UTILIZING BM25 AND COSINE SIMILARITY FOR STRUCTURED DOCUMENT RANKING IN INFORMATION RETRIEVAL

UNIVERSITY OF HAVANA

by

Franco Hernández Piloto, Carlos Mauricio Reyes Escudero, Hivan Cañizares Díaz

College of Computing and Data Science

2024

Abstract

In the past it has been common to compute scores for the individual fields (e.g.title and content) independently and then combine these scores (typically linearly) to get a final score for the document. It is applied here the way Stephen Robertson adapt the BM25 ranking formula for structured documents in the paper ”Simple BM25 Extension to Multiple Weighted Fields”, and the cosine similarity score function was applied with this modification for ranking structured documents.In this scheme, a structured document with a title weight of two is mapped to an unstructured document with the title content repeated twice. This more verbose unstructured document is then ranked in a custom way using cosine similarity and the normal BM25.

Chapter 1

Introduction

1.1 Background

1.1.1 BM25 Algorithm

The BM25 (Best Matching 25) algorithm is a widely used probabilistic model for information retrieval that ranks documents based on their relevance to a given query. It builds upon the traditional term frequency-inverse document frequency (TF-IDF) model by incorporating additional factors to improve ranking effectiveness.

Term Frequency (TF)

Term frequency measures how often a term appears in a document. It is calculated as follows:

$$TF(t,d) = \frac{f(t,d)}{|d|} \tag{1.1}$$

Where:

- $f(t,d)$ is the raw count of the term t in document d.
- $|d|$ is the total number of terms in the document.

The way to think about this is that the more times the query term(s) occur a document,

the higher its score will be. This makes intuitive sense: a document that has our name in it lots of time is more likely to be related to us than a document that has it only once.

In BM25, term frequency is adjusted to account for diminishing returns; that is, the contribution of additional occurrences of a term to the overall relevance score decreases as the term frequency increases.

Inverse Document Frequency (IDF)

Inverse document frequency measures the importance of a term across the entire document collection.The IDF component measures how often a term occurs in all of the documents and "penalizes" terms that are common. The actual formula Lucene/BM25 uses for this part is:

$$IDF(t) = \log\left(\frac{N - n + 0.5}{n + 0.5} + 1\right) \tag{1.2}$$

Where:

- N is the total number of documents in the collection.
- n is the number of documents containing the term t.

IDF helps to reduce the weight of common terms that appear in many documents, thus emphasizing more informative terms.This makes intuitive sense: the term "of" is likely to occur in nearly every English document, so when a user searches for something like "the car," "car" is probably more important than the term "of" (which will be in nearly all documents).Note the +1 , this is to avoid negative values due to the log function.

BM25 Formula

The BM25 score for a document d with respect to a query q is calculated using the following formula:

$$\text{BM25}(d, q) = \sum_{t \in q} IDF(t) \cdot \frac{TF(t, d) \cdot (k1 + 1)}{TF(t, d) + k1 \cdot (1 - b + b \cdot \frac{|d|}{\text{avgdl}})} \tag{1.3}$$

Where:

- $k1$ and b are parameters that control the influence of term frequency and document length normalization, respectively.
- $|d|$ is the length of the document d, and avgdl is the average document length across all documents.

The term frequency (TF) component of the TF-IDF increases linearly with the number of times a keyword appears in a document. This means that the increase in score when moving from 10 to 20 occurrences of a keyword is the same as the increase when moving from 100 to 110 occurrences. In other words, the scoring system treats each additional occurrence of a term uniformly, regardless of the baseline frequency. Instead of just relying on the number of occurrences of a keyword, BM25 introduces a new parameter:

$$\frac{TF(t,d)}{TF(t,d)+k1} \tag{1.4}$$

The parameter k1 above acts as the parameter to control the contribution of each incremental occurrence of a keyword into that part of the BM25 score.Which affects directly the total score. Makes that The first few occurrences of a keyword have a significant impact. However, as the keyword appears more frequently in the document, each additional occurrence contributes less. This is due to the diminishing returns mechanism in the BM25 formula.

So basically the parameter k1 controls the rate at which the contribution of each keyword occurrence grows. A higher k1 value leads to a slower increase in the contribution of each additional occurrence to the TF score. This helps address the issue of keyword saturation in the overall score.

Another key advantage of BM25 over traditional TF-IDF is that BM25 incorporates document length normalization. In BM25, a short document with few but highly relevant keywords is considered more important than a long document with many keywords but less relevance per term. This helps to prevent long documents from

dominating the search results solely based on term frequency.

$$\frac{TF(t,d)}{TF(t,d) + k1 \cdot (\frac{|d|}{\text{avgdl}})} \tag{1.5}$$

If the document is shorter than average, the value of $\frac{TF(t,d)}{TF(t,d)+k1}$ will increase, and vice versa. In other words, shorter documents will approach the saturation point quicker than longer documents.

In certain corpora, the length of documents can significantly influence results, while in others, it may not have much impact at all. To address this variability, an additional parameter b is introduced in the equation to regulate the role of document length in the overall scoring.

$$\frac{\text{TF}(t,d)}{\text{TF}(t,d) + k1 \cdot (1 - b + b \cdot \frac{|d|}{\text{avgdl}})} \tag{1.6}$$

When the parameter b is set to 0, the document length normalization term in the BM25 formula simplifies to 1, effectively eliminating the impact of document length on the scoring. In this scenario, the BM25 score is computed solely based on term frequency and inverse document frequency, disregarding document length entirely. Conversely, when b is set to 1, the overall score becomes significantly influenced by the ratio of the current document length to the average document length.

Documents that are shorter than the average will yield a smaller value in the denominator, resulting in a higher score, while longer documents will produce a larger denominator value, leading to a lower score. By varying the value of b between 0 and 1, one can adjust the significance of document length in the overall BM25 scoring. A value closer to 0 diminishes the emphasis on document length, whereas a value nearer to 1 increases the weight given to the document length normalization factor

The BM25 algorithm effectively balances the importance of term frequency and document length, allowing for more relevant document ranking in response to user queries.

1.1.2 Cosine Similarity

Cosine similarity is a metric used to measure the similarity between two non-zero vectors in an inner product space. It is particularly useful in the context of information retrieval for comparing the similarity of documents or queries represented as vectors. The cosine similarity between two vectors A and B is calculated using the following formula:

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||} \tag{1.7}$$

Where:

- $A \cdot B$ is the dot product of vectors A and B.
- $||A||$ and $||B||$ are the magnitudes (or norms) of vectors A and B, respectively, calculated as:

$$||A|| = \sqrt{\sum_{i=1}^{n} A_i^2} \tag{1.8}$$

The cosine similarity score ranges from -1 to 1 but it was clipped to be from 0 to 1:

- A score of 1 indicates that the vectors are identical.
- A score of 0 indicates that the vectors are orthogonal (no similarity).

1.1.3 Advantages of BM25 over TF-IDF

While TF-IDF is a foundational model for information retrieval, BM25 offers several advantages:

- **Diminishing Returns**: BM25 incorporates a mechanism for diminishing returns on term frequency, meaning that the relevance of additional occurrences of a term decreases. This leads to more balanced scoring.
- **Document Length Normalization**: BM25 includes a normalization factor that adjusts for document length, ensuring that longer documents do not receive an unfair advantage simply due to their size.

- **Parameter Tuning**: The parameters $k1$ and b in BM25 allow for fine-tuning of the model to better fit specific datasets and retrieval tasks, enhancing its adaptability compared to the static nature of TF-IDF.

These enhancements make BM25 a more effective and robust choice for document ranking in information retrieval systems compared to traditional TF-IDF ranking alone.

Chapter 2

Methodology

2.1 Methodology

2.1.1 Data Collection

The study utilized the Cranfield dataset, which contains 1,400 short documents. Each document was structured with several fields, including:

- **Title**: The main heading of the document.
- **Author**: The individual(s) responsible for the content.
- **Bibliography**: References cited within the document.
- **Content**: The main body text, which provides the primary information.

The documents were structured to allow for easy extraction and processing of textual data.

2.1.2 Preprocessing

Before applying the BM25 algorithm and cosine similarity, the documents underwent a preprocessing phase:

- **Tokenization**: Each field of the document was split into individual words or tokens using the `word_tokenize` function from the Natural Language Toolkit

(NLTK) framework. This process involved converting the text into lowercase and removing any unnecessary punctuation to ensure consistency in term matching.

- **Lemmatization**: To ensure that different forms of a word were treated as the same term, we employed the `WordNetLemmatizer` from the NLTK framework. For example, the terms *investigations* and *investigation* were normalized to their base form, *investigation*. This step is crucial for improving the accuracy of term matching and relevance scoring.
- **Field Weighting**: Different fields were assigned weights to reflect their importance in the context of the document. For instance, the title was given a higher weight compared to the content, acknowledging that terms in the title are often more indicative of the document's relevance. The specific field weights used were as follows:
  - `title`: 3.0 (Highest weight)
  - `author`: 2.0 (Medium weight)
  - `bibliography`: 2.0 (Medium weight)
  - `content`: 1.0 (Normal weight)

2.1.3 Implementation of BM25

The BM25 algorithm was implemented to rank the documents based on their relevance to a given query. The following steps were involved:

- **Document Frequency Calculation**: The frequency of each term across all documents was calculated. This step is crucial for determining the inverse document frequency (IDF) later.
- **Term Frequency Calculation**: For each document, the frequency of each term was computed, adjusted by the field weights to reflect their significance. The process was as follows:
  - Each document was processed field by field (title, author, bibliography, content).

- For each term in a field, its frequency was incremented by the weight of that field:
  - Title terms: frequency += 3.0
  - Author terms: frequency += 2.0
  - Bibliography terms: frequency += 2.0
  - Content terms: frequency += 1.0
- This weighting scheme effectively treated terms in important fields (like the title) as if they appeared multiple times, increasing their significance in the document representation.

- **IDF Calculation**: The IDF values for each term were computed using the formula:

$$IDF(t) = \log\left(\frac{N - n + 0.5}{n + 0.5} + 1\right) \tag{2.1}$$

where N is the total number of documents and n is the number of documents containing the term t.

- **BM25 Score Calculation**: For each document, the BM25 score was computed based on the term frequencies and IDF values using the formula:

$$\text{BM25}(d, q) = \sum_{t \in q} IDF(t) \cdot \frac{TF(t, d) \cdot (k1 + 1)}{TF(t, d) + k1 \cdot (1 - b + b \cdot \frac{|d|}{\text{avgdl}})} \tag{2.2}$$

where $k1$ and b are parameters that control the influence of term frequency and document length normalization. K1 and b were set to 1 after some poor fine-tuning,but this parameters depends completely about your dataset.

2.1.4 Cosine Similarity Calculation

In addition to BM25, cosine similarity was calculated to measure the similarity between the query and the documents:

- **TF-IDF Vector Creation**: A TF-IDF vector was generated for the query based on the same term frequency and IDF calculations used for the documents.

- **Cosine Similarity Calculation**: The cosine similarity between the TF-IDF vectors of the documents and the query vector was computed using the formula:

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||} \tag{2.3}$$

  This measure provided an additional layer of relevance assessment, focusing on the angle between the document and query vectors. Notably, the calculation of cosine similarity was enhanced by utilizing the TF-IDF values derived from the BM25 implementation, leveraging already computed term frequencies and inverse document frequencies.

2.1.5 Combining Scores

The final step involved combining the BM25 and cosine similarity scores to generate a comprehensive ranking of the documents:

- **Score Thresholds**: Documents were categorized based on predefined thresholds for both BM25 and cosine similarity scores:
  - BM25 thresholds: High (0.60), Medium (0.40), Low (0.20)
  - Cosine similarity thresholds: High (0.30), Medium (0.20), Low (0.10)
- **Document Categorization**: Based on these thresholds, documents were sorted into six categories: high BM25, high cosine, medium BM25, medium cosine, low BM25, and low cosine. Documents below all thresholds were categorized as ”bad”.
- **Interleaving Process**: The final ranking was created by interleaving documents from these categories in the following order:
  1. High BM25 and high cosine documents were interleaved first.
  2. If fewer than 30 documents were selected, medium BM25 and medium cosine documents were interleaved next.
  3. If still below 30 documents, low BM25 and low cosine documents were

interleaved.

4. If the total was still below 30, ”bad” documents were added to complete the list.

- **Final Ranking**: This interleaving process ensured that documents with high scores in either BM25 or cosine similarity were prioritized, while still maintaining a balance between the two scoring methods. The result was a ranked list of up to 30 documents, combining the strengths of both the probabilistic (BM25) and vector space (cosine similarity) models.

2.1.6 Evaluation

The effectiveness of the combined scoring mechanism was evaluated using a set of queries. The results were manually analyzed to determine the accuracy and relevance of the retrieved documents, providing insights into the performance of the BM25 and cosine similarity methods in structured document retrieval.

The algorithm was originally crafted for corporate collections that include between 5 and 30 field types, incorporating various section headings, list items, annotation fields, and authoring and editing information. Regrettably, there are no publicly available collections that match this description for evaluation purposes. Consequently, the evaluation was restricted to standard collections with a limited number of field types, primarily consisting of Title, Body, and, in the case of web collections, Anchor text. The results obtained, despite the limited number of fields, highlight the potential challenges associated with score combination.

The proposed approach involves combining the BM25 score with the cosine similarity score, integrating two distinct models—one probabilistic and the other vector-based—within structured documents. The modification of the BM25 score in the context of cosine similarity, as outlined in Stephen Robertson’s work, demonstrated promising empirical results. This combined approach is expected to outperform cosine similarity alone, as the BM25 component significantly enhances the ability to identify relevant documents that cosine similarity may overlook.

Chapter 3

Conclusion

In this study, the researchers implemented the BM25 algorithm and cosine similarity to evaluate the relevance of structured documents based on user queries. The findings revealed several important insights regarding the performance of these two methods:

1. **Effectiveness of BM25**: The BM25 scoring mechanism demonstrated considerable success in ranking documents. The results indicated that documents receiving high BM25 scores were consistently relevant to the query, showcasing the algorithm's ability to effectively prioritize documents based on term frequency and inverse document frequency.

2. **Cosine Similarity Performance**: In contrast, the cosine similarity scores were notably lower across the board. While cosine similarity is a valuable metric for measuring the angle between document vectors, it struggled to provide nuanced rankings compared to BM25. This discrepancy suggests that cosine similarity may not be as effective in differentiating between documents of varying relevance. Notably, a cosine similarity score of 0.3 is considered good on the average of observed cases, but many documents with high BM25 scores did not achieve similarly high cosine similarity scores.

3. **Merging BM25 and Cosine Similarity**: To improve the results, the researchers merged the BM25 scores with the best results from cosine similarity. The BM25 score was prioritized, but it was supported by the cosine similarity scores. This

approach helped to reaffirm that a document is truly valuable if it scores highly in both BM25 and cosine similarity. Additionally, documents with medium BM25 scores but very low cosine similarity scores were still considered usable, as the BM25 score was the primary factor in the ranking.

4. **Calculation Method for Cosine Similarity**: It is important to note that the calculation of cosine similarity in this study was not the usual approach. The researchers utilized the TF-IDF calculations derived from the BM25 implementation, which enhanced performance by leveraging already computed term frequencies and inverse document frequencies.

5. **Discrepancies in Document Ranking**: Many documents that achieved medium scores with the BM25 ranking exhibited very low scores with cosine similarity. This finding raises questions about the reliability of cosine similarity as a standalone metric for document relevance. While it can accurately rank high-quality documents, it tends to poorly rank lower-quality documents, leading to a lack of balance in the results.

6. **Manual Evaluation**: A manual evaluation of the results confirmed these observations in most cases. This qualitative assessment, often referred to as *human evaluation*, indicated that while cosine similarity could effectively identify high-quality documents, it lacked precision in ranking less relevant documents. In contrast, BM25 provided a more balanced approach, ensuring that documents across the spectrum of quality were ranked more equitably.

7. **Field Weighting Impact**: The implementation of field weighting in the BM25 algorithm played a crucial role in enhancing the relevance of the results by adjusting the term frequencies based on the importance of each field (e.g., giving higher weights to titles).

8. **Future Work**: The findings suggest that while BM25 is a robust method for document ranking, integrating cosine similarity could enhance the overall system by providing additional context. Future research could explore hybrid approaches that combine the strengths of both methods to improve the accuracy and relevance

of document retrieval systems.

In conclusion, this study highlights the importance of using appropriate ranking algorithms in information retrieval. The integration of two models—BM25, a probabilistic model, and cosine similarity, a vectorial model—demonstrated that BM25 proved to be a more effective method for ranking structured documents compared to cosine similarity, particularly in terms of balancing precision across varying document qualities. The insights gained from this research can inform future developments in search engine algorithms and information retrieval systems.