The BM25 Model

COMP3009J: Information Retrieval

Dr. David Lillis (david.lillis@ucd.ie)

UCD School of Computer Science
Beijing Dublin International College

Okapi BM25

- The BM25 Model ("BM" stands for "Best Match") came about from a series of experiments that were carried out to extend the classic probabilistic model.
- First implemented in the Okapi system that was created in London City University in the 1980s and 1990s.
- Many variants have been proposed: we will look only at BM25 itself.
- Today, BM25 is considered to be a state-of-the-art retrieval method that operates using the same principles as TF-IDF but generally performs better than the classic version we have already studied.
- Pages 104-107 of Modern Information Retrieval (2^{nd} Edition)

BM25: Reasoning

- Based on the belief that good term weighting comes from three principles:
  - Inverse document frequency: (terms that are rare across a collection should carry more weight).
  - Term frequency: (terms that are common within a document should carry more weight).
  - Document length normalisation: (so that longer documents do not get an unfair advantage if they contain query terms often simply because of their length).
- The formula evolved over time in response to many experiments being conducted.

BM25 Formula

$$sim_{BM25}(d_j, q) \sim \sum_{k_i \in d_j \wedge k_i \in q} \frac{f_{i,j} \times (1 + k)}{f_{i,j} + k\left((1 - b) + \frac{b \times len(d_j)}{avg\_doclen}\right)} \times log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)$$

- Again, we calculate a similarity score for each document d_j.

BM25 Formula

$$sim_{BM25}(d_j, q) \sim \sum_{k_i \in d_j \wedge k_i \in q} \frac{f_{i,j} \times (1 + k)}{f_{i,j} + k\left((1 - b) + \frac{b \times len(d_j)}{avg\_doclen}\right)} \times log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)$$

More words in common with the query => Good!

- The similarity score increases whenever a term k_i is in document d_j and also in the query.

BM25 Formula

$$sim_{BM25}(d_j, q) \sim \sum_{k_i \in d_j \wedge k_i \in q} \frac{f_{i,j} \times (1 + k)}{f_{i,j} + k\left((1 - b) + \frac{b \times len(d_j)}{avg\_doclen}\right)} \times log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)$$

Common words are less important (similar to IDF)

- N is the total number of documents in the collection.
- n_i is the total number of documents in the collection that contain term k_i.

BM25 Formula

$$sim_{BM25}(d_j, q) \sim \sum_{k_i \in d_j \wedge k_i \in q} \frac{f_{i,j} \times (1+k)}{f_{i,j} + k\left((1-b) + \frac{b \times len(d_j)}{avg\_doclen}\right)} \times log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)$$

Repetition of query words in the document => Good

- $f_{i,j}$ is the frequency of k_i in d_j (i.e. the number of times the term appears in the document)
- k and b are constants that can be set to suit the document collection and the desired behaviour. For general collections. k=1 and b=0.75 have been found to work well.

BM25 Formula

$$sim_{BM25}(d_j,q) \sim \sum_{k_i \in d_j \wedge k_i \in q} \frac{f_{i,j} \times (1+k)}{f_{i,j} + k\left((1-b) + \frac{b \times len(d_j)}{avg\_doclen}\right)} \times log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)$$

Repetitions are important, but are less important than different query words.

- $len(d_j)$ is the length of d_j (i.e. the number of terms in it)
- avg_doclen is the average length of a document in the collection.

BM25 Formula

$$sim_{BM25}(d_j, q) \sim \sum_{k_i \in d_j \wedge k_i \in q} \frac{f_{i,j} \times (1 + k)}{f_{i,j} + k\left((1-b) + \dfrac{b \times len(d_j)}{avg\_doclen}\right)} \times log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)$$

Repetition is more important if the document is long (relative to the average document length).

- $len(d_j)$ is the length of d_j (i.e. the number of terms in it)
- avg_doclen is the average length of a document in the collection.

BM25 Example (calculation for one document)

- **Query:** "President Lincoln"
- ***N*** = 500,000 documents
- "*president*" occurs in 40,000 documents (n_i = 40,000)
- "*lincoln*" occurs in 300 documents (n_i = 300)
- "*president*" occurs 15 times in this document ($f_{i,j}$ = 15)
- "*lincoln*" occurs 25 times in this document ($f_{i,j}$ = 25)
- The document is 90% of the length of the average ($\frac{len(d_j)}{avg\_doclen}$= 0.9)
- k = 1, b = 0.75

BM25 Calculation

$$sim_{BM25}(d_j, q) \sim \frac{15 \times 2}{15 + (0.25 + 0.75 \times 0.9)} \times log\left(\frac{500{,}000 - 40{,}000 + 0.5}{40{,}000 + 0.5}\right) + \quad \Big\} \quad k_i = \text{president}$$

$$\frac{25 \times 2}{25 + (0.25 + 0.75 \times 0.9)} \times log\left(\frac{500{,}000 - 300 + 0.5}{300 + 0.5}\right) \quad \Big\} \quad k_i = \text{lincoln}$$

$$= 27.27$$

BM25 Variations

- There are a number of variations on BM25, of which two are particularly notable:
  - BM25F: Allows different fields to be given different importance in the document (e.g. document title, headlines, main text, etc.).
  - BM25+: addresses an issue with BM25 whereby very short documents would be given scores that are too high.

BM25: Usefulness

- Unlike probabilistic model, BM25 does not require relevance information.
- Most IR researchers agree that it outperforms the vector model on general collections.