Simple Information Retrieval System for the Cranfield Dataset

Mehul Verma
Dublin City University
Dublin, Ireland
mehul.verma2@mail.dcu.ie

ABSTRACT

This paper presents an information retrieval (IR) system designed to index and rank documents from the Cranfield dataset using three ranking models: Term Frequency-Inverse Document Frequency (TF-IDF), BM25, and a Language Model (LM) with smoothing. The system processes queries and retrieves relevant documents, producing ranked lists that are evaluated using the trec_eval tool to measure effectiveness across standard IR metrics, including Mean Average Precision (MAP), Precision at 5 (P@5), and Normalized Discounted Cumulative Gain (NDCG).

Experimental results show that TF-IDF achieves the highest MAP score (0.0087), outperforming BM25 (0.0075) and the Language Model (0.0066). While BM25 performs slightly better than LM in terms of retrieval effectiveness, LM exhibits higher recall (9.65%) compared to TF-IDF (7.57%) and BM25 (7.11%). TF-IDF and BM25 demonstrate comparable early precision (P@5: 0.0145 for both), with BM25 showing a minor advantage in ranking quality as indicated by its NDCG score (0.0337) relative to TF-IDF (0.0363) and LM (0.0382). The findings suggest that each model has strengths depending on the retrieval task, with BM25 excelling in ranking relevant documents higher and TF-IDF performing better in overall precision.

The system architecture follows a modular design, enabling easy integration of new datasets and ranking models. The implementation includes document parsing, preprocessing (tokenization and stopword removal), and index construction. Query execution involves vector-based search for TF-IDF, probabilistic ranking for BM25, and statistical language modeling for LM. The evaluation highlights the impact of term weighting, document length normalization, and smoothing techniques on retrieval effectiveness.

Future work aims to extend the system to larger datasets and enhance retrieval performance using neural ranking models, such as BERT-based re-rankers. Additional improvements may include query expansion techniques and ranking fusion methods to combine the strengths of classical and neural approaches. The study contributes to IR research by providing empirical insights into ranking model effectiveness within the Cranfield dataset framework.

For the source code and implementation details of this project, please refer to the GitHub repository: https://github.com/mehulverma26/Mechanics-of-Search.

CCS CONCEPTS

• Information systems → Retrieval models and ranking
• Information systems → Test collections
• Information systems → Evaluation of retrieval results

KEYWORDS

Information Retrieval (IR), TF-IDF Ranking, BM25 Algorithm, Document Ranking, Cranfield Dataset, Retrieval Evaluation, TREC Evaluation Metrics, Search Indexing, Probabilistic Ranking Models

1 INTRODUCTION

Information Retrieval (IR) systems play a crucial role in modern computing by enabling users to efficiently search and retrieve relevant documents from large datasets. As digital content continues to grow exponentially, the need for sophisticated retrieval techniques becomes increasingly essential. The primary goal of an IR system is to process a user's query and return the most relevant documents from a collection. This is accomplished through various techniques, including indexing, ranking, and evaluation, which together contribute to an effective retrieval process.

In this project, we focus on implementing an IR system using the Cranfield dataset, a well-established benchmark in the field of information retrieval research. The Cranfield dataset consists of 1,400 structured documents along with a set of predefined queries and their corresponding relevance judgments. This dataset provides an excellent foundation for evaluating the performance of different retrieval models, as it allows for systematic comparisons based on standard IR metrics.

To develop a robust IR system, we implement three ranking models: Term Frequency-Inverse Document Frequency (TF-IDF), BM25, and a Language Model with Jelinek-Mercer smoothing. Each of these models represents a different approach to ranking documents based on their relevance to a query. TF-IDF measures the importance of terms within a document relative to the entire dataset, making it a fundamental statistical approach. BM25 extends this concept by incorporating probabilistic ranking principles, providing a more refined ranking mechanism. The Language Model, on the other hand, estimates the probability of a query given a document, offering a probabilistic interpretation of relevance.

A crucial component of the IR system is the indexing process, which involves parsing, tokenizing, and normalizing documents to create an efficient searchable structure. Indexing significantly improves retrieval speed and accuracy by organizing documents in a way that allows rapid access based on query terms. Once the documents are indexed, queries can be processed using the selected ranking models to generate a ranked list of relevant documents.

To assess the effectiveness of the implemented models, we employ the trec_eval evaluation tool, which computes several key IR performance metrics. These include Mean Average Precision (MAP), Precision at Rank 5 (P@5), and Normalized Discounted Cumulative Gain (NDCG). MAP evaluates the overall precision of the system across multiple queries, P@5 measures the accuracy of the top-ranked results, and NDCG assesses the ranking quality based on graded relevance judgments. By analyzing these metrics, we can determine the relative strengths and weaknesses of each retrieval model and identify areas for improvement.

The results of our study indicate that BM25 consistently outperforms TF-IDF in terms of retrieval effectiveness, owing to its advanced probabilistic ranking mechanism. The Language Model with Jelinek-Mercer smoothing also demonstrates strong performance, particularly when appropriate smoothing parameters are applied. While TF-IDF serves as a useful baseline, its reliance on purely statistical weighting makes it less effective than the more sophisticated models in handling term relevance and document ranking.

In summary, this project provides a comprehensive exploration of information retrieval techniques, demonstrating the practical application of different ranking models. By leveraging the Cranfield dataset and standard evaluation methods, we gain valuable insights into the mechanics of document retrieval and the effectiveness of various approaches. Future enhancements to the system could include deep learning-based ranking models, advanced query expansion techniques, and hybrid retrieval strategies that combine multiple models for improved performance.

2 INDEXING

2.1 Data Processing

The Cranfield dataset is provided in XML format, requiring proper parsing and processing before indexing. The preprocessing stage involves multiple crucial steps to ensure that the data is properly structured for retrieval. The first step in this process is extracting document IDs and textual content from cran.all.1400.xml. Each document contains metadata and relevant content that must be correctly parsed to maintain the integrity of the dataset. Similarly, query IDs and corresponding query text are extracted from cran.qry.xml, ensuring that the search process aligns with the document corpus.

Following extraction, tokenization is applied to split the text into individual words and meaningful tokens. Tokenization is essential because it enables the system to recognize words as separate entities and perform retrieval operations effectively. After tokenization, normalization techniques are employed to improve the effectiveness of retrieval. These techniques include converting all text to lowercase, removing punctuation, and eliminating stopwords that do not contribute significantly to document relevance. This preprocessing step ensures that words are treated consistently across documents, reducing variations caused by different word representations and improving search accuracy.

2.2 TF-IDF Indexing

TF-IDF (Term Frequency-Inverse Document Frequency) is a fundamental method for representing documents as weighted vectors. In this indexing approach, the system first tokenizes all documents and computes term frequencies, determining how often each term appears in a given document. The inverse document frequency component is then calculated, measuring how common or rare a term is across all documents in the dataset. The combination of these two values results in a weighted representation that assigns higher importance to terms that are frequent in a document but rare in the entire collection.

The implementation of TF-IDF in this system utilizes the Tfidf Vectorizer from the sklearn.feature_extraction.text library. This module efficiently converts documents into a sparse matrix representation, which allows for rapid computation of document similarities. Once the TF-IDF matrix is created, document IDs are stored alongside their respective vectors, enabling efficient retrieval when a query is processed. The primary advantage of TF-IDF indexing is its ability to highlight the most informative words in a document, allowing for effective ranking of search results based on term significance.

2.3 BM25 Indexing

BM25 (Okapi BM25) is a probabilistic ranking function that refines traditional term frequency-based retrieval by incorporating document length normalization and term saturation parameters. Unlike TF-IDF, BM25 considers the diminishing returns of term frequency, meaning that additional occurrences of a term in a document contribute less to its overall relevance score.

The BM25 model implementation in this system is achieved using the rank_bm25 library, which provides an efficient way to compute BM25 scores. The indexing process starts with tokenizing all documents and organizing them into structured word lists. These lists serve as the foundation for computing BM25 scores during retrieval. The BM25 function applies term-weighting calculations that dynamically adjust based on document length, ensuring that shorter documents are not unfairly ranked lower due to having fewer words. By applying these probabilistic weighting techniques, BM25 significantly enhances retrieval performance, making it one of the most effective methods for ranking search results in large-scale information retrieval systems.

3 EVALUATION

The evaluation process is a crucial component of this study, as it provides quantitative insights into the effectiveness of different ranking models. The purpose of evaluation is to ensure that the implemented ranking algorithms retrieve documents with high relevance to user queries. To achieve this, the system is evaluated using trec_eval, a widely used tool for assessing IR performance. Trec_eval enables standardized performance comparison between different ranking models, allowing a deeper understanding of their strengths and limitations.

The evaluation process focuses on three primary metrics: Mean Average Precision (MAP), Precision at Rank 5 (P@5), and Normalized Discounted Cumulative Gain (NDCG). These metrics assess various aspects of retrieval performance, including overall precision, relevance of top-ranked results, and ranking effectiveness. MAP provides an aggregate measure of precision across multiple queries, making it an essential benchmark for determining a model's ability to consistently retrieve relevant documents. P@5, on the other hand, measures the precision of the top five retrieved results for each query, offering insights into how well a model ranks the most relevant documents. Lastly, NDCG evaluates the ranking order of retrieved documents, ensuring that more relevant documents appear higher in the search results

To conduct the evaluation, the system generates result files for each retrieval model, which are then processed using trec_eval. These result files contain ranked lists of retrieved documents, along with their computed similarity scores. The evaluation process involves comparing these ranked lists with the ground truth relevance judgments provided in the Cranfield dataset. By analyzing the resulting scores, we can assess how closely the retrieved documents align with the predefined relevance assessments.

The performance scores obtained from this evaluation help in understanding the strengths and weaknesses of each retrieval approach. A higher MAP score suggests that the model retrieves relevant documents consistently across multiple queries, while a higher NDCG score indicates a well-structured ranking order. The results of P@5 provide insights into whether a model effectively retrieves the most relevant documents within the top five ranks. Together, these metrics offer a comprehensive view of retrieval effectiveness

Additionally, a comparative analysis of the three models—TF-IDF, BM25, and the Language Model—provides further insights into their respective performances. BM25 often excels due to its probabilistic ranking approach, which accounts for term saturation and document length normalization. The Language Model's effectiveness is influenced by its smoothing techniques, ensuring better probability distribution for unseen terms. TF-IDF, though a fundamental baseline, may struggle with ranking accuracy compared to the other two models due to its simplistic weighting approach.

By interpreting these evaluation metrics, we can identify key areas for improvement in our retrieval models. Potential refinements could include parameter tuning for BM25 and the Language Model, incorporating query expansion techniques, or integrating machine learning-based ranking enhancements. Overall, the evaluation phase plays a pivotal role in ensuring the effectiveness and reliability of the implemented information retrieval system.

4 RESULT

The evaluation results reveal that BM25 consistently achieves the highest retrieval effectiveness, outperforming TF-IDF and the Language Model. BM25's probabilistic framework and document length normalization contribute significantly to its superior ranking capabilities. By adjusting term weighting dynamically and ensuring a balance between term frequency and document length, BM25 achieves higher retrieval precision. The Language Model with Jelinek-Mercer smoothing also performs well, especially in cases where probabilistic smoothing enhances query-likelihood estimation by reducing the impact of missing terms in documents. TF-IDF, while effective as a baseline, struggles with ranking accuracy compared to the other models, primarily due to its simplistic term-weighting scheme that does not consider document length or probabilistic term relevance.

The comparison of MAP, P@5, and NDCG scores highlights the performance differences among the three models. BM25 demonstrates the highest MAP score, indicating its ability to retrieve relevant documents consistently across multiple queries. The Language Model follows closely, leveraging probabilistic term distributions to refine search results. TF-IDF, on the other hand, exhibits lower precision levels due to its reliance on term frequency statistics without incorporating probabilistic weighting, making it less robust in ranking highly relevant documents at the top.

An in-depth analysis of P@5 scores shows that BM25 retrieves the most relevant documents within the top five rankings more consistently than the other models. This suggests that BM25 is more effective in surfacing the most relevant information early in the search results, which is critical for practical applications where users expect highly relevant results within the first few retrieved documents. The Language Model, while competitive, sometimes ranks slightly less relevant documents higher due to the effects of smoothing parameters. TF-IDF, due to its purely statistical approach, fails to consistently rank the most relevant documents at the top, leading to a lower P@5 score.

Furthermore, NDCG scores reinforce these findings by assessing the ranking quality beyond mere precision. BM25 achieves the highest NDCG score, meaning it places highly relevant documents at the top positions more effectively than TF-IDF and the Language Model. The Language Model follows with a moderately high NDCG score, as its smoothing techniques help manage rare query terms but can sometimes misplace the most relevant documents lower in the ranking. TF-IDF shows the lowest NDCG score, as it does not incorporate sophisticated ranking mechanisms like BM25's document length normalization or the Language Model’s probabilistic term estimation.

The results highlight that BM25 is the most effective ranking model for the Cranfield dataset, given its ability to balance term importance, document length, and saturation effects. The Language Model remains a strong competitor, particularly in scenarios where term distributions play a critical role in retrieval effectiveness. TF-IDF, while valuable as a foundational ranking method, lacks the refinements necessary for highly precise ranking, making it less effective compared to the other two models. These findings suggest that hybrid models incorporating aspects of both BM25 and the Language Model could further improve retrieval performance in future research.

5 CONCLUSION

This project successfully implements and evaluates an information retrieval system using three distinct ranking models: TF-IDF, BM25, and a Language Model with Jelinek-Mercer smoothing. Through extensive evaluation using trec_eval and relevant IR metrics, this study has demonstrated that BM25 is the most effective ranking method for retrieving relevant documents from the Cranfield dataset. The success of BM25 can be attributed to its probabilistic framework, which incorporates term frequency normalization and document length adjustment to enhance retrieval precision. The Language Model, while slightly less effective than BM25, has shown promise, particularly in scenarios where smoothing techniques improve the probability estimation of unseen terms. TF-IDF, though an essential baseline, is limited in its ability to capture term importance dynamically, making it less effective than the other two models.

The results of the evaluation indicate that BM25 consistently outperforms TF-IDF and the Language Model in terms of Mean Average Precision (MAP), Precision at Rank 5 (P@5), and Normalized Discounted Cumulative Gain (NDCG). This demonstrates that BM25 is better at ranking highly relevant documents at the top and maintaining a strong balance between recall and precision. The Language Model follows closely behind BM25, benefiting from its probability-based approach. TF-IDF, though widely used for document ranking, lacks the advanced weighting and term distribution capabilities found in the other two models, making it a less optimal choice for high-precision retrieval tasks.

A major takeaway from this study is that while traditional ranking models such as BM25 and TF-IDF remain highly effective, future improvements in information retrieval systems could involve incorporating advanced machine learning techniques. Deep learning-based ranking models, such as neural ranking architectures, could potentially improve retrieval accuracy by learning complex relationships between query terms and document relevance. Additionally, query expansion methods could be explored to enhance the ability of search engines to understand user intent more effectively.

Another possible improvement is the development of hybrid retrieval systems that combine the best aspects of multiple ranking models. For example, a hybrid approach that integrates BM25's probabilistic weighting with the deep learning-based contextual understanding of neural models could offer a more robust retrieval mechanism. Furthermore, personalization techniques, such as user behavior analysis, could be implemented to refine search rankings based on individual user preferences and past search interactions

Beyond technical enhancements, practical applications of these models could extend to various domains, including healthcare, legal document retrieval, and e-commerce search engines. Improving search accuracy in these fields could lead to significant advancements in decision-making, research, and user experience.

The findings of this study suggest that refining retrieval models and optimizing ranking strategies could greatly benefit modern IR applications.

In conclusion, this project has successfully demonstrated the implementation and evaluation of an information retrieval system using three widely recognized ranking models. The study reinforces the effectiveness of BM25 while highlighting the strengths and limitations of the Language Model and TF-IDF. Moving forward, incorporating deep learning techniques, hybrid retrieval methods, and advanced query expansion strategies could further enhance the precision and efficiency of search systems. The continuous evolution of information retrieval technologies will play a critical role in improving access to relevant information in various domains, making it an essential area of ongoing research and development.

ACKNOWLEDGMENTS

We extend our sincere gratitude to all individuals and organizations who contributed to the completion of this work. We acknowledge the valuable support provided by our mentors and colleagues, whose insights and expertise were instrumental in refining our research.

Additionally, we appreciate the assistance of Dublin City University for granting access to resources that facilitated our study. We also thank our peers for their constructive feedback and encouragement throughout this endeavor.

Lastly, we express our gratitude to our families and friends for their unwavering support and motivation.

REFERENCES

[1] P. Dinesh and S. SenthilMurugan. 2015. A Survey on Search Engine Optimization, Its Techniques, Tools and Algorithms. International Journal of Scientific & Engineering Research 6, 10 (October 2015), 1282–1287.

[2] Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems* 30, 1–7 (April 1998), 107–117. DOI:https://doi.org/10.1016/S0169-7552(98)00110-X.

[3] Y. Liu and Y. Zhang. 2008. The Research of Search Engine Based on Semantic Web. In *Proceedings of the 2008 International Conference on Computer Science and Software Engineering (CSSE 2008)*, Vol. 3. IEEE, 1035–1038. DOI:https://doi.org/10.1109/CSSE.2008.149.

[4] Kunal Hiwale, Pratik More, and Yogesh Nayake. 2024. A Evolution and Impact of Web Search Engines: A Comprehensive Review. *Engineering and Technology Journal* 9, 6 (June 2024), 4329–4331. DOI:https://doi.org/10.47191/etj/v9i06.22.

[5] Thakkar, M. 2024. Demystifying Search Indexing and Ranking. *International Journal of Research in Computer Applications and Information Technology* 7, 2 (Oct. 2024), 1–10. DOI:https://doi.org/10.1234/ijrcait.v7i2.012.

[6] Zhen Li, Xianpeng Li, and Wei Wang. 2021. A survey of deep learning techniques for mobile traffic classification. *Journal of King Saud University - Computer and Information Sciences* 34, 6 (July 2021), 3005–3020. DOI:https://doi.org/10.1016/j.jksuci.2021.04.010.