Verboseness Fission for BM25 Document Length Normalization

Aldo Lipani[1] Mihai Lupu[2] Allan Hanbury[2] Akiko Aizawa[1]

[1]National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, Japan
{surname}@nii.ac.jp

[2]Inst. of Software Technology & Interactive Systems
Vienna University of Technology
Vienna, Austria
{surname}@ifs.tuwien.ac.at

ABSTRACT

BM25 is probably the most well known term weighting model in Information Retrieval. It has, depending on the formula variant at hand, 2 or 3 parameters (k_1, b, and k_3). This paper addresses b—the document length normalization parameter. Based on the observation that the two cases previously discussed for length normalization (multi-topicality and verboseness) are actually three: multi-topicality, verboseness with word repetition (repetitiveness) and verboseness with synonyms, we propose and test a new length normalization method that removes the need for a b parameter in BM25. Testing the new method on a set of purposefully varied test collections, we observe that we can obtain results statistically indistinguishable from the optimal results, therefore removing the need for ground-truth based optimization.

1. INTRODUCTION

BM25 is the most longevous weighting schema in Information Retrieval (IR), still widely used in industry and studied in research. The peculiarity of this weighting schema is its probabilistic root that is based on the 2-Poisson model of term frequencies in documents [13]. In its classic version, a document d is scored by the function:

$$S(q,d) = \sum_{t \in T_q \cap T_d} \frac{(k_3+1)tf_q}{k_3 + tf_q} \frac{(k_1+1)\overline{tf_d}}{k_1 + \overline{tf_d}} \log \frac{|D|+0.5}{df_t + 0.5}$$

with

$$\overline{tf_d}(t) = \frac{tf_d(t)}{B} \quad B = (1-b) + b\frac{L_d}{avgdl}$$

where q is the query, D is the set of documents, $d \in D$ is a document, T_d and T_q are the sets of document terms and query terms, $\overline{tf_d}$ is the normalized term frequency of the term t within the document d, tf_q is the term frequency of the term t within the query, df_t is the document frequency of the term t, L_d the length of the document d, $avgdl$ the average document length over the collection D of documents, and k_1, b and k_3 the three parameters with domains $[0,\infty[$, $[0,1]$ and $[0,\infty[$.

This semi-parametric retrieval function [10] has 3 degrees of freedom, each with a specific meaning: k_1 and k_3 tune how fast the respective tf component saturates, expressing the importance of the presence of an additional occurrence of the term t in the document or query. The parameter b controls the normalization of the tf component, varying between the two extremes of non normalized, when $b = 0$, and fully normalized by the coefficient of variation of the document length, when $b = 1$.

The tuning of the three parameters is not an easy problem, nor a resource free task, due to the required development of a test collection. Hence, in most cases, the suggested values are used: $k_1 = 1.2$, $b = 0.75$ [14] and $k_3 = 8$ [12] . Still, as shown by Chowdhury et al. [2], tuning the parameters can lead to a considerable improvement in the effectiveness of the retrieval system. However, tuning is only possible if ground truth is available, and another, more analytical approach can be taken. This consists in trying to better understand the geometry of the information space, in order to extend and improve the current model.

In this paper, we focus on the term frequency normalization, reopening the discussion described by Robertson and Zaragoza [13] about the verboseness and scope hypotheses. We propose a new parameter-free normalization, based on the features of the document collection. We test this model using a sample of five test collections from TREC and CLEF, selected on purpose from different domains: Web, News, Medical, and Patent, in order to verify experimentally the dependency between the normalization factor and the features of the document collection.

The remainder of the paper is structured as follows: in Section 2 we provide a very brief summary of the extensive work already done on the study and understanding of the term frequency normalization. Section 3 provides the intuition of our method and introduces the required concepts and the method itself. In Section 4 we present and discuss our experimental results. We conclude in Section 5.

2. RELATED WORK

The initiators of the discussion about the term frequency normalization are the early participants in TREC, with first insights appearing after TREC-3, and the first efforts on document length normalization showing improved results in TREC-4 [3]. To understand why a document is long, Robertson and Zaragoza [13, p. 358] describe two hypotheses: a) verboseness, to convey the same information using

ICTIR'15, September 27 - 30, 2015, Northampton, MA, USA.

ACM 978-1-4503-3833-2/15/09 ...$15.00.
DOI: http://dx.doi.org/10.1145/2808194.2809486 .

more words than needed; and b) scope, to convey information containing more topics, details, or aspects. These hypotheses have a conflicting effect when treating the normalization in terms of length, because while the first suggests to normalize the *tf* by the length, the second suggests the opposite. Hence, the introduction of a soft normalization based on the coefficient of variation of the document length and the introduction of the b parameter that controls the slope of the normalization factor. This is of course not the only way for length normalization. Among others, Singhal et al. [17] studied it extensively for the TF-IDF model.

Not much work has been done on the scope hypothesis, except perhaps the effort spent in passage retrieval. Here, document length is circumvented by viewing the document as a collection of concatenated shorter documents to be retrieved individually.

More work has been done to tackle the verboseness issue. Na et al. [11] briefly introduce the concept of verboseness given by repetitiveness of terms. They compare it with multi-topicality under the language modeling framework. The normalization factors are corrected based on the assumption that the vocabulary size can be used to estimate the number of topics contained in the document. He and Ounis [5] introduced a new term frequency normalization following the idea of Amati [1], who introduced the use of Dirichlet Priors. He and Ounis point out the relationship between test collection features on term frequency normalization, and introduce a new parameter, learned from the test collection. They defined the normalization effect and hypothesized that the optimal parameter is the value that makes the normalization factor give similar normalization results across different corpora [4, 6]. Lv and Zhai pointed out that the retrieval pattern of BM25 does not follow the relevancy pattern, biasing the system against long documents, and introduced a boosting parameter δ that summed to the normalized term frequency in a first version [9] and then summed to the term frequency component in a second version [8] to correct the pattern discrepancy.

Rousseau and Varzirgiannis [15] analyze the problem in terms of function composition, comparing BM25 with TF-IDF and combining the two works previously mentioned, to gain a better understanding of the similarity across the models. Some efforts have been directed towards understanding and removing the parameters of BM25: Lv and Zhai [7] pointed out that it is more effective to use a term-specific k_1, and that it is possible to estimate it using an information gain measure to quantify the contributions of repeated term occurrences. They do not address b.

Overall, a criticism of all of these works is that the studies of and experiments with new models of the term frequency normalization always use the same kind of test collection, News and Web corpora.

3. THE NORMALIZATION HYPOTHESES

Document length normalization is based on the observation that documents are long either because they are verbose, or because they cover more aspects, as discussed.

The insight at the base of this study is that we can distinguish two kinds of verboseness: a) *repetitiveness*, in which the same terms are repeated many times (e.g. legal or patent texts); and b) *non-repetitiveness*, where the writer uses different terms to describe the same thing (e.g. over-descriptive narration in Balzac's novels). In the first case, *tf* is expected to be higher, so it should be normalized more than in the second case, where it is naturally low because of the use of different terms. While non-repetitiveness implies a more semantic analysis of the text, repetitiveness can be easily identified by counting the number of times terms are repeated on average. We define this in an obvious way as the average term frequency:

$$avgtf_d = \frac{1}{|T_d|}\sum_{t\in T_d} tf_d(t) = \frac{L_d}{|T_d|} \qquad (1)$$

However, we should observe that while a high $avgtf_d$ is indicative of repetitiveness, a low $avgtf_d$ would indicate either a broader document that would fall into the scope hypothesis, or a verbose, non-repetitive document. From the observation that *tf* is expected to be higher, we have the chance to better discern the sets of elite and non-elite documents described in the 2-Poisson model. Embedding this new knowledge in the model would take into account the fact that observing a high *tf* can be due either to its relevance in the document, as an elite term, or because of its repetitiveness, as boilerplate, non-elite term.

Our intuition is that it is possible to differentiate the two kinds of verboseness for a specific document based in part on collection statistics. We can then diminish the effect of the *tf* for each document by comparing the average *tf* of a document with *mavgtf*, the mean average term frequency of the collection:

$$mavgtf = \frac{1}{|D|}\sum_{d\in D} avgtf_d \qquad (2)$$

First, a few observations on these new indicators. The average term frequency of a document d ($avgtf_d$) is an indicator of how many times the same term is repeated in the document. If a document does not have any repetitions, $avgtf_d$ is equal to 1. The average term frequency is simply document length over number of unique terms, and this makes it easy to verify that *avgtf* has domain $[1, \infty[$, assuming documents of finite length $[1, maxL]$, where $maxL$ is the length of the longest document. Since *mavgtf* is the average of the *avgtf*, it has the same domain. If the test collection is made on average of documents with a low level of repetition, then the *mavgtf* is very close to 1. The *mavgtf* is a collection specific value that summarizes the repetitiveness of the language in a specific corpus.

From the intuition above we infer first that if the language of a collection is repetitive (high *mavgtf*), we expect to need more length normalization. Therefore we define the BM25 document length normalization factor b as:

$$b = 1 - mavgtf^{-1} \qquad (3)$$

This definition of b has the required domain $[0, 1[$ and increases monotonically with *mavgtf*.

With this normalization parameter, we can define a new normalization factor, B_{-b} to be used in BM25:

$$B_{-b} = mavgtf^{-1} + (1 - mavgtf^{-1})\frac{L_d}{avgdl} \qquad (4)$$

However, the reader may have already noticed a potential issue with this new b: while in theory it has domain $[0, 1[$, it will only reach 1 as $mavgtf \to \infty$. In practice the *mavgtf* is actually small, in the range of the low single digits, and this will limit the values of b to the lower half of the normalization spectrum. The normalization factor B_{-b}

Corpus	EC	Challenge	$\|D\|$	*mavgtf*
Aquaint	TREC	Hard 2005	1,033,461	1.519
Disks 4&5	TREC	Ad Hoc 8	528,155	1.574
eHealth'13	CLEF	eHealth 2013	1,102,848	2.205
.GOV	TREC	Web 2002	1,247,753	2.481
CLEF-IP'10	CLEF	CLEF-IP 2010	2,670,678	3.008

Table 1: Corpora used, with information about the challenge and evaluation campaign (EC) to which it belongs, number of documents, and mean average term frequency.

will therefore tend to be very conservative in its document length normalization. This is partially by design: as we said before – we do not want to do strong length normalization if the collection is not using repetitive language. Even more, at this point we are still not making a distinction between repetitiveness and non-repetitiveness at document level. To do so, and at the same time to control the $(1-b)$ component in B, we need to introduce another factor in B:

$$B_{\text{VA}} = (1-b)\frac{avgtf_d}{mavgtf} + b\frac{L_d}{avgdl} \qquad (5)$$

This new factor, $\frac{avgtf_d}{mavgtf}$, boosts the normalization factor B when the document at hand is repetitive.

The new formulation for B can also be seen as a reinterpretation of document length normalization: it is now no longer a linear combination between doing or not doing length normalization, but rather a linear combination between normalizing for repetitiveness or length (non-repetitiveness), controlled by a parameter b bound to the general repetitiveness of the language of the collection. For collections that are generally repetitive, it will tend to do length normalization, and the newly added factor will reduce this normalization only for those documents that are not repetitive. For collections that are generally non-repetitive, it will tend to not do length normalization, and the newly added factor will increase this normalization only for those documents that are repetitive. Intuitively, the method compares the repetitiveness of the document with that of the collection. The proposed variant makes the repetitiveness of the document no longer a good indicator of verboseness if the collection is generally repetitive.

Finally, using our b from Eq. 3, our variant of BM25 normalization factor is:

$$B_{\text{VA}} = mavgtf^{-2}\frac{L_d}{|T_d|} + (1-mavgtf^{-1})\frac{L_d}{avgdl} \qquad (6)$$

4. EXPERIMENTS

To test our predictions we selected five ad hoc test collections from TREC and CLEF, with the aim to observe differences in the use of language, in different domains. We selected from News, Web, Medical, and Patent corpora, listed in Table 1, where we can observe how the average term frequency varies across the corpora. To assess the different experiments, we used the condensed version [16] of mean average precision (MAP') and precision at 10 (P@10') because of their better stability in case of incomplete judgments. We tested the new normalization factors, B_{-b} and B_{VA}, against two different configurations of the classic BM25: standard and ideal. The BM25 standard is characterized by having the suggested configuration of the parameters, k_1 and b. In the ideal BM25, the two parameters have been optimized using as training set and test set the same set of topics, which of course makes it an unrealistic scenario, but an interesting upper limit. In this case, k_1 varies between 0.5 and 2.5. In all experiments we set $k_3 = 0$ to avoid any potential interferences of the tf_q in the scoring of the document.

Track	P.	k1	b	MAP'	P@10'
		Standard Case			
Hard 2005	CL	1.20	0.75	0.2144‡	0.3600‡
	CL-b	1.20	-	0.2325†	0.4360†
	VA	1.20	-	0.2318†	0.4360†
Ad Hoc 8	CL	1.20	0.75	0.2504‡	0.4720
	CL-b	1.20	-	0.2578	0.4600
	VA	1.20	-	0.2677†	0.4940
eHealth'14	CL	1.20	0.75	0.5565	0.7694
	CL-b	1.20	-	0.5636‡	0.7878‡
	VA	1.20	-	0.5718‡	0.7694
Web'02	CL	1.20	0.75	0.2022	0.2460
	CL-b	1.20	-	0.1972	0.2440
	VA	1.20	-	0.2010	0.2520
CLEF-IP'10	CL	1.20	0.75	0.3562‡	0.6423‡
	CL-b	1.20	-	0.3537‡	0.6371‡
	VA	1.20	-	0.3556‡	0.6371‡
		Ideal Case			
Hard 2005	CL	1.65	0.25	0.2346†	0.4440†
	CL-b	1.70	-	0.2335†	0.4360†
	VA	1.80	-	0.2332†	0.4140†‡
Ad Hoc 8	CL	0.45	0.40	0.2715†	0.4600
	CL-b	0.45	-	0.2713†	0.4520
	VA	0.55	-	0.2744†	0.4900
eHealth'14	CL	2.30	0.55	0.5849	0.8143
	CL-b	2.30	-	0.5844†	0.8143
	VA	2.40	-	0.5922	0.7959†
Web'02	CL	2.40	0.70	0.2062	0.2460
	CL-b	2.05	-	0.2012	0.2420
	VA	2.50	-	0.2056	0.2360
CLEF-IP'10	CL	2.50	1.0	0.3713†	0.6540†
	CL-b	2.50	-	0.3615	0.6536
	VA	2.25	-	0.3643	0.6567

Table 2: Scores obtained with the classic BM25 (CL), classic BM25 with *b* as in Eq. 3 (CL-b), and our variant (VA). † indicates statistical significance (t-test, p<0.05) against the standard classic BM25 (CL) and ‡ against the ideal classic BM25 (CL).

We used the search engine Terrier[1] 4.0 for the classic BM25 and developed and integrated in it our BM25 variants[2]. All the documents have been preprocessed using the English tokenizer and Porter stemmer of the Terrier search engine. The queries are extracted from the title only, except for the CLEF-IP 2010 where the abstract has been included.

Table 2 shows the performance of each weighting scheme in the two configurations mentioned above. In only two of the five collections (eHealth and CLEF-IP) the standard VA is lower and statistically significantly different from the ideal classic BM25 (CL). This can be explained by the combination of two effects: the large influence k_1 has on the results, as shown in Fig. 1 by the size of the gray areas, and the large difference between the standard k_1 and the ideal k_1.

[1] `http://www.terrier.org`

[2] Code available on the website of the first author

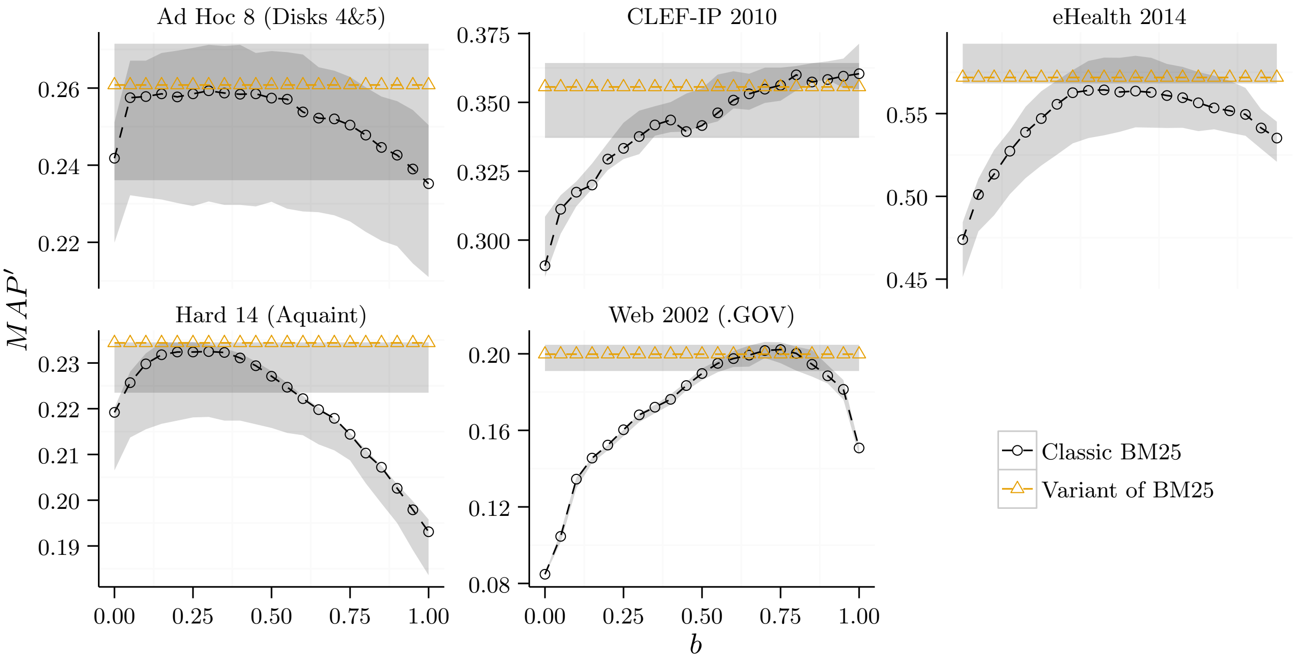

Figure 1: Performance sensitivity of the classic BM25 (CL) and our variant (VA) across all the test collections, with standard $k_1 = 1.2$. The gray area represents the range of values obtainable varying k_1 in the range $[0.5, 2.5]$.

5. CONCLUSION

We continued a discussion started 20 years ago in the context of TREC about the need for document length normalization and the nature of the document length itself. Previous studies, working on test collections of web and news corpora, failed to observe what in legal and patent collections is patently obvious: document length verbosity, in a bag-of-words model, can be expressed via repetition or via synonyms. We proposed a new factor B, including a specific value for the parameter b, and showed that, across different domains, the results are generally statistically indistinguishable from those obtained with ideal b values, without having to identify these ideal values. Together with previous works on estimation of the k_1 parameter, this brings us a step closer to a parameter-free, stable, BM25.

6. REFERENCES

[1] G. Amati and J. C. C. Van Rijsbergen. Probabilistic models for information retrieval based on divergence from randomness. *TOIS*, 20(4), 2002.

[2] A. Chowdhury, M. C. McCabe, D. Grossman, and O. Frieder. Document Normalization Revisited. In *Proc. of SIGIR*, 2002.

[3] D. Harman. Overview of the Fourth Text REtrieval Conference (TREC-4). In *Proc. of TREC 4*, 1995.

[4] B. He and I. Ounis. A Study of Parameter Tuning for Term Frequency Normalization. In *Proc. of CIKM*, 2003.

[5] B. He and I. Ounis. A Study of the Dirichlet Priors for Term Frequency Normalisation. In *Proc. of SIGIR*, 2005.

[6] B. He and I. Ounis. Term Frequency Normalisation Tuning for BM25 and DFR Models. In *Proc. of ECIR*, 2005.

[7] Y. Lv and C. Zhai. Adaptive Term Frequency Normalization for BM25. In *Proc. of CIKM*, 2011.

[8] Y. Lv and C. Zhai. Lower-bounding Term Frequency Normalization. In *Proc. of CIKM*, 2011.

[9] Y. Lv and C. Zhai. When Documents Are Very Long, BM25 Fails! In *Proc. of SIGIR*, 2011.

[10] D. Metzler and H. Zaragoza. Semi-parametric and non-parametric term weighting for information retrieval. In *Proc. of ICTIR*, 2009.

[11] S.-H. Na, I.-S. Kang, and J.-H. Lee. Improving term frequency normalization for multi-topical documents and application to language modeling approaches. In *Proc. of ECIR*, 2008.

[12] S. Robertson, S. Walker, M. Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-4. In *Proc. of TREC 4*, 1995.

[13] S. Robertson and H. Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 2009.

[14] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proc. of TREC-3*, 1994.

[15] F. Rousseau and M. Vazirgiannis. Composition of TF Normalizations: New Insights on Scoring Functions for Ad Hoc IR. In *Proc. of SIGIR*, 2013.

[16] T. Sakai. Alternatives to Bpref. In *Proc. of SIGIR*, 2007.

[17] A. Singhal, C. Buckley, and M. Mitra. Pivoted Document Length Normalization. In *Proc. of SIGIR*, 1996.