CA6005 Assignment 1 – Information Retrieval systems.

Daniel Verdejo – 23102562

Table of Contents

Abstract

In information retrieval (IR), the effectiveness of different retrieval models is important for enhancing the accuracy and relevance of retrieved information based off a user query. This report explores three prominent IR models: BM25, Vector Space Model, and the Language Model BERT. Each model will use the Cranfield dataset and undergo evaluation through the TREC (Text REtrieval Conference) Evaluation, providing a standardized platform for assessment. Finally, the paper will make conclusions along with outlining any future research endeavours recommendations.

Introduction

IR plays a crucial role in the organization, storage, retrieval, and presentation of information, with the aim to produce relevant document references in response to user queries. Relevance is central to this process, influenced by various document and user factors. While an IR system cannot precisely identify all relevant documents due to complexity, it focuses on ranking documents by estimated similarity to the query [1].

Work by researchers Salton laid the foundation for key concepts such as term weighting, document similarity, and relevance ranking. The first round of Cranfield experiments was conducted between 1958 and 1966 with the goal of evaluating existing search systems and to establish a methodology for designing and evaluating new systems [3]. In Cranfield I, four indexing methods were compared using a 'source document' approach, which revealed a high failure rate across these methods. Cranfield II then enhanced the methodology by separating recall and precision evaluations and investigating various indexing languages. TREC and other evaluations still utilize the Cranfield dataset showing its robustness and relevance almost 70 years after its inception.

This paper documents findings for an experiment implementing and evaluating the performance of 3 different IR systems. Prior to testing, standard IR practices are carried out by preprocessing the data through stopword removal and indexing techniques. Once processed, we proceed to construct the IR models for each system: VSM, BM25, and BERT Language Model. The classical approach is employed for VSM and BM25, while BERT is adapted from its text classification setting into an IR context.

The performance evaluation uses the TREC_eval suite of metrics specifically chosen for its relevance to IR system performance analysis. The metrics include the Number of Relevant Retrieved

Documents (Num_rel_ret), Mean Average Precision (MAP), Reciprocal Rank (Rprec), and precision at various cutoff points (P@n), such as P@5, P@10, and so on. Each of these metrics offers a different perspective on the effectiveness and efficiency of the IR systems under study, capturing aspects such as the ability to retrieve relevant documents, the average precision across queries, the rank of the first relevant document, and precision within the top results of the search output. Each system will be subject to evaluation and analysis of their results will be carried out. Finally, some comparative analysis between the 3 will be carried out and conclusions drawn.

Indexing

Indexing is a fundamental concept in IR, it serves as a mechanism to organize and allow for efficient access to vast amounts of data. Indexing involves creating and maintaining structured representations of information, enabling quick retrieval based on predefined criteria. In python, for example, a simple dictionary can be used as a means of indexing, this is often preferred rather than lists as the access time for values is constant rather than linear. Two popular indexing methods are signature files and inverted indices. Signature files involve generating a signature for each document using hashing on its words, facilitating faster searching in a separate file. Inverted indices on the other hand, represent each document with a list of keywords, with the index file containing pointers to qualifying documents [2].

In this experiment an inverted indices is used by first pre-processing documents using an ingestion function, which verifies that each document contains the necessary fields ('title', 'author', 'text') and combines them into a single "full_text" attribute. This consolidated text is cleaned, converting all text to lowercase for uniformity, removing non-alphanumeric characters to discard punctuation and special symbols that are generally irrelevant to the document's semantic content, and eliminating stopwords. Stopwords are common words that appear frequently across texts and carry minimal individual significance for retrieval (e.g. "the", "is", etc.), are removed to reduce the index size and focus on meaningful words for retrieval. This preprocessing is aligned with standard practices in information retrieval to enhance search efficiency and relevance by emphasizing significant query terms and minimizing noise in the dataset [9].

An inverted index is then constructed, where for every unique word present in the "full_text" attribute, a list of documents containing that word is stored as a key: value pair. This approach of indexing, enables quick retrieval of documents containing specific query terms by directly accessing lists of documents associated with those terms. The inverted index will significantly enhance the speed and efficiency of search operations, ensuring that the system scales well with the size of the document corpus [8].

Ranking

This research uses the score which will be generated by each IR system. The scores generated by BM25, VSM, and BERT indicate the relevance of documents to a given query term, with higher scores meaning greater relevance. This ranking mechanism serves as a means of prioritizing documents that are more likely to serve the user's information need. By sorting documents based on these scores, a search engine can present results in a descending order of relevance, drastically improving the search experience by ensuring users encounter the most pertinent information first. The differing methodologies of these systems from statistical calculations in BM25 and VSM to deep linguistic understanding in BERT offer varied approaches to discerning document relevance, thereby highlighting the rich landscape of techniques available for information retrieval tasks.

Evaluation

The performance evaluation uses the TREC_eval suite of metrics specifically chosen for its relevance to IR system performance analysis. The metrics include the Number of Relevant Retrieved Documents (Num_rel_ret), Mean Average Precision (MAP), Reciprocal Rank (Rprec), and precision at various cutoff points (P@n), such as P@5, P@10, and so on. Each of these metrics offers a different perspective on the effectiveness and efficiency of the IR systems in the experiment, capturing aspects such as the ability to retrieve relevant documents, the average precision across queries, the rank of the first relevant document, and precision within the top results of the search output. Notably, the analysis zeros in on the critical aspects of retrieval relevance and precision which are paramount for assessing the practical utility of IR systems. Upcoming sections will explain the IR systems, findings, and their broader significance.

Vector Space Model

The Vector Space Model (VSM), proposed by Salton and colleagues, represents documents and queries as vectors in a high-dimensional space, with each term assigned a separate dimension. The model aims to rank documents based on the similarity between the query and each document. In VSM, the similarity between documents and queries is computed using the cosine function. This model incorporates term weighting schemes which considers the term frequency (TF) and inverse document frequency (IDF), commonly known as TF-IDF weighting, to measure the importance of terms in documents or queries [1][2]. Salton theorised that the optimal indexing space for document retrieval or pattern matching environments is characterized by maximizing the distance between stored entities or documents, leading to improved retrieval performance inversely correlated with space density. This insight guides the selection of an optimal indexing vocabulary for document collections, with empirical evaluations demonstrating the effectiveness of this model [3].
Salton demonstrated that by employing index vectors of two documents D_i and D_j, it's possible to compute a similarity coefficient $s(D_i, D_j)$ denoting the extent of similarity in their respective terms and term weights [3]. This similarity metric might inversely correlate with the angle between the corresponding vector pairs; specifically, when the term assignment for both vectors is identical, the angle is zero, leading to a maximal similarity measure [3]. Consequently, if two documents share similar index terms, they would be visually represented by points in a vector space that are situated closely together [3]. This concept and VSM laid the groundwork for modern information retrieval by treating documents and queries as vectors in a multidimensional space. This conceptualization, as illustrated by Raghavan and Wong [1], enabled the computation of similarity through vector angles, fundamentally altering how relevance was quantified.

Analysing the metrics gathered during this research the following was observed:

Metric	VSM
num_ret (number of retrieved docs)	102788
Num_rel (number of relative docs)	1074
Num_rel_ret (number of relative retrieved docs)	496
Map (mean average precision)	0.0085
Rprec (reciprocal relevance rank)	0.0058
P (precision)	P@5: 0.0066 P@10: 0.0039 P@15: 0.0031 P@20: 0.0036

	P@30: 0.0033 P@100: 0.0049 P@200: 0.0048 P@500: 0.0043 P@1000: 0.0033

1. **Num_rel_ret:** Out of the 1,074 relevant documents, VSM managed to retrieve 496. While this represents a fraction of the total relevance potential, it offers a starting point to evaluate how effectively VSM retrieves relevant documents against the total number of relevant documents.
2. **MAP**: At 0.0085, the MAP score quantifies the model’s precision across all queries, on average. Given that MAP accounts for the order of retrieval, a score of 0.0085 suggests VSM places relevant documents relatively lower in the ranked list on average across queries. High MAP values are desired as they indicate higher precision at the top of the ranked list of returned documents.
3. **Rprec**: With a value of 0.0058, R-Precision reflects VSM's capability to retrieve relevant documents within the top R documents, where R is the number of relevant documents for a query. The closeness of this score to zero suggests that, on average, the top portions of VSM's results contain a low proportion of the total relevant documents for a query.
4. **P@n:** These metrics give us insights into the precision of the VSM at different cutoffs - the proportion of relevant documents within the top K retrieved documents. The precision values are quite low at earlier cutoffs (e.g., P@5 = 0.0066), indicating that relevant documents are not being ranked highly by the model. This is crucial in practical applications where users are less likely to look beyond the top few results. The scores mildly improve but remain low as K increases, which further confirms the model's challenges in effectively ranking the most relevant documents at the top of its results.

Summary

The Vector Space Model, in this context, shows challenges with retrieving and ranking relevant documents effectively for the queries involved. Both the MAP and Rprec indicate lower performance in relevance and ranking, with Precision metrics highlighting difficulty especially in the immediate top results but showing slight improvement further down the list.

BM25 Model

The BM25 retrieval method, also known as Okapi BM25 is the work of Robertson, Walker, Hancock-Beaulieu, Gull, and Lau [4], is a retrieval system designed for simplicity, robustness, and user-friendliness [4]. It operates on the principle that effective and efficient retrieval can be achieved without complex Boolean logic or extensive manual intervention. Instead, it employs best-match searching techniques, focusing on relevance feedback to enhance query results. Relevance feedback allows users to refine their initial queries based on the retrieved documents, resembling a browsing-like exploration of topics rather than precise specifications [4]. This has made BM25 a staple in many search engines and information retrieval systems [5].

The Okapi project served as the foundation for system development before its involvement in TREC-related activities. Subsequent modifications were implemented to meet the demands of TREC, including enhancements to support interactive retrieval, showcasing its adaptability. During the evaluation of the feedback run in TREC-1, a "frozen rank" approach was used, maintaining top-ranking documents before feedback to simulate real-world search scenarios. While this gave an improvement to precision, the overall system performance, utilizing basic Okapi methods, was considered respectable but not exceptional [4].

Analysing the metrics gathered during evaluation of BM25 the following was observed:

Metric	BM25
num_ret (number of retrieved docs)	102788
Num_rel (number of relative docs)	1074
Num_rel_ret (number of relative retrieved docs)	553
Map (mean average precision)	0.0113
Rprec (reciprocal relevance rank)	0.0059
P (precision)	P@5: 0.0145 P@10: 0.0099 P@15: 0.0101 P@20: 0.0102 P@30: 0.0099 P@100: 0.0073 P@200: 0.0066 P@500: 0.0053 P@1000: 0.0036

Num_rel_ret: This reflects the number of relevant documents BM25 successfully retrieved. Although this constitutes just over half of the available relevant documents (553 / 1074), it's indicative of a reasonable recall given the systems more focused retrieval approach.

MAP: BM25 boasts the highest MAP among the systems discussed at 0.0113, indicating it has relatively high precision across the ranking lists for all queries. This metric highlights BM25's effectiveness at ranking relevant documents higher, on average, across the diverse set of queries.
Rprec: While not significantly high, the score 0.0059 suggests that BM25 does manage to place relevant documents higher in its search results, though there is room for improvement in pushing the most relevant documents to the top ranks.
P@n: The various precision scores at different cutoffs demonstrate BM25 starting strong with P@5 at 0.0145, indicating decent precision in the very top results, but sees a gradual decrease as n increases. Notably, the precision doesn't plummet drastically, maintaining a semi-consistent descent through P@1000 = 0.0036, evidencing BM25's ability to retain some degree of relevancy deep into its retrieval list.

The BM25 model demonstrates a balanced performance in recall and precision across various metrics, with a particular strength in maintaining higher average precision across query rankings (MAP). It exhibits a commendable ability to retrieve a significant chunk of relevant documents (Num_rel_ret) out of the total available (Num_rel), and its precision at early retrieval stages (P@5, P@10) suggests it effectively prioritizes highly relevant documents. Given BM25's algorithmic simplicity and robustness, these metrics underscore its utility in search and information retrieval systems, where blending relevancy with retrieval efficiency remains a priority. BM25's methodology, focusing on term frequency and document length, provides a practical approach that achieves a high degree of relevance across a broad spectrum of information retrieval tasks.

Language Model (BERT)

BERT (Bidirectional Encoder Representations from Transformers)

The introduction of BERT (Bidirectional Encoder Representations from Transformers) in 2018 by researchers at Google revolutionized the field of natural language processing (NLP) by enabling models to understand the context of words in a sentence like never before. This deep learning model employs the transformer architecture to process words in relation to all the other words in a sentence, rather than one-by-one in order [10]. As a language model that deeply understands context and semantics, BERT transcends traditional word-matching techniques, as surveyed by Wang, Huang, Tu, Huang, Laskar, and Bhuiyan [11]. This method represents a paradigm shift towards leveraging neural networks to grasp the nuanced meanings of queries and documents, setting a new benchmark in the accuracy of retrieval results. BERT is significantly more complex than the other 2 systems being researched in this paper, it is a transformer model which are most popularised by GPT, which requires significant computational resources for training and inference compared to others. This could make it less feasible for scenarios where limited resources are available or real-time performance is necessary. Regardless of this it is one of the latest models and is worth seeing how older systems hold up to modern systems.

Analysing the metrics gathered for BERT during this research the following was observed:

Metric	BERT
num_ret (number of retrieved docs)	204744
Num_rel (number of relative docs)	1074
Num_rel_ret (number of relative retrieved docs)	1003
Map (mean average precision)	0.0091
Rprec (reciprocal relevance rank)	0.0061
P (precision)	P@5: 0.0039 P@10: 0.0059 P@15: 0.0053 P@20: 0.0046 P@30: 0.0042 P@100: 0.0042 P@200: 0.0043 P@500: 0.0048 P@1000: 0.0048

1. **num_ret:** BERT retrieved 204,744 documents which is just under double of that which the others retrieved. BERT's higher retrieval count suggests a wider search net, potentially increasing the chance to retrieve more relevant documents but also raising the possibility of introducing more noise (irrelevant documents).
2. **Num_rel_ret:** 1,003 of the 1074 relevant documents (93%) have been retrieved, indicating a high recall suggesting BERT's effectiveness at identifying relevant documents across the dataset.
3. **MAP:** Although not very high at 0.0091, this score suggests that BERT can retrieve relevant documents with a reasonable level of precision.
4. **Rprec:** A score of 0.0061 indicates moderate success in ranking highly relevant documents at the top of the result set, suggesting an area for improvement in precision at higher ranks.
5. **P@n:** The lower precision scores (e.g., P@5 = 0.0039) at early ranks indicate that relevant documents are not consistently positioned at the very top of the search results. However, as the number of considered documents increases (up to P@1000 = 0.0048), the precision slightly improves but remains moderate, though it is unlikely that a user will search through 1000 results to find a relevant document. This highlights a potential challenge in accurately ranking the most relevant documents within the topmost results.

BERT demonstrates a broad capacity to retrieve many relevant documents, suggesting a strong ability to recall pertinent information. However, its precision, particularly in the top ranks of search results, indicates moderate performance. While BERT excels in understanding complex language structure and context, translating this comprehension into precise rankings at the top of search results is a nuanced challenge that could benefit from further optimization or hybrid approaches that incorporate BERT's contextual understanding with traditional IR systems’ strengths in precision-focused retrieval tasks.

Conclusions

Based on the results of the research, for the BM25, VSM, and BERT systems, several observations and comparisons can be made:

Metric	BM25	VSM	BERT
num_ret (number of retrieved docs)	102788	102788	204744
Num_rel (number of relative docs)	1074	1074	1074
Num_rel_ret (number of relative retrieved docs)	553	496	1003
Map (mean average precision)	0.0113	0.0085	0.0091
Rprec (reciprocal relevance rank)	0.0059	0.0058	0.0061
P (precision)	P@5: 0.0145 P@10: 0.0099 P@15: 0.0101 P@20: 0.0102 P@30: 0.0099 P@100: 0.0073 P@200: 0.0066 P@500: 0.0053 P@1000: 0.0036	P@5: 0.0066 P@10: 0.0039 P@15: 0.0031 P@20: 0.0036 P@30: 0.0033 P@100: 0.0049 P@200: 0.0048 P@500: 0.0043 P@1000: 0.0033	P@5: 0.0039 P@10: 0.0059 P@15: 0.0053 P@20: 0.0046 P@30: 0.0042 P@100: 0.0042 P@200: 0.0043 P@500: 0.0048 P@1000: 0.0048

1. **Num_ret**: BM25 and VSM both retrieved 102,788 documents, while BERT retrieved significantly more at 204,744 documents. BERT's higher retrieval count could indicate a broader search but may also increase the potential for retrieving irrelevant documents.
2. **Num_rel_ret:** BERT retrieved the most relevant documents 1,003, nearly retrieving the number of documents retrieved by VSM 496 and BM25 553 combined. This indicates BERT's strong capability in identifying relevant documents in the dataset. There may be a correlation between the **Num_ret** and **num_rel_ret** values here which could be explanatory as to why both the VSM and BM25 systems have retrieved around half that of BERT.
3. **Map**: BM25 achieved the highest MAP at 0.0113, this indicates it has the highest precision across all queries on average, followed by BERT 0.0091 and VSM 0.0085. A higher MAP suggests that BM25 more consistently ranks relevant documents higher across queries.
4. **Rprec**: The R-Precision is very similar across the three methods, with BM25 and BERT slightly outperforming VSM. This metric, being quite close for all three, suggests a relatively comparable ability to retrieve relevant documents up to the Rth rank, where R is the total number of relevant documents. Rprec is sensitive to the position of the first relevant result in the ranked list, it may not be a definitive indicator of overall performance [6].
5. **P@n**: BM25 showed the strongest result at between P@5 0.0145 to P@30 0.0099 outperforming the precision scores of both BERT and VSM. All three systems have similar precision values for P@100 onwards. However, the VSM has a slightly lower P@30 value

compared to others. This indicates that BM25 and BERT may provide better results when focusing on the top 30 results.

In conclusion, based on the results of the TREC evaluation on the 3 systems, all 3 systems have their strengths and weaknesses. BM25 provides the best precision overall and provides better ranking results when focusing on the MAP value. This research shows marginal difference in performance between the 3 systems, meaning that the choice of which to use may come down to an external factor such as computational resources, dataset size or domain knowledge.

Bibliography

[1] V. V. Raghavan and S. K. M. Wong, "A Critical Analysis of Vector Space Model for Information Retrieval", Journal of the American Society for Information Science, vol. 37, no. 5, pp. 279-287, 1986.

[2] A. Roshdi and A. Roohparvar, "Review: Information Retrieval Techniques and Applications", International Journal of Computer Networks and Communications Security, vol. 3, no. 9, pp. 373-377, 2015.

[3] G. Salton, "INFORMATION STORAGE AND RETRIEVAL", Department of Computer Science, Cornell University, Ithaca, New York, Scientific Report No. ISR-22, 1974.

[4] D. K. Harman, Ed., "The First Text REtrieval Conference (TREC-1)", NIST Special Publication 500-207, Computer Systems Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899, 1993.

[5] M. Géry and C. Largeron, "BM25t: a BM25 extension for focused information retrieval", Knowledge and information systems, vol 32, pp. 217–241, Springer-Verlag London Limited, 2011.

[6] S. Bütcher, C. L. A. Clarke, and G. V. Cormack, "Information Retrieval: Implementing and Evaluating Search Engines,", book, The MIT Press, 2010.

[7] S. Robertson, H. Zaragoza, and M. Taylor, "Simple BM25 extension to multiple weighted fields," in Proceedings of the thirteenth ACM international conference on Information and knowledge management (CIKM '04), pp. 42-49, 2004.

[8] J. Zobel and A. Moffat, "Inverted files for text search engines," ACM Computing Surveys, vol. 38, no. 2, pp. 6-es, 2006.

[9] C. D. Manning, P. Raghavan, & H. Schütze, "Introduction to information retrieval," Cambridge University Press, 2008.

[10] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805 [cs.CL], 2018.

[11] J. Wang, J. X. Huang, X. Tu, J. Wang, A. J. Huang, M. T. R. Laskar, and A. Bhuiyan, "Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges," ACM Computing Surveys, accepted January 2024.