A Tutorial on the BM25F Model

Abstract – This is a tutorial on the Best Match 25 Model with Extension to Multiple Weighted Fields, also known as the BM25F Model. Unlike BM25, the model is applicable to structured documents consisting of multiple fields. The model preserves term frequency nonlinearity and removes the independence assumption between same term occurrences.

Keywords: bm25f, bm25, weighted fields, structured documents, nonlinearity, independence assumption
Published: 08-02-2011; Updated: 10-11-2016

Note: This article is part of a legacy series that the author published circa 2011 at http://www.miislita.com, now a search engine site. It is now republished in pdf format here at http://www.minerazzi.com, with its content edited and updated.

Introduction

In a previous tutorial (Garcia, 2016a), we discussed Okapi Best Match 25 Model, commonly referred to as BM25 (Robertson & Zaragoza, 2009). The model computes local weights as parameterized term frequencies and global weights as RSJ weights.

BM25, however, lefts out the structure of documents in the weighting process. To address this issue, Robertson and Zaragoza (Robertson, Zaragoza, & Taylor, 2004) proposed a simple BM25 extension for weighting terms present in multiple fields that they referred to as the BM25F Model. The purpose of this article is to briefly describe BM25F, considered by many an elegant approach.

Background

Let us first review what we know about BM25. This model defines the weight of term i in document j as the product of local, $L_{i,j}$, and global, G_i, weights; i.e., $w_{i,j} = L_{i,j}G_i$ where

$$L_{i,j} = \left(\frac{f_{i,j}\,(k_1+1)}{k_1\left((1-b) + b\left(\frac{dl_j}{dl_{ave}}\right)\right)+f_{i,j}} \right) \tag{1}$$

$$G_i = \mathrm{F4} = \log\left(\frac{(\mathrm{r}+\mathrm{k})/(\mathrm{R}-\mathrm{r}+\mathrm{k})}{(\mathrm{n}-\mathrm{r}+\mathrm{k})/(\mathrm{N}-\mathrm{n}-\mathrm{R}+\mathrm{r}+\mathrm{k})}\right) \tag{2}$$

where F4 is an RSJ weight (Robertson & Spärck-Jones, 1976). Therefore,

$$w_{i,j} = L_{i,j} G_i = \left(\frac{f_{i,j}(k_1+1)}{k_1\left((1-b) + b\left(\frac{dl_j}{dl_{ave}}\right)\right)+f_{i,j}} \right) \log\left(\frac{(r+k)/(R-r+k)}{(n-r+k)/(N-n-R+r+k)}\right) \quad (3)$$

where

$f_{i,j}$	=	frequency of term *i* in document *j*
k_1	=	a smoothing parameter for adjusting term frequencies saturation
b	=	parameter for achieving full, soft, or zero document length normalization
dl_j	=	length of document *j* as the sum of all of its *m* terms, $dl_j = \sum_i^m f_{i,j}$
dl_{ave}	=	average document length in a collection of documents
r	=	number of relevant documents in a collection containing term *i*
n - r	=	number of nonrelevant documents in a collection containing term *i*
n	=	number of documents in a collection containing term *i*
R – r	=	number of relevant documents that do not contain term *i*
N – n– R + r	=	number of nonrelevant documents of a collection that do not contain term *i*
N – n	=	number of documents in a collection that do not contain term *i*
R	=	number of relevant documents in a collection
N – R	=	number of nonrelevant documents in a collection
N	=	number of documents in a collection
k	=	a smoothing correction

Experimental Conditions

The model provides no guidance on how k, *b*, and k_1 should be set. Typical values are k = 0.5, $0.5 < b < 0.8$, and $1.2 < k_1 < 2$. Acceptable results are obtained in most cases with $\frac{b}{k_1} < 1$.

If R = 0, r = 0, and k = 0, (2) reduces to a probabilistic inverse document frequency (IDFP); i.e. $G_i = \text{IDFP} = \log\left(\frac{N-n}{n}\right)$. However, if a term is mentioned by all, half, or more than half of the documents of a collection, this approximation gives $G_i = \text{undefined}$, $G_i = 0$, or $G_i < 0$. In addition, setting k = 0.5 does not really solve the question of negative weights.

As mentioned in previous articles, the use of logs for transforming data is just one type of Box-Cox Power Transformations (Garcia, 2016b; 2016c; 2016d). In these transformations, an extra term is added to the data to be transformed, to offset any zero or negative value.

We can do the same here by adding an extra term *l* like this

$$G_i = \log\left(\frac{\mathrm{N}-\mathrm{n}}{\mathrm{n}} + l\right) \tag{4}$$

where *l* can be viewed as a *lift* as that's what it does to a curve of G_i vs. n values. This simple modification avoids all of the above complications and makes G_i robust against negative weights.

Notice that $l = 1$ returns $G_i = \log\left(\frac{\mathrm{N}}{\mathrm{n}}\right) = \mathrm{IDF}$, $l > 1$ avoids negative weights altogether, and $l = 2$ yields $G_i = \log\left(1 + \frac{\mathrm{N}}{\mathrm{n}}\right)$, an expression derived by Lee (2007). The difference between Lee's derivation and ours is that she used some layers of abstractions while we simply applied Box-Cox data transformation theory.

With regard to the setting of b and k_1, the following table lists the family of BM weights that can be derived by setting these parameters (Garcia, 2016a).

Table 1. Family of Best Match Models.

Model	Weight, $w_{i,j} = L_{i,j}G_i$	Parameters
BM25	$w_{i,j} = \left(\frac{f_{i,j}\,(k_1+1)}{k_1\left((1-b) + b\left(\frac{dl_j}{dl_{ave}}\right)\right) + f_{i,j}}\right)\mathrm{F4}$	$0 < b < 1$ $k_1 > 0$
BM15	$w_{i,j} = \left(\frac{f_{i,j}(k_1+1)}{k_1 + f_{i,j}}\right)\mathrm{F4}$	$b = 0$ $k_1 > 0$
BM11	$w_{i,j} = \left(\frac{f_{i,j}\,(k_1+1)}{k_1\left(\frac{dl_j}{dl_{ave}}\right) + f_{i,j}}\right)\mathrm{F4}$	$b = 1$ $k_1 > 0$
BM1	$w_{i,j} = \mathrm{F4}$	$k_1 = 0$
BM0	$w_{i,j} = 1$	-

As Table 1 shows, BM25 = BM15 = BM11 for documents of average lengths while these reduce to BM1 for $k_1 = 0$ which is a binary weight; i.e., $w_{i,j} = 1$ if $f_{i,j} > 0$; otherwise $w_{i,j} = 0$.

The Challenge of Structured Documents

Most, though not all, information retrieval models consider documents as unstructured text. For instance, (1) and (2) were originally developed for scoring unstructured text. However, Web documents are frequently structured, consisting of specific fields or sections like markup tags (head, body, title, description, paragraphs, divisions, forms, etc).

How could we apply a BM25 derivative model to these documents? Trying to combine or concatenate all document fields into a non-structured pseudo-document is certainly out of the question. How about treating each field type as collections of unstructured *document fields*? For example, for scientific journal papers we may construct collections of titles, abstracts, and body sections. For Web documents, we may include anchor text incoming links (in-links). This is the text of links pointing to the document to be scored.

We could apply BM25 to each field collection and form a linear combination of these scores. Unfortunately, this approach destroys the nonlinearity relationship between term weights and term frequencies, restoring term independence.

The second problem is that scoring term weights by combining field types opens the question of how to collect global field statistics like IDF values for the individual fields. For instance, titles are short fields and body large fields. Since frequently used terms in body fields may rarely occur in title fields these should receive a high weight in the title score. It turns out that because a frequently used term is defined in relation to a field type, the result would be very unstable IDF statistics.

The third problem is that there would be no easy way of interpreting the meaning of merging evenly weighted field types. For instance due to the nonlinearity nature of term frequencies, setting all field weights to 1 does not restore the unstructured scenario of equivalently merging all fields into a large unstructured field.

A fourth problem that arises from constructing collections of field types is how to normalize the length of the fields. As noted by Robertson, et al. (2004, 2009), the initial reason for document length normalization in BM25 is to account for the verbosity and scope of the documents. It is not that clear whether normalization based on verbosity and scope should apply to the different fields, or if the whole document length should be used.

Finally, there is the question of how to optimize k_1 and b for each field type.

An Elegant Approach: BM25F

An elegant treatment consists in weighting term frequencies accordingly to their field importance, combining them, and then using the resulting pseudo-frequencies.

To illustrate, let *s* be a field or stream and *v* its weight. Suppose a document is decomposed into two fields or streams: title and body. If we assign a weight of 6 to terms in the title and a weight of 2 to terms in the body, this is equivalent to replacing the document by itself but this time repeating the title six times and the original body twice.

The resulting pseudo-frequency of a term is now a linear combination of these weighted fields

$$\widetilde{f_{i,j}} = \sum_{s=1}^{S} v_s f_{i,s} \tag{5}$$

Applying this to all terms, the new document length is

$$\widetilde{dl_j} = \sum_{i}^{m} \widetilde{f_{i,j}} \tag{6}$$

The new average document length over this pseudo-collection is

$$\widetilde{dl_{ave}} = \frac{\sum_{j=1}^{N} \widetilde{dl_j}}{N} \tag{7}$$

We can now use (5 – 7) in (3). This makes term weights nonlinear with the pseudo-frequencies, preserving the statistical dependence of terms and their IDF values in the original collection.

Finally, the unstructured scenario is restored by setting $v = 1$ for all fields, without compromising statistical dependence. This is essentially the beauty of the BM25F Model. That's why we refer to the model as a simple, elegant approach to a very complex problem.

Combining BM25F with other Models

When BM25F is combined with other models the result is a superior model. For instance, Najork, Zaragoza, and Taylor (2007) found that a combination of BM25F and simple in-degree link analysis outperforms the combination of BM25F with PageRank or HITS authority scores. Scores

were also much easier and faster to compute. In addition, Najork (2007) found that the combination of SALSA and BM25F outperforms the combination of HITS and BM25F.

Robertson, et al (2009) have made a distinction between the Simple BM25F and a modified version presented at TREC-13 (Zaragoza, Craswell, Taylor, Saria, & Robertson, 2004) in which the several free parameters of BM25F are allowed to vary between fields. This variable BM25F presents new challenges.

BM25F, LETOR, and Simplex Optimization

To address the problem of optimizing k_1 and b, and of convergence around a global optimum, Robertson et al. have tried some heuristic techniques and tricks, including scaling of k_1 according to changes in the average term frequencies and document lengths. They have also developed what they call a *robust linear search optimization* technique (Robertson, et. al, 2004, 2009).

Although problems associated with the optimization of ranking functions is the focus of LETOR (learning to rank) models (Joachims, Li, Liu, & Zhai, 2007; Li, Liu, & Zhai, 2008), these types of problems can also be examined with some of the robust algorithms described in the Sequential Simplex Optimization (SSO) literature.

Developed in the ‘60s by Nelder and Mead, and since its introduction to Chemometrics by Deming and Morgan in the early ‘70s, SSO has been used extensively in science and engineering disciplines for the optimization of experimental parameters and working conditions (Walters, Parker, Morgan, & Deming, 1991; Allman, 1995; Michałowska-Kaczmarczyk & Michałowski, 2014).

Conclusion

Robertson and co-workers have developed the family of BM25 models in stages over a period of several decades, with BM25 being the best known of these. Unlike vector space models found in the IR literature, these models account for the verbosity and scope of documents and their lengths.

The incorporation of weighted frequencies based on the structure of documents as captured by BM25F represents a remarkable improvement, but is not free from drawbacks. Unfortunately, BM25-based models are in practice difficult to implement efficiently, requiring of parameterized functions. In recent years, binned or document-centric impact models have been developed to overcome some of these efficiency issues (Anh & Moffat, 2004; 2005; Metzler, Strohman, & Croft, 2008).

References

Allman, M. C. (1995). Sequential Simplex Optimization. Retrieved from http://www.allmanpc.com/site_files/papers_presentations/Seq_Simplex.pdf

Anh, V.N. and Moffat, A. (2004). Collection-independent document-centric impacts. In: Proc. Australian Document Computing Symposium. 25 – 32. Retrieved from https://pdfs.semanticscholar.org/1779/d097851fdbbe1b3c5fe182ddab43bc4c136c.pdf.

Anh, V.N. and Moffat, A. (2005). Simplified similarity scoring using term ranks. In: Proc. 28th SIGIR. 226 – 233. Retrieved from http://www2.dcc.ufmg.br/eventos/sigir2005/files/talks-papers-2005-08-11/AlistairMoffat.pdf

Garcia, E. (2016a). A Tutorial on Okapi BM25 Model. Retrieved from http://www.minerazzi.com/tutorials/okapi-bm25-model.pdf

Garcia, E. (2016b). Robertson-Spärck-Jones Probabilistic Model Tutorial. Retrieved from http://www.minerazzi.com/tutorials/probabilistic-model-tutorial.pdf

Garcia, E. (2016c). Development of BM25IR: A Best Match Model based on Inverse Regression. Retrieved from http://www.minerazzi.com/tutorials/bm25ir.pdf

Garcia, E. (2016d). A Tutorial on Quantile-Quantile Plots. Retrieved from http://www.minerazzi.com/tutorials/quantile-quantile-tutorial.pdf

Joachims, T., Li, H., Liu, & Zhai, C. (2007). *Learning to rank for information retrieval* (LR4IR 2007). SIGIR Forum, vol. 41, no. 2, pp. 58–62, 2007. Retrieved from http://dl.acm.org/citation.cfm?id=1328974&dl=ACM&coll=DL&CFID=849711571&CFTOKEN=99377890 See also http://research.microsoft.com/en-us/um/beijing/events/lr4ir-2007/

Lee, L. (2007). IDF Revisited: A Simple New Derivation within the Robertson-Spärck Jones Probabilistic Model. SIGIR 07, July 23-27,2007. Amsterdam, The Netherlands. ACM 978-1-59593-597-7/07/0007. Retrieved from http://www.cs.cornell.edu/home/llee/papers/idf.pdf

Li, H., Liu, T. Y., & Zhai, C. (2008). *Learning to rank for information retrieval* (LR4IR 2008). SIGIR Forum, vol. 42, no. 2, pp. 76–79, 2008. Retrieved from http://research.microsoft.com/en-us/um/beijing/events/lr4ir-2008/PROCEEDINGS-LR4IR%202008.PDF
See also https://pdfs.semanticscholar.org/7028/c325f161611f6bce252083c4815b61be9927.pdf and http://www2009.org/pdf/T7A-LEARNING%20TO%20RANK%20TUTORIAL.pdf

Metzler, D., Strohman, T., and Croft, W. B. (2008). A Statistical View of Binned Retrieval Models. Advances in Information Retrieval. Volume 4956 of the series Lecture Notes in Computer Science pp 175-186. Springer. Retrieved from https://pdfs.semanticscholar.org/0814/4a3c963e7af72c0756abf01fbc2551ed7f89.pdf

Michałowska-Kaczmarczyk, A. M. & Michałowski, T. (2014). Simplex Optimization and Its Applicability for Solving Analytical Problems. Journal of Applied Mathematics and Physics, 2, 723-736. Retrieved from http://file.scirp.org/pdf/JAMP_2014062509510476.pdf

Najork, M., Zaragoza, H., & Taylor M. (2007). HITS on the Web: How does it Compare? SIGIR, 2007. Retrieved from http://research.microsoft.com/apps/pubs/default.aspx?id=65139 See also https://www.microsoft.com/en-us/research/wp-content/uploads/2007/07/sigir2007.pdf

Robertson, S. E., & Spärck-Jones, K. (1976). Relevance weighting of search terms, Journal of the American Society for Information Science, Volume 27, 1976 pp. 129–146. Retrieved from http://www.staff.city.ac.uk/~sb317/papers/RSJ76.pdf

Robertson, S. E., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval, Vol. 3, No. 4 (2009) 333–389. http://www.gbv.de/dms/tib-ub-hannover/632343664.pdf

Robertson, S. E., Zaragoza, H., & Taylor, M. (2004). Simple BM25 Extension to Multiple Weighted Fields (2004). Retrieved from http://www.hugo-zaragoza.net/academic/pdf/robertson_cikm04.pdf See also http://citeseer.ist.psu.edu/viewdoc/download;jsessionid=549257EC3BC5B8BD1D0195241A931602?doi=10.1.1.9.5255&rep=rep1&type=pdf

Walters, F. H., Parker, L. R., Morgan, S. L., & Deming, S. N. (1991). Sequential Simplex Optimization. CRC Press. Retrieved from http://www.chemeng.kmutt.ac.th/cheps/Sequential_Simplex_Optimization.pdf

Zaragoza, H., Craswell, N., Taylor, M., Saria, S., & Robertson, S. (2004). Web and HARD track. Microsoft Cambridge at TREC-13 (2004). In Proceedings of 13th Annual Text Retrieval Conference. Retrieved from http://trec.nist.gov/pubs/trec13/papers/microsoft-cambridge.web.hard.pdf