A Tutorial on OKAPI BM25 Model

Abstract – This is a light tutorial on OKAPI BM25, a Best Match model where local weights are computed as parameterized frequencies and global weights as RSJ weights. Local weights are based on a 2-Poison model and the verbosity and scope hypotheses and global weights on the Robertson-Spärck-Jones Probabilistic Model.

Keywords: okapi, best match models, bm25, poisson-2 model, verbosity hypothesis, scope hypothesis
Published: June 30, 2011; Updated: October 22, 2018

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 the early 80s Gillian Venner, Nathalie Mitev, and Stephen Walker (1985, 1987) conducted research work that led to the design and evaluation of online public access catalogs (OPACs) at Polytechnic of Central London (PCL).

The project initial phases spanned from November 1982 to May 1985. The prototype was named OKAPI (Online Keyword Access to Public Information). As Mitev (1985) wrote:

> "*Designing an online public access catalogue* [OPAC]: *Okapi, a catalogue on a local area network* [LAN] is the final report of a two-year research project "Microprocessor networking in libraries" which was funded by the British Library and the Department of Trade and Industry, and based at the Polytechnic of Central London."
>
> "The aim was to produce an OPAC on a LAN, that would be readily usable without training or experience, without sacrificing effectiveness or being tedious for experienced users."
>
> "The result was a functioning prototype OPAC called Okapi, which has a number of distinctive features: use is eased by coloured keys and a lack of jargon; the system uses search decision trees to select a suitable action at each stage of a search, and it performs automatic Boolean and hyper-Boolean functions where appropriate. The OPAC was installed and evaluated in one of the Polytechnic site libraries."

That research predates the first Web search engines, including ARCHIE. Back then OKAPI operated on a LAN (local area network) using Apple IIe computers. In July 1989, OKAPI moved from PCL to the Centre for Interactive Systems Research at City University (Walker & Beaulieu, 1991). So the idea that OKAPI originated at City University is not accurate.

Under Robertson, Walker, and others OKAPI used NIST's TREC to improve its term weighting functions and algorithms (Robertson, Walker, Jones, Hancock-Beaulieu, & Gatford, 1996).

Implementations included, though not limited to, Best Match models, a combination of global and local term weighting functions, with global weights computed as RSJ weights and local weights as parameterized term frequencies (Robertson & Spärck-Jones, 1976; Robertson, Walker, & Beaulieu, 2000). The best known of these models is BM25 (Best Match 25). SIGIR's digital museum (SIGIR, 2016) provides an interesting account of the origins of OKAPI. An OKAPI-PACK used to be available for download from Macfarlane (2001).

As nowadays new generations of computer science students are discovering information retrieval systems, we believe that writing a light tutorial on BM25 is more than appropriate.

Best Match Components

BM models incorporate local and global weight components. The precursor of the ranking functions of these models is a formula of the general form (Robertson & Zaragoza, 2009)

$$w_{i,j} = L_{i,j} G_i = \left(\frac{f_{i,j}}{k + f_{i,j}}\right) \text{F4} \qquad (1)$$

In (1) the terms are defined as follows:

$w_{i,j}$ = weight of term i in document j
$L_{i,j}$ = local weight of term i in document j
G_i = global weight of term i
$f_{i,j}$ = frequency of term i in document j
k = a smoothing correction
F4 = a best match scoring function that compute RSJ weights

where

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

and where

r	=	number of relevant documents that contain the term.
n – r	=	number of non-relevant documents that contain the term.
n	=	number of documents that contain the term.
R – r	=	number of relevant documents that do not contain the term.
N – n – R + r	=	number of non-relevant documents that do not contain the term.
N - n	=	number of documents that do not contain the term.
R	=	number of relevant documents.
N – R	=	number of non-relevant documents.
N	=	number of documents in the collection.
k	=	a smoothing correction usually set to 0.5

If R = 0, r = 0, and no smoothing correction is used, (2) reduces to a probabilistic inverse document frequency (IDFP).

$$F4 = IDFP = \log\left(\frac{N-n}{n}\right) \quad (3)$$

Thus IDFP, and in general the IDF concept, is an RSJ probabilistic weight in the absence of relevance information (Robertson, 2004).

Since the global weight components of these models were discussed in a previous tutorial (Garcia 2016), this time the discussion is focused on their local weight components; that is, in defining a best match scoring function for $L_{i,j}$.

BM Local Weight Components

In (1), $L_{i,j}$ is an approximation of a 2-Poison model where it is assumed that all documents are of same length and the distribution of within-document term frequencies is Poison for *elite* and *non-elite* documents (Robertson & Walker, 1994).

Elite documents are those that are "about" the concept represented by a term. Eliteness is a binary property: a document is either elite to a term or not. If a document is elite to terms it mentions and these are query terms, more likely the document is *relevant* to the query.

Representing local weights as a 2-Poisson model has the following characteristics:

- $L_{i,j}$ is zero when $f_{i,j} = 0$;
- $L_{i,j}$ increases monotonically with $f_{i,j}$;
- $L_{i,j}$ approaches an asymptotic maximum value of 1.

Figure 1 depicts profile curves of $L_{i,j}$ in terms of k and $f_{i,j}$.

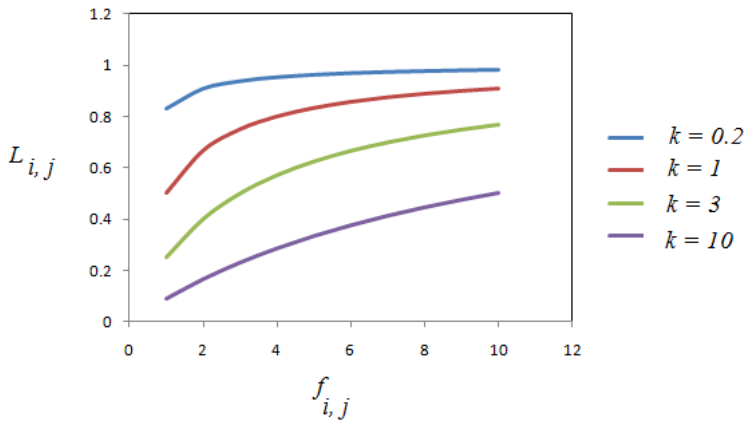

Figure 1. Profile curves of $L_{i,j}$ as a function of k and $f_{i,j}$.

The absolute positions of these curves in the graph are not important. What is important is the relative increments for different increments in $f_{i,j}$.

For high k, increments in $f_{i,j}$ continue to contribute significantly to the local weights. For low k, the additional contribution of a newly observed occurrence quickly reaches a saturation point.

That makes sense. According to Robertson, Zaragoza, and Taylor (2004)

> "Most modern weighting functions based on term frequencies (*tf*) are nonlinear in this parameter. This is desirable because of the statistical dependence of term occurrences: the information gained on observing a term the first time is greater than the information gained on subsequently seeing the same term."

Accordingly, the largest gain in information, then in $L_{i,j}$, occurs from $f_{i,j} = 0$ to $f_{i,j} = 1$. Subsequent changes in the frequency of a term by a given factor should not change by the same factor its weight and the relevance of a document to said term.

That assertion contradicts the idea that repeating a term x times makes the document x times more relevant. It also works against the concept of *keyword density* promoted by SEOs. In any case, we should expect short documents to reach the saturation scenario quicker than long documents.

What remains to be addressed now is the question of how to incorporate document length, *dl*, into the model.

Document Length: Definitions and Assumptions

There are different ways of defining *dl*. We might define it by counting

- bytes, text lines, sentences, or paragraphs.
- characters, including or excluding spaces.
- unique terms including or excluding stopwords.
- all terms regardless of their nature.

For instance, let *l* be the number of unique terms in a document (including stopwords) and *m* be the number of *index terms*. If "negative" terms and stopwords are excluded from the inverted index, $l > m$. This is more than reasonable. Actually, in most IR models including the family of BM25 models, document length are customarily computed as

$$dl_j = \sum_i^m f_{i,j} \tag{4}$$

where

$$\sum_i^l f_{i,j} > \sum_i^m f_{i,j} \quad (5)$$

Regardless of how *dl* is defined, it is the result of writing styles. To account for this, Robertson & Walker (1994) introduced two important hypotheses:

- The Verbosity Hypothesis
- The Scope Hypothesis

The Verbosity and Scope Hypotheses

Consider any two documents written by different authors but equally relevant and elite to the same term(s).

The Verbosity Hypothesis states that some authors are more verbose than others, using more words to say the same thing, and thus that we can simply normalize term frequencies by document length without affecting eliteness and relevance. This is the Verbosity Hypothesis.

The Scope Hypothesis suggests the opposite: that we should not normalize document lengths because some authors have more to say about a given topic or topics than others. The extreme scenario would be an author writing like concatenating several documents into a single one (Robertson & Zaragoza, 2009).

Both hypotheses are ideal scenarios. In reality, a mixture of these scenarios is frequently present in document collections. We can assume that each hypothesis is a partial explanation and thus that a kind of soft or adjustable normalization is appropriate.

To insure that the definition used in Equation 5 is not critical, each document length could be normalized with the average document length,

$$dl_{ave} = \frac{\sum_j^{\mathrm{N}} dl_j}{\mathrm{N}} \quad (6)$$

The normalization can then be adjusted with a function of the form

$$B = 1 - b + b\left(\frac{dl_j}{dl_{ave}}\right) \tag{7}$$

The normalization function, *B*, is then used to normalize term frequencies,

$$f'_{i,j} = \frac{f_{i,j}}{B} \tag{8}$$

So the overall local weight component in (1) is

$$L_{i,j} = \left(\frac{f'_{i,j}}{k_1 + f'_{i,j}}\right) = \left(\frac{f_{i,j}}{k_1\left((1-b) + b\left(\frac{dl_j}{dl_{ave}}\right)\right) + f_{i,j}}\right) \tag{9}$$

where k_1 is an adjustable parameter.

A common modification consists in multiplying (9) by the scaling factor (k_1 + 1). Since this is the same for all terms, it does not affect the results so we can write

$$L_{i,j} = \left(\frac{f_{i,j}\,(k_1+1)}{k_1\left((1-b) + b\left(\frac{dl_j}{dl_{ave}}\right)\right) + f_{i,j}}\right) \tag{10}$$

and for the grand finale (1) becomes

$$w_{i,j} = L_{i,j}G_i = \left(\frac{f_{i,j}\,(k_1+1)}{k_1\left((1-b) + b\left(\frac{dl_j}{dl_{ave}}\right)\right) + f_{i,j}}\right)\mathrm{F4} \tag{11}$$

which is commonly referred to as the OKAPI BM25 Model.

Applying the $(k_1 + 1)$ scaling factor makes the final scores more compatible with RSJ weights. Thus for $f_{i,j} = 1$ and $b = 0$, a single occurrence of a term and no normalization, (11) reduces to an RSJ weight

$$w_{i,j} = \text{F4} \tag{12}$$

Table 1 lists a family of BM models that are obtained from BM25 by tuning the b and k_1 parameters (Robertson & Walker, 1994).

Table 1. Family of Best Match Models.

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

Without loss of generality, BM25 = BM15 = BM11 for documents of average lengths while these reduce to BM1 for $k_1 = 0$. This brings up the question of how b and k_1 affect (11). The BM25 model provides no guidance on how these parameters should be set.

Figure 2 depicts BM25 $L_{i,j}$ vs. $f_{i,j}$ curves for several dl_j/dl_{ave} ratios, b, and k_1 values. In (a) and (b), setting $k_1 = 1$ and changing b, from $b = 0.5$ (partial normalization) to $b = 1$ (zero

normalization), does not change the curves corresponding to $dl_j = dl_{ave}$. A similar result is observed between (c) and (d) for curves where $dl_j = dl_{ave}$.

Compare now (a) with (c). Setting $b = 0.5$ and changing k_1, from $k_1 = 1$ to $k_1 = 2$, does not change the curves corresponding to $dl_j = dl_{ave}$, except that the curves are now shifted downward.

Similarly in (b) and (d), setting $b = 0.8$ and changing k_1, from $k_1 = 1$ to $k_1 = 2$, does not changes the curves corresponding to $dl_j = dl_{ave}$, except that the curves are equally shifted down.

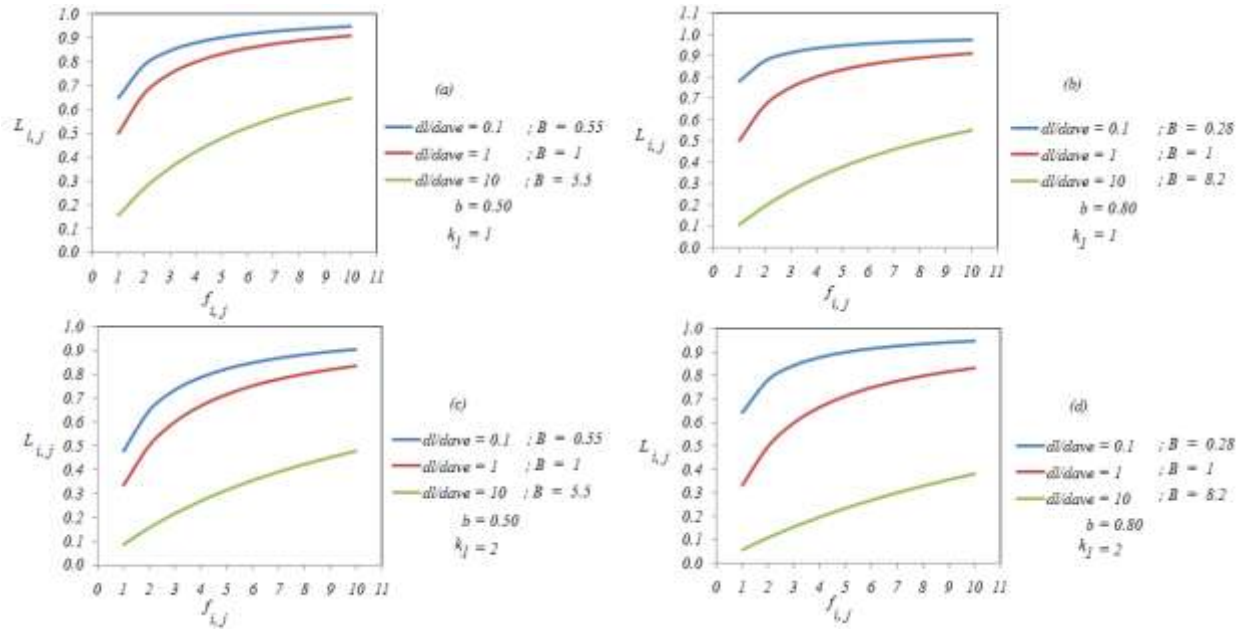

Figure 2. $L_{i,j}$ as a function of $f_{i,j}$ for some combinations of k_1 and b values.

Figure 2 shows that changing b and k_1 attenuates the normalization function B and improves the resolution of the curves, with the settings $0.5 < b < 0.8$ and $1.2 < k_1 < 2$ working fairly well. Other studies suggest $0.75 < k_1 < 2$ are equally good marks. It seems that acceptable results are obtained in most cases with

$$\frac{b}{k_1} < 1 \qquad (13)$$

Conclusion

Okapi BM25 is a model where local weights are computed as parameterized frequencies based on a 2-Poison model and global weights as RSJ weights.

Absent from the discussion was the scoring of query terms. It is not hard to realize that the query can be treated as another document, so we may treat within-query term frequency in a similar fashion to within-document term frequency (Robertson and Walker, 1994).

Also absent from the discussion was the question of scoring terms from structured documents with multiple fields. We hope to cover this in an upcoming tutorial on an advanced model: BM25F (Robertson & Zaragoza, 2004).

Exercises

1. Construct profile curves as in Figure 2 for $0.5 < b < 0.8$ and $0.5 < k_1 < 1$. Explain the relative shape of the curves and which of these setting conditions are acceptable.
2. Repeat exercise 1 using $b = 0$ and $1 < k_1 < 5$.

References

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

Macfarlane, A. (2001). Okapi-Pack. Centre For Interactive Systems Research. City University, London EC1V 0BH. Retrieved from http://www.staff.city.ac.uk/~andym/OKAPI-PACK/

Mitev, N. N., Venner, G. M., & Walker, S. (1985). Designing an Online Public Access Catalogue: Okapi, a Catalogue on a Local Area Network. Retrieved from http://sigir.org/files/museum/pub-28/pub-28-frontmatter.pdf

Robertson, S. E. (2004). Understanding Inverse Document Frequency: On theoretical arguments for IDF. Journal of Documentation, 60, 5, 503-520. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.438.2284&rep=rep1&type=pdf

Robertson, S. E., & Walker, S. (1994). Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In W. B. Croft and C. J. van Rijsbergen, editors, SIGIR'94: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 345-354. Springer-Verlag, 1994. Retrieved from https://pdfs.semanticscholar.org/c0a4/8ed7577a7b48288dfb2711cbd86e30636b5f.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., Walker, S., & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. Information Processing and Management 36, 95-108. https://pdfs.semanticscholar.org/25ca/cc430c97148f743aa0e75c75edf6fba4acbf.pdf

Robertson, S. E., Walker, S., Jones, Hancock-Beaulieu, M. M., & Gatford, M. (1996). Okapi at TREC 3. Retrieved from https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/okapi_trec3.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

SIGIR (2016). Museum. Retrieved from http://sigir.org/resources/museum/

Walker, S. (1987). OKAPI: Evaluating and Enhancing an Experimental Online Catalog. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/7503/librarytrendsv35i4j_opt.pdf?sequence=1

Walker, S., & Beaulieu, M. H. (1991). Okapi at City: An evaluation facility for interactive IR. Centre for Information Science City University, British Library Research and Development Report No. 6056, 1991. Retrieved from http://sigir.org/files/museum/pub-26/frontmatter.pdf