Maciej Wieczorek
maciejw2@illinois.edu

BM25, its derivatives and some extensions

CS410 - Technology review

Introduction to BM25

The *Best Match* family of functions are term-weighting functions and algorithms used as ranking functions in search engines to order documents for a given query. They operate on the bag-of-words model, where documents and queries are tokenized into shorter, usually word-length components though longer n-grams are possible, and larger structures like sentences and word order are disregarded. The functions incorporate Term Frequency, Inverse Document Frequency and Document Length Normalization heuristics to improve ranking accuracy and these heuristics include parameters to tune their weights.

The best known and most commonly used of the BM family is BM25 (Best Match 25), which when its parameters are set to extreme values results in special case models BM0, BM1, BM11 and BM15, and which has many extensions such as BM25+ and BM25t which address specific shortcoming. The general BM25 model is composed of local, $L_{i,j}$, and global, G_i, weight components:

$$L_{i,j} = \left(\frac{f_{i,j}\,(k_1 + 1)}{f_{i,j} + k_1 \left((1 - b) + b \left(\frac{\mathrm{dl}_j}{\mathrm{dl}_{ave}} \right) \right)} \right)$$

$$G_i = \log \left(\frac{(r + k)/(R - r + k)}{(n - r + k)/(N - n - R + r + k)} \right)$$

$$w_{i,j} = G_i L_{i,j}$$

$$= \left(\frac{(r + k)/(R - r + k)}{(n - r + k)/(N - n - R + r + k)} \right) \left(\frac{f_{i,j}\,(k_1 + 1)}{f_{i,j} + k_1 \left((1 - b) + b \left(\frac{\mathrm{dl}_j}{\mathrm{dl}_{ave}} \right) \right)} \right)$$

where
$f_{i,j}$ = frequency of term *i* in document *j*
k_1 = smoothing parameter for term frequency saturation
b = parameter for document length normalization
dl_j = length of document *j*
dl_{ave} = average document length in collection
r = number of relevant documents that contain term *i*

$n - r$ = number of non-relevant documents that contain term *i*
n = number of documents that contain term *i*
$R - r$ = number of relevant documents that do not contain term *i*
$N - n - R + r$ = number of non-relevant documents that do not contain term *i*
$N - n$ = number of documents that do not contain term *i*
R = number of relevant documents
$N - R$ = number of non-relevant documents
N = number of documents
k = smoothing factor

Best Match model family

Derivates:

BM0

This base case applies equal weights to all terms:

$$w_{i,j} = 1$$

BM1

In BM1, no relevance information is provided from $L_{i,j}$ and only the probabilistic document weights are used:

$$w_i = \log\left(\frac{(r+k)/(R-r+k)}{(n-r+k)/(N-n-R+r+k)}\right)$$

BM11

In BM11, b is set to the extreme value of 1 and document lengths are fully normalized, so that longer documents are fully penalized for their length.

$$w_{i,j} = \left(\frac{(r+k)/(R-r+k)}{(n-r+k)/(N-n-R+r+k)}\right)\left(\frac{f_{i,j}\,(k_1+1)}{f_{i,j} + k_1\left(\frac{\mathrm{dl}_j}{\mathrm{dl}_{ave}}\right)}\right)$$

BM15

In BM15, b is set to the extreme value of 0 which removes the document length normalization and any penalty associated with relatively longer documents compared to the corpus.

$$w_{i,j} = \left(\frac{(r+k)/(R-r+k)}{(n-r+k)/(N-n-R+r+k)}\right)\left(\frac{f_{i,j}\,(k_1+1)}{f_{i,j} + k_1}\right)$$

Extensions:

BM25F

BM25F is an extension to BM25 which adjusts the weights of terms in different types of fields in a document. The base BM25 function is built around unstructured documents and does not differentiate or account for the value that different sections of a document may bring. For example, a term found in a document title, summary or abstract is more likely to indicate the topic of a document than if it was found strictly in the body. The approach that BM25F takes to address this is to allow different weights to be assigned to each document section. For example, an abstract section may be assigned a weight of 4, and the body section a weight of 1. This would be equivalent to multiplying the count of terms in the summary by 4 and using this pseudo-frequency in place of the actual frequency in the general BM25 function. Note that this effectively lengthens the document and affects its length, now pseudo-length.

BM25+

BM25+ is an extension to BM25 which adjusts the lower bound of document length normalization. When $\mathrm{dl}_j \gg \mathrm{dl}_{ave}$, the document length normalization may penalize such long documents such that even those that contain terms matching the query may be ranked below non-relevant documents. The approach that BM25+ takes to address this is to add a pseudo TF value δ to control the scale of the lower bound to ensure a suitable gap between documents matching and missing a term.

$$\sum_{t \in Q \cap D} \frac{(k_3+1)c(t,Q)}{k_3 + c(t,Q)} \times \left[\frac{(k_1+1)c(t,D)}{k_1\left(1 - b + b\frac{|D|}{\mathrm{avdl}} + c(t,D)\right)} + \delta \right] \times \log \frac{N+1}{df(t)}$$

BM25t

BM25t is an extension to BM25 which adjusts the weights of terms in structured document formats like XML according to the tag (eg: bold, italic, centred) or section (eg: title, heading, body). This is similar BM25F, but extends the idea to more structured documents. The approach that BM25t takes to address this is to add a vector m_j with dimensions $(T_{10}, \cdots, T_{1k}, \cdots, T_{1m}, \cdots, T_{n0}, \cdots, T_{nk}, \cdots, T_{nm})$ which correspond to tags within the document, and where

$T_{ik} = 1$ if term t occurs in tag b_k
$T_{ik} = 0$ if term t does not occur in tag b_k
$T_{i0} = 1$ if term t occurs without any tags
$T_{i0} = 0$ if term t does not occur without any tags

Appropriate weights are then found for each T_{ik} through training.

Conclusion

BM25 is state-of-the-art ranking function based on bag-of-words representations and considering TF, IDF and DL normalizations. Its generalization can be constrained to derive special cases BM1, BM0, BM11 and BM15, or can be extended through various means and ways to address any shortcomings, for example to account for structured documents (BM25F, BM25t) or document edge cases when $\mathrm{dl}_j \gg \mathrm{dl}_{ave}$ (BM25+). Its flexibility and extendibility should ensure that it remains relevant and performant in the future.

References

S. E. Robertson and S. Walker. Some Simple Effective Approximations to the 2–Poisson Model for Probabilistic Weighted Retrieval. In *SIGIR '94*, pages 345–354, 1994

S. E. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *CIKM '04*, pages 42–49, 2004

Y. Lv and C. Zhai. Lower-bounding term frequency normalization. In *CIKM'11*, pages 7-16, 2011

M. Géry, C. Largeron, and F. Thollard. BM25t, une extension de BM25 pour la recherche d'information ciblée, In *Document numérique*, vol. 13, no. 1, pages 83-110, 2010

S.E. Robertson and H. Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. In *Foundations and Trends in Information Retrieval: vol. 3: no. 4*, pages 333–389, 2009

E. Garcia. A Tutorial on the BM25F model. http://www.minerazzi.com/tutorials/bm25f-model-tutorial.pdf

H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok. Probabilistic document-context based relevance feedback with limited relevance judgments. In *CIKM'06*, pages 854–855, 2006