



Okapi BM25

In information retrieval, **Okapi BM25** (*BM* is an abbreviation of *best matching*) is a ranking function used by search engines to estimate the relevance of documents 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*. The fuller name, *Okapi BM25*, includes the name of the first system to use it, which was the Okapi information retrieval system, implemented at London's City University^[1] in the 1980s and 1990s. BM25 and its newer variants, e.g. BM25F (a version of BM25 that can take document structure and anchor text into account), represent TF-IDF-like retrieval functions used in document retrieval.^[2]

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 their proximity within the document. It is a 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 $q_1, \dots, q_n$, the BM25 score of a document D is:

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

where $f(q_i, D)$ is the number of times that the keyword q_i occurs 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 $b = 0.75$.^[3] $\text{IDF}(q_i)$ is the IDF (inverse document frequency) weight of the query term q_i . It is usually computed as:

$$\text{IDF}(q_i) = \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right)$$

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.

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 $b = 0$).^[4]
- BM25F**^{[5][2]} (or the **BM25 model with Extension to Multiple Weighted Fields**^[6]) 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, term relevance saturation and length normalization. BM25F defines each type of field as a *stream*, applying a per-stream weighting to scale each stream against the calculated score.
- BM25+**^[7] 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:

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

IPR2025-01398

References

1. "OKAPI" (<https://web.archive.org/web/20231207112813/https://smcse.city.ac.uk/doc/cisr/web/okapi/okapi.html>). *smcse.city.ac.uk*. Archived from the original (<https://smcse.city.ac.uk/doc/cisr/web/okapi/okapi.html>) on 2023-12-07. Retrieved 2023-10-16.
2. Stephen Robertson & Hugo Zaragoza (2009). "The Probabilistic Relevance Framework: BM25 and Beyond" (<http://dl.acm.org/citation.cfm?id=1704810>). *Foundations and Trends in Information Retrieval*. **3** (4): 333–389. CiteSeerX 10.1.1.156.5282 (<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.156.5282>). doi:10.1561/15000000019 (<https://doi.org/10.1561%2F15000000019>). S2CID 207178704 (<https://api.semanticscholar.org/CorpusID:207178704>).
3. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. *An Introduction to Information Retrieval*, Cambridge University Press, 2009, p. 233.
4. "The BM25 Weighting Scheme" (<http://xapian.org/docs/bm25.html>).
5. 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.
6. Robertson, Stephen; Zaragoza, Hugo; Taylor, Michael (2004-11-13). "Simple BM25 extension to multiple weighted fields" (<https://doi.org/10.1145/1031171.1031181>). *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. CIKM '04. New York, NY, USA: Association for Computing Machinery. pp. 42–49. doi:10.1145/1031171.1031181 (<https://doi.org/10.1145%2F1031171.1031181>). ISBN 978-1-58113-874-0. S2CID 16628332 (<https://api.semanticscholar.org/CorpusID:16628332>).
7. 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.

General references

- Stephen E. Robertson; Steve Walker; Susan Jones; Micheline Hancock-Beaulieu & 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 & 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. CiteSeerX 10.1.1.134.6108 (<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.134.6108>). doi:10.1016/S0306-4573(00)00015-7 (<https://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://doi.org/10.1016%2FS0306-4573%2800%2900016-9>).
- Stephen Robertson & Hugo Zaragoza (2009). "The Probabilistic Relevance Framework: BM25 and Beyond" (<http://dl.acm.org/citation.cfm?id=1704810>). *Foundations and Trends in Information Retrieval*. **3** (4): 333–389. CiteSeerX 10.1.1.156.5282 (<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.156.5282>).

ary?doi=10.1.1.156.5282). doi:10.1561/15000000019 (<https://doi.org/10.1561%2F15000000019>).
S2CID 207178704 (<https://api.semanticscholar.org/CorpusID:207178704>).

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) (PDF). NOW Publishers, Inc. ISBN 978-1-60198-308-4.
-

Retrieved from "https://en.wikipedia.org/w/index.php?title=Okapi_BM25&oldid=1302762902"