Improving Retrieval Performance for Verbose Queries via Axiomatic Analysis of Term Discrimination Heuristic

Mozhdeh Ariannezhad
School of Eelectrical and Computer Engineering
College of Engineering, University of Tehran
m.ariannezhad@ut.ac.ir

Ali Montazeralghaem
School of Eelectrical and Computer Engineering
College of Engineering, University of Tehran
ali.montazer@ut.ac.ir

Hamed Zamani
Center for Intelligent Information Retrieval
University of Massachusetts Amherst
zamani@cs.umass.edu

Azadeh Shakery
School of Eelectrical and Computer Engineering
College of Engineering, University of Tehran
School of Computer Science
Institute for Research in Fundamental Sciences (IPM)
shakery@ut.ac.ir

ABSTRACT

Number of terms in a query is a query-specific constant that is typically ignored in retrieval functions. However, previous studies have shown that the performance of retrieval models varies for different query lengths, and it usually degrades when query length increases. A possible reason for this issue can be the extraneous terms in longer queries that makes it a challenge for the retrieval models to distinguish between the key and complementary concepts of the query. As a signal to understand the importance of a term, inverse document frequency (IDF) can be used to discriminate query terms. In this paper, we propose a constraint to model the interaction between query length and IDF. Our theoretical analysis shows that current state-of-the-art retrieval models, such as BM25, do not satisfy the proposed constraint. We further analyze the BM25 model and suggest a modification to adapt BM25 so that it adheres to the new constraint. Our experiments on three TREC collections demonstrate that the proposed modification outperforms the baselines, especially for verbose queries.

KEYWORDS

Verbose queries, query length, axiomatic analysis, theoretical analysis, term discrimination

ACM Reference format:
Mozhdeh Ariannezhad, Ali Montazeralghaem, Hamed Zamani, and Azadeh Shakery. 2017. Improving Retrieval Performance for Verbose Queries via Axiomatic Analysis of Term Discrimination Heuristic. In *Proceedings of SIGIR'17, August 7-11, 2017, Shinjuku, Tokyo, Japan, ,* 4 pages.
DOI: 10.1145/3077136.3080761

1 INTRODUCTION

Modern retrieval models use different query-based, document-based, and corpus-based properties to compute the relevance score of a document with respect to a query. Term frequency (TF), inverse document frequency (IDF), and document length are the main factors that are typically present in a retrieval function. When optimizing a retrieval model, the main focus is generally on the document-side of the function that weighs the query terms with respect to a document. The query-side part of retrieval models, however, is usually a simple term-weighting function based on the count of a term in the query. As a result, query-specific properties, such as query length, have been widely assumed as constants that do not affect document ranking, henceforth ignored in the retrieval models [9].

While most of the existing retrieval models are targeted mainly at short keyword queries, their poor performance on longer queries led a large number of efforts that try to understand the properties of verbose queries [1, 8, 11]. Recently, several studies revealed that query length affects the performance of retrieval models through interaction with TF and document length normalization [2, 4, 9]. Chung et al. [2] proposed an adaptive method to estimate the parameters of pivoted document length normalization based on query length. Lv [9] proposed a formal constraint to model the relation between query length and the TF decay speed. He further modified the BM25 retrieval function to satisfy the proposed constraint and showed that the modified version improves the standard BM25 model. The work of Cummins and Riordan [4] is also based on constraint analysis. Similarly, they formalized a constraint to capture the interaction of query length and document length normalization. However, their method only performs comparably to the baseline retrieval models.

In this paper, we focus on the interaction of the discrimination value of query terms with the length of query. Usually, verbose queries contain some unessential terms, while short queries consist of keywords that are almost equally important. We argue that the effect of term discrimination value, which is generally modeled with IDF, should differ in verbose and short keyword queries. We hypothesize that when query length increases, the effect of IDF should be highlighted, in order to facilitate distinguishing more

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR'17, August 7-11, 2017, Shinjuku, Tokyo, Japan
© 2017 ACM. 978-1-4503-5022-8/17/08...$15.00
DOI: 10.1145/3077136.3080761

important terms. We propose a formal constraint to model this hypothesis mathematically and use axiomatic analysis to examine BM25 [12], a state-of-the-art retrieval model, and find that it does not satisfy the proposed constraint. We further modify BM25 so that it adheres to the new constraint. More specifically, we learn that the constraint requires the difference of IDF values to increase with query length and propose a simple function to adapt BM25 to the constraint, which can be computed as efficient as the BM25 model.

We evaluate the new version of BM25 using three TREC collections: AP (Associated Press 1988-89), Robust (TREC 2004 Robust track), and WT10g (TREC 9-10 Web track). To further study the effect of query length, we test two types of queries, short queries and verbose queries, which are borrowed from the title and the description fields of the TREC topics, respectively. Our experimental results demonstrate that our proposed method outperforms the baselines, especially in verbose queries.

2 DEFINITION OF THE FORMAL CONSTRAINT

Axiomatic analysis or constraint analysis of retrieval models provides a formal framework to evaluate existing retrieval models and diagnose their deficiencies. It can be further employed to improve the retrieval models by introducing new developed versions that address the previously found shortcomings [3, 5–7, 10]. This goal is usually achieved by describing a set of desirable properties that a retrieval model should have, and characterizing a reasonable retrieval formula by listing formal constraints that it must satisfy.

Inspired by previous work [5, 9], we propose a formal constraint that retrieval models should satisfy. This constraint, namely QLN-IDF, captures the interaction between IDF and query length. We then provide an analytical analysis of BM25, a state-of-the-art retrieval model, to show that it does not satisfy the proposed constraint. We then propose a modification to BM25 so that the modified version adheres to the constraint.

Notation. We first introduce our notation. $S(Q, D)$ denotes the relevance score of document D for a given query Q, computed by a retrieval model. $|Q|$ and $|D|$ denote the length of Q and D, respectively. $dtf(q, D)$ weighs query term q with respect to the document D based on the count of q in D, i.e., $c(q, D)$. $idf(w)$ denotes the inverse document frequency for a given term w.

In the following, we introduce our formal constraint, named QLN-IDF.

QLN-IDF: Let $Q = \{q_1, q_2\}$ be a query with two terms where $idf(q_1) > idf(q_2)$. Assume that D_1 and D_2 are two documents such that $|D_1| = |D_2|$, $c(q_1, D_1) = c(q_2, D_2) > 0$ and $c(q_1, D_2) = c(q_2, D_1) = 0$ with the document relevance scores $S(Q, D_1)$ and $S(Q, D_2)$. If we reformulate the query $Q^{'} = Q \cup \{q_3\}$ by adding a term to the query, such that $c(q_3, D_1) = c(q_3, D_2) = 0$, then:

$$S(Q, D_1) - S(Q, D_2) < S(Q^{'}, D_1) - S(Q^{'}, D_2) \tag{1}$$

Term discrimination constraint [5] states that when two documents contain the same number of occurrences of query terms, the document that has more discriminative terms should get a higher relevancy score. According to QLN-IDF, the difference of scores between such documents should increase when query gets longer. From an information theoretic perspective, adding a term to the query is equivalent to increasing the information provided by the query. We hypothesize that when query length increases, the effect of IDF should be highlighted, in order to facilitate distinguishing more important terms. Suppose that two documents, say D' and D'', contain q_1 and q_2, respectively. When additional information is given by inclusion of a new query term q_3, which does not occur in either of the mentioned documents, D' should not be penalized as much as D'', because the highlighted effect of IDF means that D' contains more information compared to D''.

3 A MODIFICATION TO BM25

We now analyze the BM25 retrieval model to examine whether it satisfies the proposed constraint or not. Following previous work [9], we use the BM25 formula presented in [5]. The score of a document D with respect to a query Q is calculated as follows:

$$S(Q, D) = \sum_{q \in Q \cap D} qtf(q, Q) \times dtf(q, D) \times idf(q) \tag{2}$$

$$= \sum_{q \in Q \cap D} \frac{(k_3 + 1) \times c(q, Q)}{k_3 + c(q, Q)} \times \frac{(k_1 + 1) \times c(q, D)}{k_1((1 - b) + b\frac{|d|}{avdl}) + c(q, D)} \times \log \frac{N + 1}{df(q)},$$

where k_1, k_3 and b are free hyper-parameters. $avdl$ and N respectively denote the average document length and total number of documents in the collection.

It can be easily proved that the BM25 model does not satisfy the QLN-IDF constraint, since there is no query length related property in the BM25 model. In more details, it can be shown that $S(Q, D_1) - S(Q, D_2) < S(Q^{'}, D_1) - S(Q^{'}, D_2)$ implies the following inequality:

$$idf(q_1) - idf(q_2)\Big|_{|Q|} < idf(q_1) - idf(q_2)\Big|_{|Q+1|}.$$

The above inequality states that the effect of $idf(q)$ should change with respect to the query length. In this regard, we propose the following modification to the BM25 formula:

$$S(Q, D) = \sum_{q \in Q \cap D} qtf(q, Q) \times dtf(q, D) \times (idf(q) + 1)^{\log(|Q|+1)}.$$

The 1 that is added to the log in the power is to prevent it from becoming zero when $|Q|$ is one. The 1 that is added to the idf formula is to ensure that the value of idf is greater than 1, which is necessary for satisfaction of the QLN-IDF constraint. In order to prove that the new BM25 formula results in the satisfaction of QLN-IDF, the following condition should always be true:

$$(idf(q_2) + 1)^{\log(|Q+1|+1)} - (idf(q_2) + 1)^{\log(|Q|+1)}$$
$$< (idf(q_1) + 1)^{\log(|Q+1|+1)} - (idf(q_1) + 1)^{\log(|Q|+1)}.$$

With some basic mathematical derivations, it can be shown that this is always the case when $idf(q_2) < idf(q_1)$, which is implied by the constraint. We refer to this modified version of BM25 as *BM25-QI*.

4 EXPERIMENTS

In this section, we first introduce our collections, experimental setup, and evaluation metrics. We further report and discuss the experimental results.

Table 1: Summary of TREC collections and topics.

ID	Collection	Queries	#docs	#qrels	average query length	
					title (short)	description
AP	TREC 1-3 Ad-hoc Track, Associated Press 88-89	topics 51-200	165k	15,838	4.17	11.31
Robust	TREC 2004 Robust Track Collection	topics 301-450 & 601-700	528k	17,412	2.59	8.43
WT10g	TREC 9-10 Web Track Collection	topics 451-550	1692k	5931	2.48	6.47

Table 2: Comparison of the proposed modification to the BM25 model compared to the baselines. Superscripts 1/2 indicate that the improvements over BM25/BM25-QL are statistically significant.

Short Queries						
Method	AP		Robust		WT10g	
	MAP	P@10	MAP	P@10	MAP	P@10
BM25	0.2717	0.4275	0.2540	0.4353	0.1938	0.2768
BM25-QL	0.2729	0.4262	0.2496	0.4353	0.1959	0.3071
BM25-QI	0.2731	0.4188	0.2550^{2}	0.4353	0.2015	0.2980

Verbose Queries						
Method	AP		Robust		WT10g	
	MAP	P@10	MAP	P@10	MAP	P@10
BM25	0.2468	0.4154	0.2367	0.4100	0.1876	0.3180
BM25-QL	0.2495^{1}	0.4242	0.2340	0.4129	0.1854	0.3330
BM25-QI	0.2691^{12}	0.4201	0.2530^{12}	0.4157	0.2109^{12}	0.3280

4.1 Experimental Design

Collections. We use three standard TREC collections in our experiments: AP, Robust04 and WT10g. AP and Robust are newswire collections, whereas WT10g is a Web collection containing more noisy documents. Statistics of the collections are shown in Table 1.

Experimental Setup. We use two types of queries, short queries and verbose queries, which are taken from the title and the description fields of the TREC topics, respectively. Average query length of title and description fields for all collections are shown in Table 1. All documents are stemmed using the Porter stemmer and stopped using the standard INQUERY stopword list. The experiments were carried out using the Lemur toolkit[1].

Parameter Setting. The parameters b and k_1 of BM25 and BM25-QI are set using 2-fold cross-validation over the queries of each collection. We changed the parameter b from 0 to 1, and the parameter k_1 from 0 to 5 in increments of 0.1. k_3 has no effect in our experiments, because for almost all of the query terms $c(q,Q)$ is equal to 1. Therefore, $qtf(q,Q)$ will be equal to $c(q,Q)$.

Evaluation Metrics. We use two metrics to measure the retrieval quality: (1) mean average precision (MAP) of the top ranked 1000 documents, and (2) the precision of the top 10 retrieved documents (P@10). MAP also serves as the objective function for parameter tuning. Statistically significant differences of performances are determined using the two-tailed paired t-test at a 95% confidence level.

4.2 Results and Discussion

In this subsection, we evaluate the performance of our proposed method (BM25-QI) and compare its performance to those obtained by the baselines. We further study the sensitivity of our method to the input parameters.

4.2.1 *Evaluation of the Proposed Method*

We consider two baselines: (1) the original BM25 method, and (2) the enhanced version of BM25 (BM25-QL) which satisfies QLN-TFC constraint proposed in [9]. This model computes the retrieval score as follows:

$$S(Q,D) = \sum_{q \in Q \cap D} c(q,Q) \times \frac{(\alpha . \log(|Q|) + \beta + 1)c'(q,D)}{\alpha . \log(|Q|) + \beta + c'(q,D)} \times \log \frac{N+1}{df(q)},$$

where

$$c'(q,D) = \frac{c(q,D)}{1 - ((\alpha' . \log(|Q|) + \beta')) + ((\alpha' . \log(|Q|) + \beta') \frac{|D|}{avdl}},$$

and α, β, α', and β' are free parameters that are estimated using supervised learning to find the optimal k_1 and b parameters in the original BM25 retrieval model [9]. The parameters are optimized using the linear regression model implemented in the scikit-learn toolkit[2].

Table 2 summarizes the results achieved by the proposed method and the baselines. We separate title and description queries to demonstrate the behavior of methods for different query lengths.

The results show that BM25-QI outperforms BM25 and BM25-QL consistently in terms of MAP and also achieves better or comparable P@10 scores compared to BM25. The MAP improvements of BM25-QI over the baselines are much larger on verbose queries. In particular, the MAP improvements on all collections are statistically significant for verbose queries. This is likely because when the number of terms in a query is high, the range of IDF values is wide. In this situation, IDF can be a good signal of term discrimination value. However, short queries usually consist of keywords and the variance of their IDF values are smaller, which degrades the effect of our proposed constraint and modification.

Another interesting observation is that, while the performance of the baselines is substantially better on short queries in all collections, the performance of BM25-QI on verbose queries is comparable to its performance on short queries on AP and Robust, and has improved on WT10g. This result empirically confirms that the proposed method better adapts to different query lengths. In other words, while existing retrieval models [9, 12] are targeted mainly at short keyword queries and perform poorly on longer queries, our proposed model significantly improves the performance for verbose queries while achieving comparable and in some cases better results on short queries.

[1] http://lemurproject.org/

[2] http://scikit-learn.org/

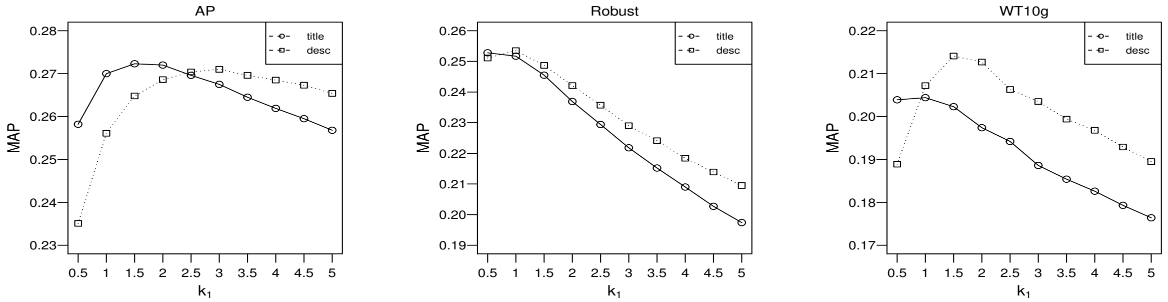

Figure 1: Sensitivity of the proposed method to the parameter k_1 in different collections, for title and description queries.

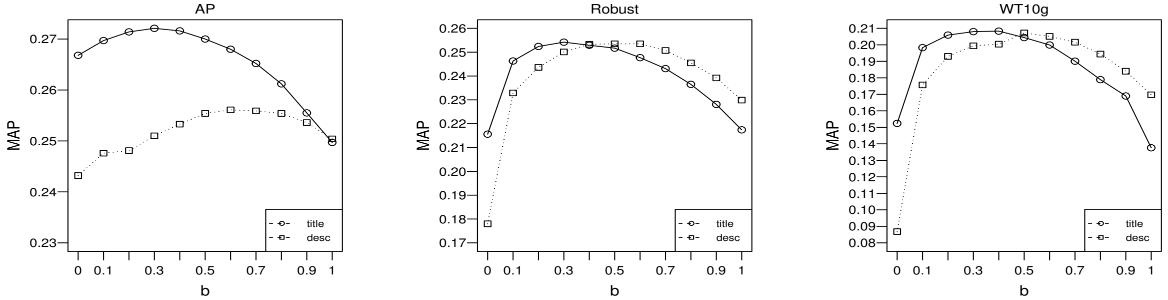

Figure 2: Sensitivity of the proposed method to the parameter b in different collections, for title and description queries.

4.2.2 Parameter Sensitivity

In this set of experiments, we study the performance of BM25-QI in terms of MAP with respect to the parameters k_1 and b for both query sets on all of the collections. The results are shown in Figures 1 and 2. According to these figures, large values of k_1 hurt the performance of BM25-QI on both title and description queries, and description queries almost always achieve higher MAP values compared to title queries. These results emphasize the advantage of our model for verbose queries. The performance of BM25-QI is more stable with respect to k_1 on AP, and higher values of k_1 seem to be more suitable for longer queries. A similar behavior on Robust and WT10g is again observed with respect to b. For lower values of b, the performance is better on title queries and as value of b goes up from 0.5, description queries reach to higher values of MAP. However, the overall performance is stable in the $[0.1, 0.5]$ interval for title queries and in the $[0.3, 0.7]$ interval for description queries.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we analyzed the interaction between query length and IDF, a term discrimination heuristic that can be thought as a signal that measures the relative importance of query terms. We proposed the idea that the effect of discrimination value of query terms should not be the same in verbose queries and short queries. We hypothesized that when query length increases, the effect of IDF should be highlighted, in order to facilitate distinguishing more important terms. To formalize this idea, we proposed a new constraint that any reasonable retrieval function should satisfy. We have then studied the BM25 model and revealed that it does not adhere to the constraint. We proposed a modification to BM25 based on our axiomatic analysis. Our experimental results showed that the modified version of BM25 outperforms the original one, in particular for the cases where queries are long. In the future, we intend to study other retrieval models, such as query likelihood, and analyze their adherence to the proposed constraint. Investigating a more effective approach to model the interaction between query length and IDF is also an interesting research direction for future work.

Acknowledgements. This work was supported in part by the Center for Intelligent Information Retrieval and in part by a grant from the Institute for Research in Fundamental Sciences (No. CS1396-4-51). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

REFERENCES

[1] Michael Bendersky and W. Bruce Croft. 2008. Discovering Key Concepts in Verbose Queries. In *SIGIR '08*. Singapore, Singapore, 491–498.

[2] Tze Leung Chung, Robert Wing Pong Luk, Kam Fai Wong, Kui Lam Kwok, and Dik Lun Lee. 2006. Adapting Pivoted Document-length Normalization for Query Size: Experiments in Chinese and English. *ACM Transactions on Asian Language Information Processing (TALIP)* 5, 3 (2006), 245–263.

[3] Ronan Cummins. 2016. A Study of Retrieval Models for Long Documents and Queries in Information Retrieval. In *WWW '16*. Montreal, Quebec, Canada, 795–805.

[4] Ronan Cummins and Colm O'Riordan. 2012. A Constraint to Automatically Regulate Document-length Normalisation. In *CIKM '12*. Maui, Hawaii, USA, 2443–2446.

[5] Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A Formal Study of Information Retrieval Heuristics. In *SIGIR '04*. Sheffield, United Kingdom, 49–56.

[6] Hui Fang and ChengXiang Zhai. 2005. An Exploration of Axiomatic Approaches to Information Retrieval. In *SIGIR '05*. Salvador, Brazil, 480–487.

[7] Hui Fang and ChengXiang Zhai. 2006. Semantic Term Matching in Axiomatic Approaches to Information Retrieval. In *SIGIR '06*. Seattle, Washington, USA, 115–122.

[8] Manish Gupta and Michael Bendersky. 2015. Information Retrieval with Verbose Queries. *Foundations and Trends in Information Retrieval* 9, 3-4 (2015), 209–354.

[9] Yuanhua Lv. 2015. A Study of Query Length Heuristics in Information Retrieval. In *CIKM '15*. Melbourne, Australia, 1747–1750.

[10] Yuanhua Lv and ChengXiang Zhai. 2011. Lower-bounding Term Frequency Normalization. In *CIKM '11*. Glasgow, Scotland, UK, 7–16.

[11] Jiaul H. Paik and Douglas W. Oard. 2014. A Fixed-Point Method for Weighting Terms in Verbose Informational Queries. In *CIKM '14*. Shanghai, China, 131–140.

[12] S. E. Robertson and S. Walker. 1994. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *SIGIR '94*. Dublin, Ireland, 232–241.