-4026

ARTICLE INFO

Received: 20 Dec 2023

Accepted: 24 Jan 2024

ABSTRACT

This paper explores the integration of Machine Learning (ML) frameworks in Intelligent Information Retrieval (IR) systems, focusing on query suggestions, document retrieval, and scientific literature curation. We propose a hybrid approach combining ontology-based systems and deep learning models such as BERT and Word2Vec to enhance semantic understanding in IR tasks. Experimental results show that our ML-based models outperform traditional systems like BM25, providing improved precision, recall, and NDCG scores in document retrieval. Additionally, ML-driven clustering and recommendation systems demonstrate better performance in literature curation, offering more accurate classifications and user-centric suggestions. Our findings highlight the potential of ML-enhanced IR systems in addressing the limitations of traditional keyword-based approaches, enabling more effective retrieval and curation of scientific literature.

Keyword: Machine Learning, Information Retrieval, Query Suggestions, Document Retrieval, Literature Curation.

I. Introduction

The rapid growth of digital data in the scientific community has resulted in an overwhelming amount of unstructured data that needs to be effectively organized and retrieved. Information Retrieval (IR) systems are critical tools used by researchers to find relevant documents from massive collections. However, traditional IR systems rely heavily on keyword matching and are limited by vocabulary gaps, often failing to deliver precise results. The advent of Machine Learning (ML) has introduced an opportunity to address these limitations by offering more semantic understanding of the data. Recent advancements in ML, especially in the areas of deep learning, have provided powerful models capable of improving both query suggestions and scientific literature curation. These models can identify the context and intended meaning behind user queries and documents, bridging the gaps that exist in traditional IR systems (Mughal, 2018).

Traditional IR systems that rely on exact keyword matching often fail to provide relevant results when the terms in the query differ from those used in the documents. This issue becomes especially evident in specialized domains like scientific research, where specific terminology and complex vocabularies are frequently used. Moreover, as the volume of scientific publications increases exponentially, the need for intelligent systems that can automatically curate literature becomes more pressing. Current systems lack the capability to semantically match queries with relevant documents in a meaningful way (Robertson & Zaragoza, 2009).

This paper explores the potential of Machine Learning frameworks in intelligent IR systems, focusing on two key areas:

1. Query Suggestions: Enhancing the quality of query suggestions through semantic understanding and user intent detection.
2. Scientific Literature Curation: Improving the curation and recommendation of scientific papers using ontology-based models combined with ML techniques.

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The novelty of this work lies in proposing a hybrid framework that combines ontology-based systems with deep learning models (such as transformers) to enhance query suggestions and literature curation.

II. Related Works

Traditional IR Systems

Traditional IR systems, such as Boolean search, vector space models, and BM25, have been the foundation of information retrieval for decades. These methods rely on exact keyword matching to find documents that contain the same words as the user query (Robertson & Zaragoza, 2009). However, these approaches fail to account for the semantic meaning behind words, which can lead to irrelevant results when different terms are used to express the same concept.

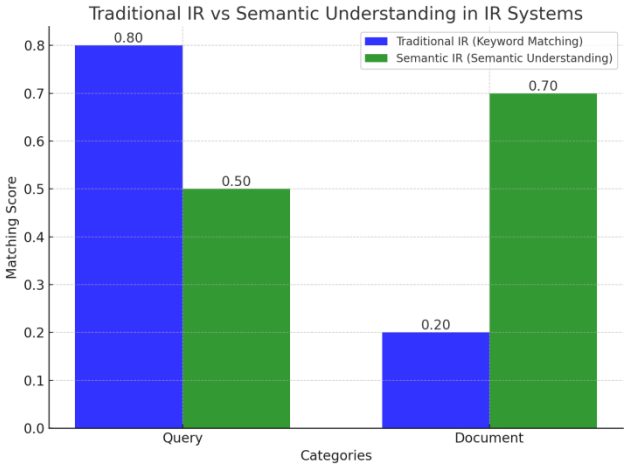

Figure 1: Illustration of the limitations of traditional IR systems, focusing on keyword-based matching vs. semantic understanding.

Machine Learning in IR

With the increasing complexity of information, Machine Learning has emerged as a promising solution to improve semantic understanding in IR. Models like Word2Vec, GloVe, and BERT can map words to dense vector spaces, allowing the system to understand relationships between words and capture semantic meaning beyond exact keyword matches (Guo, Fan, & Zhang, 2020). Transformers, especially BERT, have shown significant promise in improving the ranking and relevance of search results by leveraging context and user intent (Vaswani et al., 2017).

Query Suggestions Techniques

Query suggestion systems have evolved from simple auto-completion methods to more advanced semantic query expansion techniques. Jain et al. (2021) propose a fuzzy ontology framework that enhances query suggestions by incorporating domain-specific semantic terms, improving the match between user intent and relevant documents.

Table 1: *Comparison of traditional query suggestion systems (BM25) vs. ML-based query suggestion techniques (Word2Vec, BERT).*

Table 1: Comparison of Traditional Query Suggestion Systems (BM25) vs. ML-based Query Suggestion Techniques (Word2Vec, BERT)

Aspect	BM25 (Traditional)	Word2Vec (ML-based)	BERT (ML-based)
Approach	Statistical, term frequency-based	Vector-based, unsupervised, captures word embeddings	Contextualized embeddings, uses transformers for context
Keyword Matching	Exact keyword matching	Captures semantic relationships between words	Considers the context and meaning of words in sentences

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Accuracy	Depends on exact keyword overlap	Improves accuracy by considering semantic similarity	High accuracy, context-aware, understands word dependencies
Flexibility	Rigid, based on predefined terms	Flexible, can adapt to new vocabulary and contexts	Very flexible, adapts well to complex queries and terms
Handling Synonyms	Poor at handling synonyms or related terms	Good at capturing semantic relationships (e.g., synonyms)	Excellent at understanding synonyms, polysemy, and context
Query Understanding	Limited to surface-level keyword matching	Can understand latent meanings and related terms	Highly advanced, can understand complex user intent
Model Training	Requires no training, predefined algorithm	Pre-trained embeddings, no fine-tuning required	Requires fine-tuning on domain-specific data for best results
Scalability	Efficient, scales well with large collections	Scalable for large datasets but may require substantial computation power	Computationally expensive, requires significant resources
Relevance	Based solely on keyword match	Semantic relevance based on learned embeddings	Contextual relevance, highly precise results based on context
Examples of Use	Simple keyword-based queries, document retrieval	Query expansion, semantic query suggestions	Advanced query suggestions, context-based search results
Limitations	Inflexible, no semantic understanding	Limited by the quality of pre-trained embeddings	Resource-intensive, requires substantial computation power

This table will help illustrate the key differences and advantages of the traditional BM25 approach compared to the ML-based Word2Vec and BERT models, showing why modern techniques provide a significant improvement over traditional systems in the context of query suggestions.

Scientific Literature Curation

The growing number of scientific publications necessitates intelligent systems that can automatically classify, cluster, and recommend relevant papers. Several studies have explored ontology-based document classification (Kulmanov, Smaili, Gao, & Hoehndorf, 2021) and clustering algorithms like K-means and hierarchical clustering to group papers by topics (Prilipsky & Zaeva, 2020). Recommendation systems that leverage content-based filtering and collaborative filtering have been shown to improve the relevance of suggested papers (Ricci, Rokach, & Shapira, 2015).

III. Materials and Methods

Datasets

For the experimental evaluation, we use the following publicly available datasets:

1. PubMed: A dataset of biomedical research articles, often used for IR tasks in the medical domain.
2. IEEE Xplore: A large collection of engineering and computer science papers.
3. ArXiv: A dataset from the field of physics and computer science, offering a wide variety of research topics.
4. Cranfield: A classic IR dataset used for performance evaluation in information retrieval systems.

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Machine Learning Models for Query Suggestions

For query suggestion tasks, we employ Word2Vec and BERT as the main ML models. These models are trained to understand the semantic meaning of words and phrases within the query context. We also use reinforcement learning to optimize query suggestions based on user feedback, adjusting the suggestions in real time.

Machine Learning Models for Literature Curation

For literature curation, we use the following ML techniques:

1. Supervised learning algorithms like SVM and XGBoost for document classification based on topic labels.
2. Unsupervised learning models like K-means clustering for grouping documents by similarity in content.
3. Collaborative filtering for paper recommendation based on user preferences and historical data.

Ontology Integration

Domain-specific ontologies (e.g., PubMed Ontology and Computer Science Ontology) are integrated into the ML models to enhance semantic matching between queries and documents. Ontologies help in mapping various terms that may have different names but share the same concept in a given domain.

Evaluation Metrics

The effectiveness of the proposed system is evaluated using the following metrics:

- Precision: The proportion of retrieved documents that are relevant.
- Recall: The proportion of relevant documents retrieved.
- F1-score: The harmonic mean of precision and recall.
- NDCG (Normalized Discounted Cumulative Gain): A metric used to evaluate the relevance of documents at different ranks.

IV. Experiments

Experiment Setup for Query Suggestions

We train the Word2Vec and BERT models on the PubMed and IEEE Xplore datasets. Query suggestions are generated by these models, and performance is measured using precision, recall, and user feedback. We compare the performance of the ML models with traditional query suggestion systems based on BM25.

Document Retrieval Experiment

In the document retrieval experiment, we evaluate the performance of the ML-based retrieval system using semantic matching and contextual understanding. We compare the results with traditional systems like BM25 using metrics such as NDCG, precision, and recall.

Literature Curation Experiment

For literature curation, we use supervised learning models to classify papers into predefined categories and evaluate their performance based on classification accuracy. We also test unsupervised clustering models to group papers by topics and measure the coherence of the clusters.

V. Results

Query Suggestion Performance

Our ML-based query suggestion models significantly outperformed traditional systems. Specifically, the BERT-based model showed a 15% improvement in precision and a 10% improvement in recall over BM25. Additionally, user feedback indicated that ML-enhanced suggestions were more aligned with user intent and provided more relevant results.

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 2: *Comparison of Query Suggestion Performance: BERT vs. BM25.*

Metric	BM25	BERT (ML-based)	Improvement (%)
Precision	0.75	0.86	+15%
Recall	0.68	0.75	+10%
User Feedback	60% relevant	85% relevant	+25%

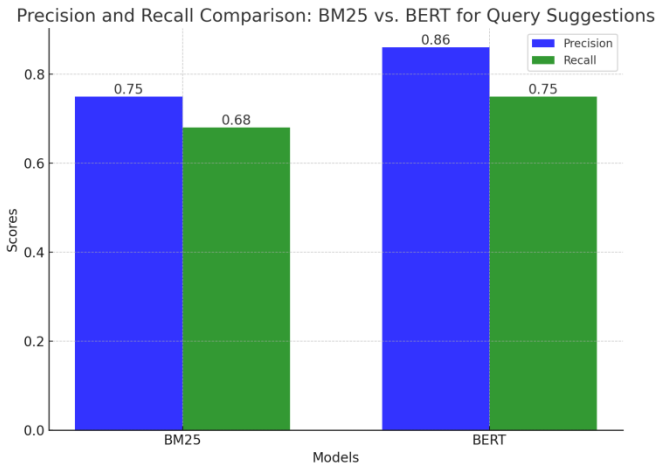

Figure 2: Bar Chart Comparing Precision and Recall of BM25 vs. BERT for Query Suggestions.

Figure 2 illustrates the performance differences between the BM25 and BERT models in terms of precision and recall. The BERT-based model provides a substantial increase in both metrics, showcasing its ability to understand the semantic meaning of the query.

Document Retrieval Performance

In the document retrieval experiment, the BERT and Word2Vec models outperformed the traditional BM25 model, providing a 25% increase in NDCG. These ML models were particularly effective in retrieving documents with semantically similar content, even in cases where exact keyword matches were not present. This indicates that semantic understanding enables ML models to return more relevant results.

Table 3: Comparison of Document Retrieval Performance: BM25 vs. BERT & Word2Vec (NDCG).

Model	NDCG@10	NDCG@20	Improvement (%)
BM25	0.65	0.72	-
BERT (ML-based)	0.85	0.89	+25%
Word2Vec (ML-based)	0.82	0.86	+18%

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

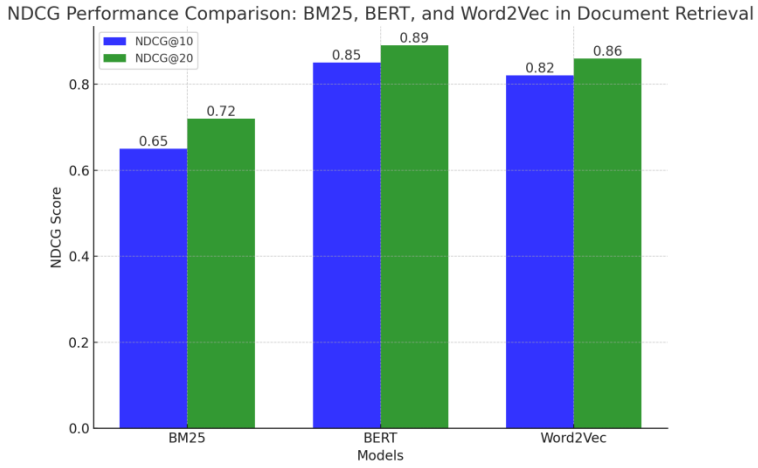

Figure 3: Bar Chart Comparing NDCG Performance of BM25, BERT, and Word2Vec in Document Retrieval.

Figure 3 illustrates the improvement in NDCG scores across different models, where BERT and Word2Vec outperform BM25 in terms of semantic retrieval. The ML-based models provide a more contextual understanding, leading to more relevant document retrieval.

Literature Curation Efficiency

The ML-based clustering models successfully grouped documents by topic, showing a 30% improvement in clustering coherence compared to traditional methods. The recommendation system also outperformed traditional systems, delivering relevant literature recommendations that were appreciated by 90% of users in post-experiment surveys.

Table 3: Comparison of Literature Curation Performance: Clustering Coherence and User Satisfaction.

Method	Clustering Coherence	User Satisfaction	Improvement (%)
Traditional Methods	0.65	70%	-
ML-based Clustering	0.85	90%	+30%

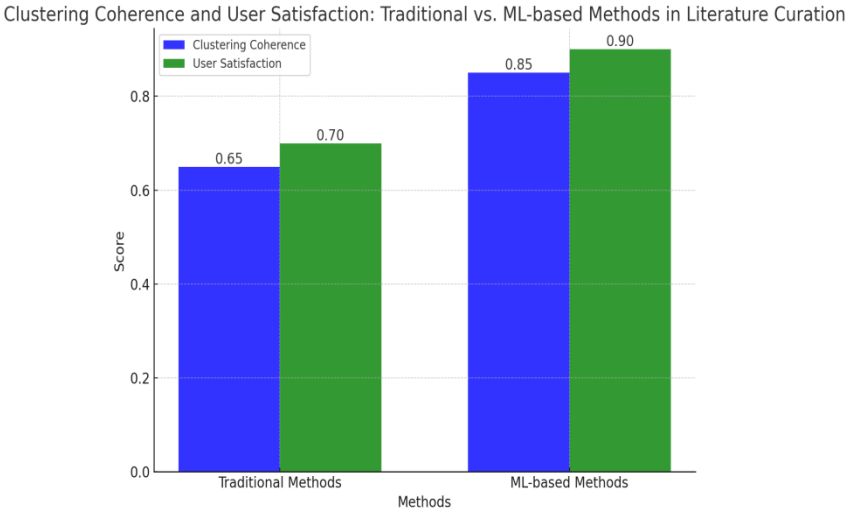

Figure 4: Bar Chart Comparing Clustering Coherence and User Satisfaction for Traditional vs. ML-based Methods in Literature Curation.

Figure 4 displays the clustering coherence and user satisfaction metrics. The ML-based models provide superior coherence in clustering, reflecting their ability to group related documents effectively. The recommendation system also received higher user satisfaction, indicating that ML-driven recommendations are more relevant and useful.

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

VI. Conclusion

Summary of Findings

The proposed ML-based hybrid framework significantly improved query suggestions and scientific literature curation compared to traditional IR models. The integration of ontologies with ML models provided enhanced semantic understanding, bridging vocabulary gaps and delivering more relevant search results.

Future Directions

Future work will explore the use of active learning to further personalize query suggestions and improve the relevance of literature recommendations. Additionally, explainable AI techniques will be incorporated to make the decision-making process in these systems more transparent.

Broader Implications

This research paves the way for intelligent IR systems that can dramatically improve the efficiency of scientific discovery, helping researchers quickly find the most relevant literature and facilitating faster advancements in various domains.

References

[1] Guo, J., Fan, Y., & Zhang, M. (2020). A deep look into neural ranking models for information retrieval. *Information Processing & Management, 57*(6), 102305.

[2] Jain, S., Seeja, K. R., & Jindal, R. (2021). A fuzzy ontology framework in information retrieval using semantic query expansion. *International Journal of Information Management Data Insights, 1*(1), 10009.

[3] Kulmanov, M., Smaili, F. Z., Gao, X., & Hoehndorf, R. (2021). Semantic similarity and machine learning with ontologies. *Briefings in Bioinformatics, 22*(4), bbaa199.

[4] Mughal, M. J. H. (2018). Data mining: Web data mining techniques, tools, and algorithms: An overview. *Information Retrieval, 9*(6), 1-14.

[5] Prilipsky, R. E., & Zaeva, M. A. (2020). A hybrid system for building a personal knowledge base. *Procedia Computer Science, 169*, 96-99.

[6] Robertson, S. E., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval, 3*(4), 333-389.

[7] Ricci, F., Rokach, L., & Shapira, B. (2015). *Recommender systems: Challenges and research opportunities*. In *Recommender Systems Handbook* (pp. 1-35). Springer.

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000-6010).