A New Term Frequency Normalization Model for Probabilistic Information Retrieval

Fanghong Jian, Jimmy Xiangji Huang, Jiashu Zhao and Tingting He *
Information Retrieval and Knowledge Management Research Lab
[1]National Engineering Research Center for E-Learning, [3]School of Computer, Central China Normal University, Wuhan, China; [2]School of Information Technology, York University, Toronto, Canada
jfhrecoba@mails.ccnu.edu.cn,jhuang@yorku.ca,zhaojiashu@gmail.com,tthe@mail.ccnu.edu.cn

ABSTRACT

In probabilistic BM25, term frequency normalization is one of the key components. It is often controlled by parameters k_1 and b, which need to be optimized for each given data set. In this paper, we assume and show empirically that term frequency normalization should be specific with query length in order to optimize retrieval performance. Following this intuition, we first propose a new term frequency normalization with query length for probabilistic information retrieval, namely BM25_{QL}. Then BM25_{QL} is incorporated into the state-of-the-art models CRTER_2 and LDA-BM25, denoted as $\text{CRTER}_2^{\text{QL}}$ and $\text{LDA-BM25}_{\text{QL}}$ respectively. A series of experiments show that our proposed approaches BM25_{QL}, $\text{CRTER}_2^{\text{QL}}$ and $\text{LDA-BM25}_{\text{QL}}$ are comparable to BM25, CRTER_2 and LDA-BM25 with the optimal b setting in terms of MAP on all the data sets.

KEYWORDS

Term Frequency Normalization, BM25, Probabilistic Model

ACM Reference Format:
Fanghong Jian, Jimmy Xiangji Huang, Jiashu Zhao and Tingting He . 2018. A New Term Frequency Normalization Model for Probabilistic Information Retrieval. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210147

1 INTRODUCTION AND RELATED WORK

Term frequency (TF) normalization is very important in information retrieval (IR) models. There are kinds of term frequency normalization achieving success. Sub-linear term frequency normalization in BM25 [10] is one of state-of-the-art approaches in the last two decades. It has two hyper-parameters (k_1 and b), which are as term independent constants and often need to be optimized for each given data set [4]. In recent years, much research work started to focus on the automatic tuning of document length normalization. TF normalization approaches in [4, 8, 9, 14] are document and collection dependent, and fixed term-independent parameter setting remained the same as the original BM25. Cummins et al. [1] first investigated the effect of query length on normalization but didn't measure the effect. [2, 6, 7] used query length normalization constraints to estimate term-specific parameters, which may be expensive and overfitted. Chung et al. [15] have incorporated the query-length into vector space model and conducted experiments on Chinese and English corpora, suggesting that the query-length should be incorporated in other existing ranking functions. So it is worth studying how to simply and effectively incorporate query length into probabilistic model.

In this paper, we propose a new term frequency normalization for probabilistic BM25, and integrate it into state-of-the-art BM25-based models with proximity and topic modeling. We also present experiments on TREC data sets to investigate the effect of three term frequency normalization functions.

The remainder of this paper is organized as follows. We propose a modified BM25 via a new term frequency normalization method in Section 2. In Section 3, we set up our experimental environment on eight TREC data sets. In Section 4, the experimental results are presented and discussed. Finally, we conclude our work briefly and present future research directions in Section 5.

2 OUR PROPOSED APPROACH

In this section, we first introduce a new term frequency normalization approach, and then describe how to integrate it into probabilistic BM25. For clarification, Table 1 outlines the notations used throughout the paper.

Table 1: Notations

Notations	Description
c	collection
d	document
q	query
q_i	query term
ql	query length
$\frac{\partial b_{QL}}{\partial ql}$	first order partial derivative ∂b_{QL} with respect to ql
$\frac{\partial^2 b_{QL}}{\partial ql^2}$	second order partial derivative ∂b_{QL} with respect to ql
dl	length of document
$avdl$	average document length
N	number of indexed documents in collection
n	number of indexed documents containing a term
tf	within-document term frequency
qtf	within-query term frequency
IDF	inverse document frequency, equals to $\log_2 \frac{N-n+0.5}{n+0.5}$
b, k_1, k_3	parameters in BM25

2.1 A New Method for TF Normalization

BM25 is a well-known probabilistic IR model, which scores a document d with respect to a query q as follows.

$$BM25(q, d) = \sum_{q_i \in q \cap d} \frac{(k_1+1) \cdot TF}{k_1 + TF} \cdot \frac{(k_3+1) \cdot qtf}{k_3 + qtf} \cdot IDF \quad (1)$$

where $TF = \frac{tf}{(1-b)+b \cdot \frac{dl}{avdl}}$ is pivoted document length normalization, which is proved to be effective for term frequency normalization. b is a parameter used to balance the impact of document

*The corresponding author is Jimmy Xiangji Huang.
The affiliation 1 is for Fanghong Jian, 2 for Jiashu Zhao and 3 for Tingting He.

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210147

length dl. In practice, b is usually set to a default value or optimized for each individual data set. Generally, parameter b should be optimized for each given collection [4], so it is worth exploring a modified term frequency normalization.

In previous work [11], the query length, i.e. ql, the number of terms in a query q, is used to balance two kinds of TF normalization. From an information theoretic perspective, adding a term to the query is equivalent to increasing the information provided by the query. We assume that when query length increases, the effect of TF Normalization should be boosted, in order to facilitate preference to shorter documents. Based on this assumption, we propose a new method for document TF normalization using query length.

$$TF_{QL} = \frac{tf}{(1 - b_{QL}(ql)) + b_{QL}(ql) \cdot \frac{dl}{avdl}} \tag{2}$$

where $b_{QL}(ql)$ is a given function of query length ql. Heuristically, this function $b_{QL}(ql)$ should increase with the growth of query length, while it must lie between 0 and 1. In addition, when a term is added to a shorter query, it is more likely to show more search intent than added to a longer query. Thus, $b_{QL}(ql)$ should be less affected with the change of ql for larger ql. Specifically, we characterize $b_{QL}(ql)$ as follows.

- Boundedness: $b_{QL}(1) = 0$, and $b_{QL}(\infty) = 1$
- Monotonicity: $b_{QL}(ql) < b_{QL}(ql + 1) \Rightarrow \frac{\partial b_{QL}}{\partial ql} > 0$
- Convexity: $b_{QL}(ql + 1) - b_{QL}(ql) > b_{QL}(ql + 2) - b_{QL}(ql + 1) \Rightarrow \frac{\partial^2 b_{QL}}{\partial ql^2} < 0$

To satisfy the above characteristics, we propose several different types of functions as in Formula (3)-(5). These three functions are proposed to satisfy all the required characteristics for $b_{QL}(ql)$. In addition, the proposed functions grow differently when the query length ql increases: $b_{QL}^{LOG}(ql)$ is based on the logarithm function which grows the slowest; $b_{QL}^{REC}(ql)$ is based on the reciprocal function which grows with a median speed; $b_{QL}^{EXP}(ql)$ is based on the exponential function which grows the fastest. In this paper, we only consider these three types of functions and more functions will be evaluated in the future.

$$b_{QL}^{LOG}(ql) = 1 - \frac{2}{1 + \log_2(1 + ql)} \tag{3}$$

$$b_{QL}^{REC}(ql) = 1 - \frac{4}{3 + ql} \tag{4}$$

$$b_{QL}^{EXP}(ql) = 1 - \exp(-\frac{ql - 1}{6}) \tag{5}$$

2.2 A New Model: BM25$_{QL}$

We use the query length for term frequency normalization in BM25 and propose a new BM25$_{QL}$ formula as follows.

$$BM25_{QL}(q, d) = \sum_{q_i \in q \cap d} \frac{(k_1+1) \cdot TF_{QL}}{k_1 + TF_{QL}} \cdot \frac{(k_3+1) \cdot qtf}{k_3 + qtf} \cdot IDF \tag{6}$$

In this paper, we explore three term frequency normalization functions in BM25, and the corresponding BM25$_{QL}$ are denoted as BM25$_{QL}^{LOG}$, BM25$_{QL}^{REC}$ and BM25$_{QL}^{EXP}$ respectively.

Recent years, there are some state-of-the-art BM25-based models succeeded in IR. For example, bigram cross term model CRTER$_2$ in [5] is a well known probabilistic proximity model, and LDA-BM25 in [3] is a strong topic based hybrid model. We use BM25$_{QL}$ in the same way as the BM25 in CRTER$_2$ and LDA-BM25, and propose CRTER$_2^{QL}$ and LDA-BM25$_{QL}$ respectively. Similarly, we also investigate term frequency normalization functions in CRTER$_2^{QL}$ and LDA-BM25$_{QL}$.

3 EXPERIMENTAL SETTINGS

We conduct experiments on eight standard TREC data sets, which include AP88-89 with queries 51-100, LA with queries 301-400, WSJ(87-92) with queries 151-200, DISK1&2 with queries 51-200, DISK4&5 no CR with queries 301-450, Robust04 with queries 301-450 & 601-700, WT2G with queries 401-450 and WT10G with queries 451-550. These data sets are different in sizes and genres, including high-quality newswire collections and Web collections containing many noisy documents. In all the experiments, we only use the title field of the TREC queries for retrieval. Queries without judgments are removed. For all test data sets used, each term is stemmed by using Porter's English stemmer. Standard English stopwords are removed. The official TREC evaluation measure is used in our experiments, namely Mean Average Precision (MAP).

For fair comparisons, we use the following parameter settings for both the baselines and our proposed models, which are popular in the IR domain for building strong baselines. First, in BM25, k_1 and k_3 are set to be 1.2 and 8. Meanwhile, we sweep the values of b for BM25 from 0 to 1.0 with an interval of 0.05. Second, in CRTER$_2$, we sweep the values of normalization parameter σ in a group of different values 2, 5, 10, 20, 25, 50, 75, 100, and triangle kernel was shown in [5] to achieve best MAP for most data sets. Thirdly, in LDA modeling, we use symmetric Dirichlet priors with $\alpha = 50/K_t$ and $\beta = 0.01$, which are common settings in the literature and shown in [3, 16] that retrieval results were not very sensitive to the values of these parameters. The number of topics K_t is set to be 400 as recommended in [3, 16]. Finally, we sweep the values of balancing parameter λ from 0.1 to 0.9 with an interval of 0.1 in CRTER$_2$, CRTER$_2^{QL}$, LDA-BM25 and LDA-BM25$_{QL}$.

4 EXPERIMENTAL RESULTS

4.1 Comparison with BM25

We first investigate the performance of our proposed BM25$_{QL}$ compared with BM25. The experimental results are presented in Figure 1. As shown by the results, our proposed BM25$_{QL}$ models are comparable to BM25 with optimal b on almost all data sets in terms of MAP. Moreover, according to the results in Figure 1, each new term frequency normalization function has its advantage on some aspects. There is no single function can outperform others on all the data sets. Without much knowledge of a new data set, logarithmic function is recommended for BM25$_{QL}$.

4.2 Comparison with CRTER$_2$

In order to test the robustness, we incorporate our proposed BM25$_{QL}$ models into various types of BM25-based models. Firstly, we use BM25$_{QL}$ to tune the parameter b in the state-of-the-art BM25-based proximity approaches. Zhao et al. [5] showed that bigram cross term model CRTER$_2$ is at least comparable to major probabilistic proximity models PPM [12] and BM25TP [13] in BM25-based framework. We compare our proposed CRTER$_2^{QL}$ with CRTER$_2$. The results are presented in Figure 2. Figure 2 shows that the proposed CRTER$_2^{QL}$ models are also comparable to CRTER$_2$ with optimal b on almost all data sets.

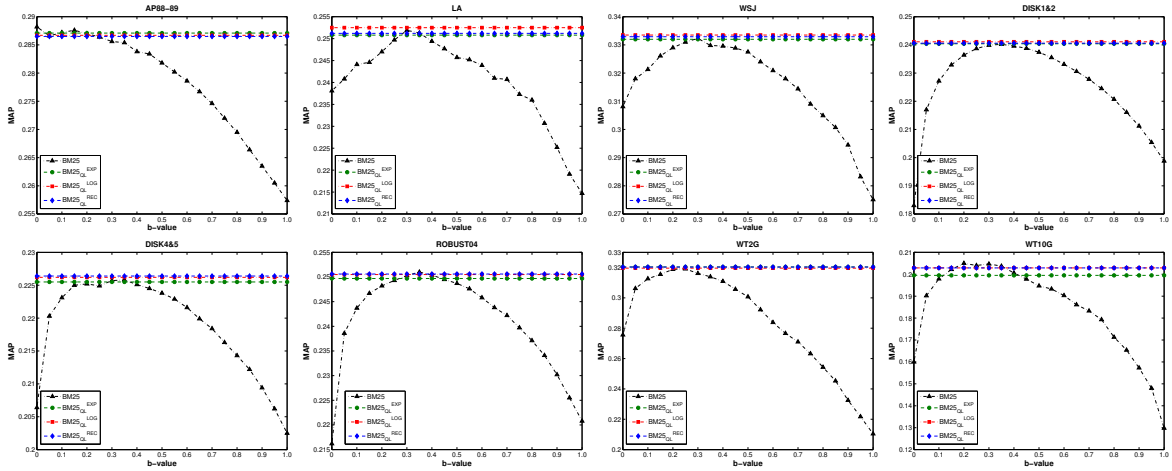

Figure 1: MAP Comparison between $\texttt{BM25}_{\texttt{QL}}$ and BM25

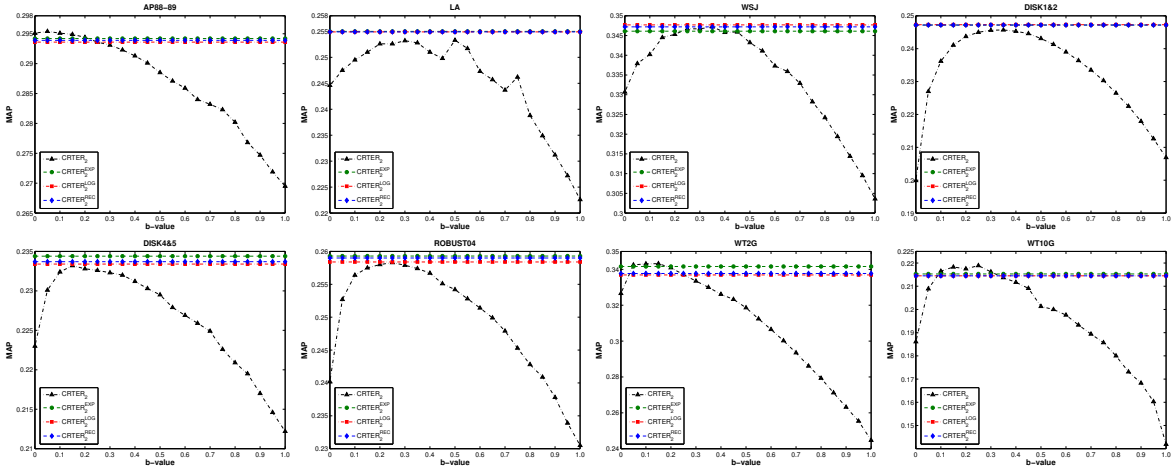

Figure 2: MAP Comparison between $\texttt{CRTER}_2^{\texttt{QL}}$ and CRTER_2

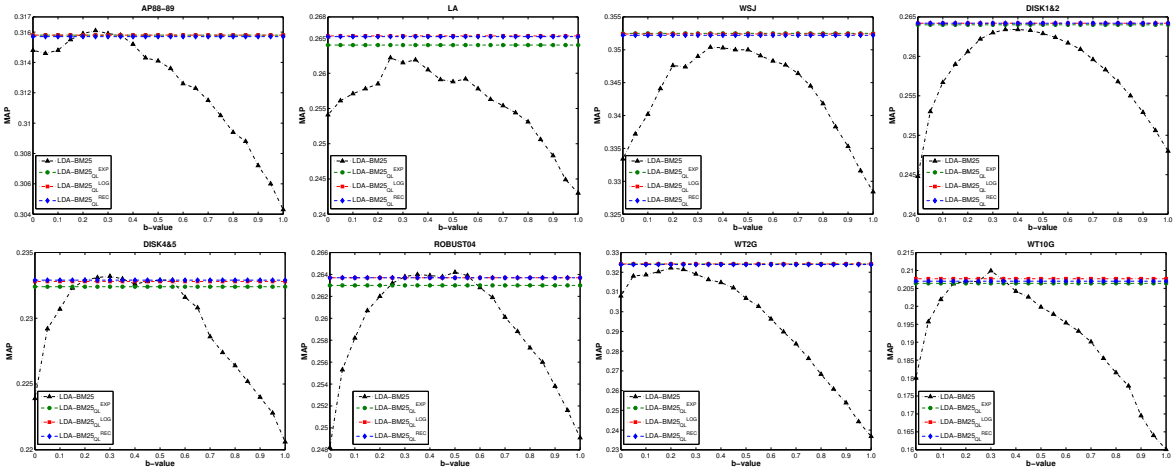

Figure 3: MAP Comparison between $\texttt{LDA-BM25}_{\texttt{QL}}$ and LDA-BM25

Table 2: Summary of Comparison with BM25_{QL} and BM25, $\text{CRTER}_2{}^{\text{QL}}$ and CRTER_2, $\text{LDA-BM25}_{\text{QL}}$ and LDA-BM25. The bold phase style means that it is the best result in each group. "$1, 2, 3, 4$" denotes our proposed models outperform the corresponding models with the settings for b as 0.35, 0.4, 0.75 and optimal respectively.

	AP88-89	LA	WSJ	DISK1&2	DISK4&5	ROBUST04	WT2G	WT10G
BM25-b=0.35	0.2854	0.2513	0.3298	0.2402	0.2258	**0.2510**	0.3139	0.2037
BM25-b=0.4	0.2838	0.2494	0.3296	0.2396	0.2251	0.2504	0.3109	0.2006
BM25-b=0.75	0.2720	0.2373	0.3090	0.2245	0.2163	0.2397	0.2632	0.1793
BM25-optimal b	**0.2882**	0.2519	0.3323	0.2402	0.2258	**0.2510**	0.3191	**0.2050**
$\text{BM25}_{\text{QL}}{}^{\text{EXP}}$	0.2871^{123}	0.2508^{23}	0.3320^{123}	0.2404^{1234}	0.2255^{23}	0.2497^{3}	$\mathbf{0.3204}^{1234}$	0.1995^{3}
$\text{BM25}_{\text{QL}}{}^{\text{LOG}}$	0.2867^{123}	$\mathbf{0.2525}^{1234}$	$\mathbf{0.3335}^{1234}$	$\mathbf{0.2411}^{1234}$	0.2262^{1234}	0.2505^{23}	0.3196^{1234}	0.2029^{23}
$\text{BM25}_{\text{QL}}{}^{\text{REC}}$	0.2865^{123}	0.2512^{23}	0.3329^{1234}	0.2406^{1234}	$\mathbf{0.2264}^{1234}$	0.2506^{23}	0.3203^{1234}	0.2029^{23}
CRTER_2-b=0.35	0.2923	0.2528	0.3472	0.2457	0.2320	0.2574	0.3300	0.2137
CRTER_2-b=0.4	0.2913	0.2510	0.3458	0.2453	0.2312	0.2567	0.3261	0.2117
CRTER_2-b=0.75	0.2823	0.2462	0.3282	0.2303	0.2226	0.2453	0.2861	0.1857
CRTER_2-optimal b	**0.2954**	0.2533	0.3472	0.2457	0.2332	0.2583	**0.3432**	**0.2189**
$\text{CRTER2}^{\text{EXP}}$	0.2942^{123}	$\mathbf{0.2550}^{1234}$	0.3461^{23}	0.2472^{1234}	$\mathbf{0.2344}^{1234}$	$\mathbf{0.2593}^{1234}$	0.3416^{123}	0.2153^{123}
$\text{CRTER2}^{\text{LOG}}$	0.2936^{123}	0.2549^{1234}	$\mathbf{0.3477}^{1234}$	$\mathbf{0.2473}^{1234}$	0.2334^{1234}	0.2584^{1234}	0.3369^{123}	0.2145^{123}
$\text{CRTER2}^{\text{REC}}$	0.2939^{123}	0.2549^{1234}	0.3472^{23}	0.2472^{1234}	0.2337^{1234}	0.2590^{1234}	0.3377^{123}	0.2145^{123}
LDA-BM25-b=0.35	0.3158	0.2619	0.3504	0.2634	0.2330	0.2640	0.3163	0.2074
LDA-BM25-b=0.4	0.3152	0.2605	0.3503	0.2634	0.2326	0.2639	0.3148	0.2042
LDA-BM25-b=0.75	0.3105	0.2544	0.3445	0.2583	0.2274	0.2588	0.2763	0.1855
LDA-BM25-optimal b	**0.3161**	0.2622	0.3504	0.2634	**0.2332**	**0.2642**	0.3222	**0.2099**
$\text{LDA-BM25}_{\text{QL}}{}^{\text{EXP}}$	0.3158^{23}	0.2640^{1234}	$\mathbf{0.3525}^{1234}$	0.2640^{1234}	0.2324^{3}	0.2630^{3}	$\mathbf{0.3242}^{1234}$	0.2064^{23}
$\text{LDA-BM25}_{\text{QL}}{}^{\text{LOG}}$	0.3158^{23}	$\mathbf{0.2653}^{1234}$	0.3524^{1234}	$\mathbf{0.2642}^{1234}$	0.2328^{23}	0.2637^{3}	$\mathbf{0.3242}^{1234}$	0.2077^{123}
$\text{LDA-BM25}_{\text{QL}}{}^{\text{REC}}$	0.3157^{23}	0.2652^{1234}	0.3522^{1234}	$\mathbf{0.2642}^{1234}$	0.2329^{23}	0.2637^{3}	0.3239^{1234}	0.2070^{23}

4.3 Comparison with LDA-BM25

Finally, we further incorporate our proposed BM25_{QL} models into state-of-the-art BM25-based model with topic modeling. Jian et al. [3] showed that LDA-BM25 is at least comparable to the state-of-the-art model CRTER_2. The performance of our proposed $\text{LDA-BM25}_{\text{QL}}$ and LDA-BM25 is presented in Figure 3. From Figure 3, we can find that $\text{LDA-BM25}_{\text{QL}}$ models are also comparable to LDA-BM25 with optimal b in MAP on almost all data sets. The performance is even better than searching the parameter space on several data sets, such as LA, WSJ, DISK1&2 and WT2G.

4.4 Analysis and Discussion

The experimental results show that our proposed models have consistent good performance in all scenarios on all data sets. In some occasions, the performance is even better than the heuristic best b-value. This is because that the new variable b_{QL} is self-adjusted for each query, while given the heuristic b-value is tested for all queries on an entire collection. b_{QL} is more adaptive, especially in real applications when the queries are quite different from each other. The functions proposed in Formula (3)-(5) perform similarly in terms of MAP. Although more functions could be considered to define the b_{QL}, most of the functions grow faster than the logarithm function and slower compared with the exponential function. According to the experimental results, we can see that the retrieval performance can be guaranteed using any of the proposed functions in Formula (3)-(5).

5 CONCLUSIONS AND FUTURE WORK

In this paper, we propose a new term frequency normalization model BM25_{QL} for probabilistic IR. Specifically, we present three term frequency normalization functions: logarithmic function, reciprocal function and exponential function. We also incorporate BM25_{QL} into two state-of-the-art BM25-based models CRTER_2 and LDA-BM25. Experimental results on eight standard TREC data sets show that BM25_{QL}, $\text{CRTER}_2{}^{\text{QL}}$ and $\text{LDA-BM25}_{\text{QL}}$ at least comparable to and sometimes even better than BM25, CRTER_2 and LDA-BM25 with the optimal b in terms of MAP.

In the future, we will conduct experiments on more large data sets with different types, such as GOV2 and ClueWeb09. There are also several interesting future research directions for us to explore. First, it is interesting to conduct an in-depth study on complete new term frequency normalization without hyper-parameters k_1 and b. Second, we will investigate the optimal term frequency normalization function. Third, we also plan to evaluate our models on more data sets including some real data sets and apply our models into real world applications.

ACKNOWLEDGMENTS

This research is supported by a Discovery grant from the Natural Sciences & Engineering Research Council (NSERC) of Canada, an Ontario Research Foundation award and also supported by the National Natural Science Foundation of China under grants number 61572223. We thank anonymous reviewers for their thorough comments, and greatly appreciate Dr. Xinhui Tu's help and support.

REFERENCES

[1] R. Cummins and C. O'Riordan. 2009. The Effect of Query Length on Normalisation in Information Retrieval. In *Proc. of the 2009 AICS*. 26–32.
[2] R. Cummins and C. O'Riordan. 2012. A Constraint to Automatically Regulate Document-length Normalisation. In *Proc. of the 21st ACM CIKM*. 2443–2446.
[3] F. Jian, J. X. Huang, J. Zhao, T. He and P. Hu. 2016. A Simple Enhancement for Ad-hoc Information Retrieval via Topic Modelling. In *Proc. of the 39th ACM SIGIR*. 733–736.
[4] B. He and I. Ounis. 2007. On Setting the Hyper-parameters of Term Frequency Normalization for Information Retrieval. *ACM TOIS* 25, 3 (2007), 13.
[5] J. X. Huang J. Zhao and B. He. 2011. CRTER: Using Cross Terms to Enhance Probabilistic IR. In *Proc. of the 34th ACM SIGIR*. 155–164.
[6] Y. Lv. 2015. A Study of Query Length Heuristics in Information Retrieval. In *Proc. of the 24th ACM CIKM*. 1747–1750.
[7] Y. Lv and C. Zhai. 2011. Adaptive Term Frequency Normalization for BM25. In *Proc. of the 20th ACM CIKM*. 1985–1988.
[8] Y. Lv and C. Zhai. 2011. Lower-bounding Term Frequency Normalization. In *Proc. of the 20th ACM CIKM*. 7–16.
[9] Y. Lv and C. Zhai. 2011. When Documents Are Very Long, BM25 Fails!. In *Proc. of the 34th ACM SIGIR*. 1103–1104.
[10] X. Huang S. Robertson S. Walker M. Beaulieu, M. Gatford and P. Williams. 1996. Okapi at TREC-5. In *Proc. of the 5th TREC*. 143–166.
[11] Jiaul H. Paik. 2013. A Novel TF-IDF Weighting Scheme for Effective Ranking. In *Proc. of the 36th ACM SIGIR*. 343–352.
[12] J.R. Wen R. Song, L. Yu and W.H. Hon. 2011. A Proximity Probabilistic Model for Information Retrieval. *Tech. Rep., Microsoft Research* (2011).
[13] C. Clarke S. Buttcher and B. Lushman. 2006. Term Proximity Scoring for Ad-hoc Retrieval on Very Large Text Collections. In *Proc. of the 29th ACM SIGIR*. 621 – 622.
[14] H. Zaragoza S. Robertson and M. Taylor. 2004. Simple BM25 Extension to Multiple Weighted Fields. In *Proc. of the 13th ACM CIKM*. 42–49.
[15] K.F. Wong K.L. Kwok T.L. Chung, R.W.P. Luk and D.L. Lee. 2006. Adapting Pivoted Document-length Normalization for Query Size: Experiments in Chinese and English. *ACM TALIP* 5, 3 (2006), 245–263.
[16] X. Wei and W. B. Croft. 2006. LDA-Based Document Models for Ad-hoc Retrieval. In *Proc. of the 29th ACM SIGIR*. 178–185.