Okapi BM25

From Wikipedia, the free encyclopedia

In information retrieval, **Okapi BM25** (BM stands for Best Matching) is a ranking function used by search engines to rank matching documents according to their relevance to a given search query. It is based on the probabilistic retrieval framework developed in the 1970s and 1980s by Stephen E. Robertson, Karen Spärck Jones, and others.

The name of the actual ranking function is BM25. To set the right context, however, it usually referred to as "Okapi BM25", since the Okapi information retrieval system, implemented at London's City University in the 1980s and 1990s, was the first system to implement this function.

BM25, and its newer variants, e.g. BM25F (a version of BM25 that can take document structure and anchor text into account), represent state-of-the-art TF-IDF-like retrieval functions used in document retrieval, such as web search.

Contents

The ranking function

BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document (e.g., their relative proximity). It is not a single function, but actually a whole family of scoring functions, with slightly different components and parameters. One of the most prominent instantiations of the function is as follows.

Given a query Q, containing keywords $q_1, ..., q_n$, the BM25 score of a document D is:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})},$$

where $f(q_i, D)$ is q_i's term frequency in the document D, $|D|$ is the length of the document D in words, and $avgdl$ is the average document length in the text collection from which documents are drawn. k_1 and b are free parameters, usually chosen, in absence of an advanced optimization, as $k_1 \in [1.2, 2.0]$ and $b = 0.75$.[1] $\text{IDF}(q_i)$ is the IDF (inverse document frequency) weight of the query term q_i. It is usually computed as:

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

where N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing q_i.

There are several interpretations for IDF and slight variations on its formula. In the original BM25 derivation, the IDF component is derived from the Binary Independence Model.

Please note that the above formula for IDF shows potentially major drawbacks when using it for terms appearing in more than half of the corpus documents. These terms' IDF is negative, so for any two almost-identical documents, one which contains the term and one which does not contain it, the latter will possibly get a larger score. This means that terms appearing in more than half of the corpus will provide negative contributions to the final document score. This is often an undesirable behavior, so many real-world applications would deal with this IDF formula in a different way:

- Each summand can be given a floor of 0, to trim out common terms;
- The IDF function can be given a floor of a constant ϵ, to avoid common terms being ignored at all;
- The IDF function can be replaced with a similarly shaped one which is non-negative, or strictly positive to avoid terms being ignored at all.

IDF information theoretic interpretation

Here is an interpretation from information theory. Suppose a query term q appears in $n(q)$ documents. Then a randomly picked document D will contain the term with probability $\frac{n(q)}{N}$ (where N is again the cardinality of the set of documents in the collection). Therefore, the information content of the message "D contains q" is:

$$-\log \frac{n(q)}{N} = \log \frac{N}{n(q)}.$$

Now suppose we have two query terms q_1 and q_2. If the two terms occur in documents entirely independently of each other, then the probability of seeing both q_1 and q_2 in a randomly picked document D is:

$$\frac{n(q_1)}{N} \cdot \frac{n(q_2)}{N},$$

and the information content of such an event is:

$$\sum_{i=1}^{2} \log \frac{N}{n(q_i)}.$$

With a small variation, this is exactly what is expressed by the IDF component of BM25.

Modifications

- At the extreme values of the coefficient b BM25 turns into ranking functions known as **BM11** (for $b = 1$) and **BM15** (for $b = 0$).[2]
- **BM25F**[3] is a modification of BM25 in which the document is considered to be composed from several fields (such as headlines, main text, anchor text) with possibly different degrees of importance.
- **BM25+**[4] is an extension of BM25. BM25+ was developed to address one deficiency of the standard BM25 in which the component of term frequency normalization by document length is not properly lower-bounded; as a result of this deficiency, long documents which do match the query term can often be scored unfairly by BM25 as having a similar relevancy to shorter documents that do not contain the query term at all. The scoring formula of BM25+ only has one additional free parameter δ (a default value is 1.0 in absence of a training data) as compared with BM25:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \left[\frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})} + \delta \right]$$

Footnotes

1. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. *An Introduction to Information Retrieval*, Cambridge University Press, 2009, p. 233.
2. http://xapian.org/docs/bm25.html
3. Hugo Zaragoza, Nick Craswell, Michael Taylor, Suchi Saria, and Stephen Robertson. *Microsoft Cambridge at TREC-13: Web and HARD tracks.* (http://trec.nist.gov/pubs/trec13/papers/microsoft-cambridge.web.hard.pdf) In Proceedings of TREC-2004.
4. Yuanhua Lv and ChengXiang Zhai. *Lower-bounding term frequency normalization.* (http://sifaka.cs.uiuc.edu/~ylv2/pub/cikm11-lowerbound.pdf) In Proceedings of CIKM'2011, pages 7-16.

References

- Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford (November 1994). *Okapi at TREC-3* (http://trec.nist.gov/pubs/trec3/papers/city.ps.gz). Proceedings of the Third Text REtrieval Conference (TREC 1994) (http://trec.nist.gov/pubs/trec3/t3_proceedings.html). Gaithersburg, USA.

- Stephen E. Robertson, Steve Walker, and Micheline Hancock-Beaulieu (November 1998). *Okapi at TREC-7* (http://trec.nist.gov/pubs/trec7/papers/okapi_proc.pdf.gz). Proceedings of the Seventh Text REtrieval Conference (http://trec.nist.gov/pubs/trec7/t7_proceedings.html). Gaithersburg, USA.

- Spärck Jones, K.; Walker, S.; Robertson, S. E. (2000). "A probabilistic model of information retrieval: Development and comparative experiments: Part 1". *Information Processing &*

Management **36** (6): 779–808. doi:10.1016/S0306-4573(00)00015-7 (https://dx.doi.org/10.1016%2FS0306-4573%2800%2900015-7).

- Spärck Jones, K.; Walker, S.; Robertson, S. E. (2000). "A probabilistic model of information retrieval: Development and comparative experiments: Part 2". *Information Processing & Management* **36** (6): 809–840. doi:10.1016/S0306-4573(00)00016-9 (https://dx.doi.org/10.1016%2FS0306-4573%2800%2900016-9).

External links

- Robertson, Stephen; Zaragoza, Hugo (2009). *The Probabilistic Relevance Framework: BM25 and Beyond* (http://staff.city.ac.uk/~sb317/papers/foundations_bm25_review.pdf). NOW Publishers, Inc. ISBN 978-1-60198-308-4.

Retrieved from "http://en.wikipedia.org/w/index.php?title=Okapi_BM25&oldid=649835179"

Categories: Ranking functions | Information retrieval