Verboseness Fission for BM25 Document Length Normalization

Aldo Lipani\textsuperscript{1} \hspace{0.5cm} Mihai Lupu\textsuperscript{2} \hspace{0.5cm} Allan Hanbury\textsuperscript{2} \hspace{0.5cm} Akiko Aizawa\textsuperscript{1}

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

\textsuperscript{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 \((k_1, b, k_3)\). 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 purposely 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:

\[
S(q, d) = \sum_{t \in \mathcal{T}_d \cap \mathcal{T}_q} \frac{(k_1 + 1) \cdot tf_t \cdot \log \frac{|D| + 0.5}{df_t + 0.5}}{k_3 + tf_t} \cdot \frac{(k_1 + 1) \cdot \log df_t}{k_3 + df_t} \cdot \log \frac{|D|-|D_t^d|}{|D|-|D_q|}
\]

where \(q\) is the query, \(D\) is the set of documents, \(d \in D\) is a document, \(\mathcal{T}_d\) and \(\mathcal{T}_q\) are the sets of document terms and query terms, \(T_f^d\) is the normalized term frequency of the term \(t\) within the document \(d\), \(T_q\) is the term frequency of the term \(t\) within the query, \(df_t\) is the document frequency of the term \(t\), \(L_d\) is the length of the document \(d\), \(avgdl\) is 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. \([12]\), 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 \([5]\). To understand why a document is long, Robertson and Zaragoza \([13]\) p. 358 describe two hypotheses: a) verboseness, to convey the same information using...
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 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 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 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 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 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.
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.

| Corpus       | EC      | Challenge | \(|D|\) | \(mavgtf\) |
|--------------|---------|-----------|------|-----------|
| Aquaint      | TREC    | Hard05    | 1,034,361 | 1.571 |
| 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 2: Scores obtained with the classic BM25 (CL), classic BM25 with \(b\) as in Eq. 5 (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).

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 [10] 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, \(B_{\text{VA}}\) and \(B_{\text{VA}}\), 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 \(t_f^i\) in the scoring of the document.

We used the search engine Terrier [4] for the classic BM25 and developed and integrated in it our BM25 variants. 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 of \(k_1\) 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 \(k_1\).
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