A Study of Query Length Heuristics in Information Retrieval

Yuanhua Lv
Microsoft Research
Redmond, WA 98052
yuanhual@microsoft.com

ABSTRACT

Query length has generally been regarded as a query-specific constant that does not affect document ranking. In this paper, we reveal that query length actually interacts with term frequency (TF) normalization, a key component of all effective retrieval models. Specifically, the longer the query is, the smaller the TF decay speed should be. In order to study the impact of query length, we present a desirable formal constraint to capture the heuristic of query length for retrieval. Our constraint analysis shows that current state-of-the-art retrieval functions, including BM25 and language models, fail to satisfy the constraint, and that, in order to solve this problem, the TF normalization component in a retrieval function should be adapted to query length. As an application, we develop a simple regression algorithm to adapt BM25 to query length, and demonstrate its effectiveness on several representative TREC collections.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Algorithms, Theory

Keywords

Query length, term frequency normalization, constraints

1. INTRODUCTION

Optimization of retrieval models is a fundamental research problem in information retrieval. Although many studies have attempted to improve retrieval models from various perspectives, e.g., [10, 11, 3, 6], they mainly focused on either document properties (e.g., term frequency and document length) or the estimation of term statistics from the corpus (e.g., IDF and background model), while little attention has been paid to the query length (i.e., the count of terms in a query). In fact, query length has largely been regarded as a query-specific constant that does not affect document ranking.

In this paper, we reveal that query length actually plays a role in retrieval models through influencing term frequency (TF) normalization. It is widely recognized that TF should be normalized to properly weaken the contribution of repeated term occurrences (as compared to the first occurrence) of a term, based on the intuition that the first occurrence of the term often brings more relevance evidence. However, we argue that the first occurrence of a term may not be that important when the query is long, as the terms of a longer query generally tends to be more semantically redundant (i.e., one query term may be semantically overlapping with other query terms), and the relative importance of the first occurrence of a new query term (as compared to its repeated occurrences) in a document from a longer query may not be as high as that from a shorter query. We thus hypothesize that the longer the query is, the less penalization the repeated term occurrences should receive. In other words, a document that mis-matches a query term from a longer query should not be penalized as much as that of another document that mis-matches the query term from a shorter query.

We propose a formal constraint to model the interaction between query-length and TF normalization mathematically, so that it is possible to apply it to not only diagnose a retrieval function analytically and but also guide us to fix the problem. We then use constraint analysis to examine BM25 [8, 9] and language models [11], and find that neither of them satisfies the constraint. More specifically, for BM25, the proposed constraint is equivalent to a requirement that its parameter k_1 should increase monotonically with the query length; for the language modeling approach with Dirichlet prior smoothing [11], the constraint requires that the Dirichlet prior μ should increase monotonically with the query length. In other words, the optimal k_1 and μ values for a longer query should generally be larger than that for a shorter query; as a result, a query-independent parameter setting would not be optimal for all queries. This has also been confirmed by the empirical evidences in Section 3.2.

Motivated by this understanding, we propose a simple methodology for adapting retrieval models to query length using a linear regression method, which incurs almost no additional computational cost. As an application, we apply the proposed method to estimate a query-length aware k_1 for BM25. Our experimental results on multiple representative TREC collections demonstrate its effectiveness.

2. RELATED WORK

In our previous works [5, 7], we explored how to adapt TF normalization to the global distribution of individual terms in the whole corpus. Although related, those works are orthogonal to the current work, as they only used global term statistics but did not explore query length. We emphasize that our major contributions in this paper are the query

CIKM'15, October 19–23, 2015, Melbourne, Australia.
© 2015 ACM. ISBN 978-1-4503-3794-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806416.2806592.

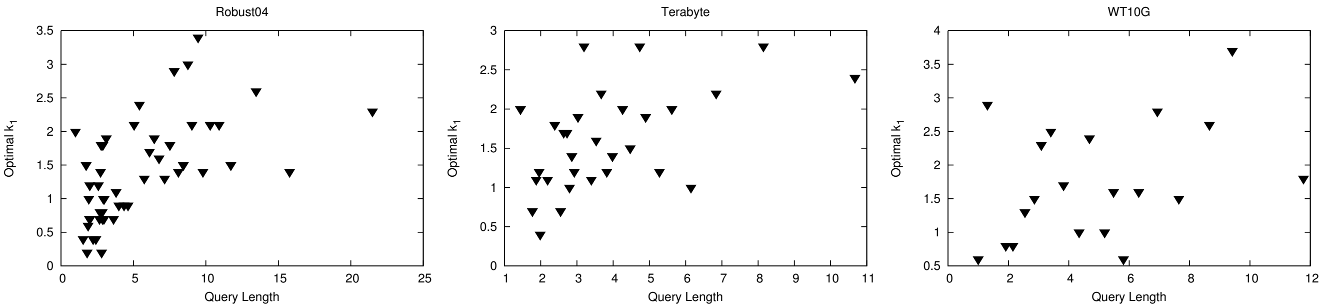

Figure 1: Plots of the optimal k_1 w.r.t. the query length (as described in Section 3.2).

length heuristics and constraints, which are novel. Nevertheless, we have compared the proposed method with our previous work [5] in the experiments.

The interaction between query length and document length normalization has been studied [1, 2]. Chung et al. [1] have incorporated the query length into the pivoted document length normalization in the vector space model (rather than BM25), but they did not reach any conclusive results; Cummins and Riordan's recent work [2] examined a similar heuristic, and formalized a constraint to describe the heuristic, but their method only performs comparably to the baseline retrieval models. Moreover, both works focused only on document length normalization, but did not pay attention to the effect of query length in TF normalization.

Constraint analysis has been explored in information retrieval to diagnostically evaluate existing retrieval models [3], introduce novel retrieval signals into existing retrieval models [6, 2], and guide the development of new retrieval models [4]. Our constraint analysis is inspired by these previous works, but the proposed constraint QLN-TFC is novel.

3. FORMAL QUERY LENGTH CONSTRAINT

How can we regulate the interactions between TF normalization and query length as discussed in Section 1 so that we can adapt a retrieval function to different queries? To answer this question, we first propose a desirable formal constraint, namely QLN-TFC, that any reasonable retrieval function should satisfy.

QLN-TFC: Let $Q = \{q_1, q_2\}$ be a query with two terms q_1 and q_2. Assume D_1 and D_2 are two documents such that $|D_1| = |D_2|$, $c(q_1, D_1) = n > 1$, $c(q_2, D_1) = 0$, $c(q_1, D_2) = c(q_2, D_2) = 1$, and the document relevance scores $S(Q, D_1) = S(Q, D_2)$. If we reformulate the query $Q' = Q \cup \{q_3\}$, by adding another term $q_3 \notin Q$ into the query, where $c(q_3, D_1) = c(q_3, D_2) = 1$, then $S(Q', D_1) > S(Q', D_2)$.

This constraint ensures that the relative importance of the first occurrence of a new query term (as compared to its repeated occurrences) in a longer query should not be as high as that in a shorter query. To illustrate, a document, say D_1, that mis-matches a query term q_2 from a longer query Q' should not be penalized as much as that D_1 mis-matches the query term q_2 from a shorter query Q.

3.1 Analytical Analysis

We now diagnose retrieval models using the proposed constraint QLN-TFC. We use the widely accepted BM25 [8, 9] as an example. The BM25 formula, as presented in [3], scores a document D with respect to query Q as follows:

$$\sum_{q \in Q \cap D} c(q, Q) \cdot dtf(q, D) \cdot \log \frac{N+1}{df(q)} \qquad (1)$$

where $c(q, Q)/c(q, D)$ represents the frequency of term q in Q/D, $df(q)$ is the number of documents containing term q, N is the total number of documents in the collection, and $dtf(q, D)$ is the key component of BM25 contributing to its success, i.e., the sub-linear TF normalization formula, which prevents the contribution of repeated occurrences of a term from growing too large:

$$dtf(q, D) = \frac{(k_1 + 1) \cdot c(q, D)}{k_1 \left(1 - b + b\frac{|D|}{avdl}\right) + c(q, D)} = \frac{(k_1 + 1) \cdot c'(q, D)}{k_1 + c'(q, D)}$$

where

$$c'(q, D) = \frac{c(q, D)}{1 - b + b\frac{|D|}{avdl}}$$

is the pivoted normalization method [10] for document length normalization, in which, $|D|$ is the length of document D, and $avdl$ is the average document length. There are two important parameters: (1) k_1 is used to control the shape of this TF normalization component, and a larger k_1 penalizes the repeated term occurrences less; (2) $b \in [0, 1]$ is the slope parameter. Both k_1 and b are usually set independently on query length.

Let's look at the QLN-TFC constraint. Consider the average case when $|D_1| = |D_2| = avdl$. It can be shown that the $S(Q, D_1) = S(Q, D_2)$ implies the following equality:

$$\left.\frac{(k_1+1)n}{k_1+n} \log \frac{N+1}{df(q_1)}\right|_{|Q|} = \left.\log \frac{N+1}{df(q_1)} + \log \frac{N+1}{df(q_2)}\right|_{|Q|}$$

And after query reformulation, the constraint $S(Q', D_1) > S(Q', D_2)$ requires

$$\left.\frac{(k_1+1)n}{k_1+n} \log \frac{N+1}{df(q_1)} + \log \frac{N+1}{df(q_3)}\right|_{|Q'|} > \left.\sum_{i \in \{1,2,3\}} \log \frac{N+1}{df(q_i)}\right|_{|Q'|}$$

With some basic mathematical derivations, the constraint QLN-TFC is equivalent to the following inequality:

$$\left.\frac{(k_1+1)n}{k_1+n}\right|_{|Q|} < \left.\frac{(k_1+1)n}{k_1+n}\right|_{|Q|+1} \qquad (2)$$

which clearly cannot be satisfied by the standard BM25, given that k_1 is independent on query length $|Q|$. To satisfy the constraint, *it is required that the parameter k_1 should increase monotonically with the query length $|Q|$.*

Similarly, we also analyze another representative retrieval function, the language modeling approach with Dirichlet prior smoothing method [11], for which the constraint QLN-TFC is equivalent to the following inequality:

$$\left.\frac{1}{\mu \cdot p(q_1|C)p(q_2|C)}\right|_{|Q|} > \left.\frac{1}{\mu \cdot p(q_1|C)p(q_2|C)}\right|_{|Q|+1} \qquad (3)$$

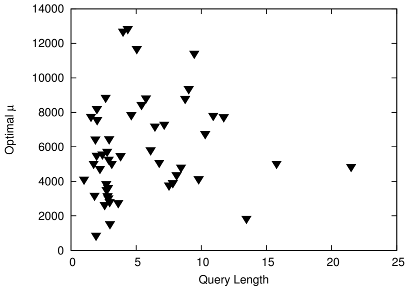

Figure 2: Plots of optimal μ w.r.t. the query length.

	WT10G	Terabyte	Robust04
queries	451-550	701-850	301-450, 601-700
#qry(with qrel)	198	298	498
mean(ql)	7.93	7.34	9.10
std(ql)	5.44	5.16	8.53
max(ql)	30	24	62
#total_qrel	$5,981$	$28,640$	$17,412$
#documents	$1692k$	$25205k$	$528k$

Table 1: Characteristics of document and query sets

where μ is the Dirichlet prior, and $p(q_i|C)$ is the probability of q_i estimated using the whole collection. This shows that *in order for the language modeling approach to satisfy the constraint, an adaptive Dirichlet prior μ that increases monotonically with the query length $|Q|$ is desired.*

3.2 Empirical Analysis

Our above analysis has shown that, analytically, query length interacts with TF normalization. Now we turn to seeking empirical evidence to see if this is a common behavior of modern information retrieval models.

We first plot the optimal k_1 [1] in BM25 for each query w.r.t. the query length on different TREC collections in Figure 1. We do a binning analysis, where we rank all queries according to their lengths and then group every 10 continuous queries into a bin. The "query length" of a bin is the average length of all queries in the bin, and the "k_1" of the bin is the median of all k_1 values in the bin. We can see there is indeed clear correlation between the query length and the optimal k_1 value, especially on Robust04 and Terabyte, probably because Robust04 and Terabyte consist of more query topics (498 and 298 respectively) than WT10G (only 198). Anyway, this confirms our intuition that TF normalization interacts with query length, and thus a query length independent k_1 would not be optimal for all queries.

In addition, we also analyze the optimal Dirichlet prior μ in the language modeling approach w.r.t. the query length in a similar way based on Robust04, as shown in Figure 2, and find μ also has a positive correlation with query length.

4. QUERY-LENGTH AWARE BM25

4.1 Query-Length Aware k_1

In order to improve current retrieval models to satisfy QLN-TFC, we need to make their TF normalization components dependent on query-length. However, we do not want that the addition of this new constraint changes the implementations of other existing retrieval heuristics in these retrieval functions that have been shown to work quite well [3]. We propose a heuristic approach to achieve this goal by making the corresponding parameters in the TF normalization component aware of query-length using a simple regression method. Specifically, $k_1 = f(|Q|)$ in BM25 and the Dirichlet prior $\mu = g(|Q|)$ in the language modeling approach, where $f(\cdot)$ and $g(\cdot)$ are functions to be estimated. In this paper, we focus only on the improvement of BM25 due to the space reason. With $k_1 = f(|Q|)$, the Inequality 2 is equivalent to:

$$f(|Q|) < f(|Q| + 1) \tag{4}$$

suggesting that a monotonically increasing function $f(\cdot)$ will make BM25 satisfy QLN-TFC unconditionally. And interestingly, this is consistent with the data analysis results in Figure 1, confirming empirically that QLN-TFC is a desirable retrieval constraint. Observing the **sublinear** curves of k_1 w.r.t. the query length in Figure 1, a heuristic approximation of $f(\cdot)$ is thus obtained as follows:

$$k_1 = f(|Q|) = \alpha \cdot \log(|Q|) + \beta \tag{5}$$

where α and β are two free parameters. The logarithm function in $f(\cdot)$ intuitively makes sense: $k_1 = f(|Q|)$ should increase more when the query length increases from 3 to 4 than when the query length increases from 10 to 11.

As we have already collected the optimal k_1 value for each training query, we can use the standard curve-fitting technique and the least square method to fit $f(\cdot)$ to the ground truth to estimate both α and β. Finally, substituting Equation 5 into 1, we get the following retrieval function,

$$\sum_{q \in Q \cap D} c(q,Q) \cdot \frac{(\alpha \cdot \log(|Q|) + \beta + 1) \cdot c'(q,D)}{\alpha \cdot \log(|Q|) + \beta + c'(q,D)} \cdot \log \frac{N+1}{df(q)} \tag{6}$$

4.2 Query-Length Aware b

In addition, we also revisit the heuristic of query length interacting with document length normalization. This heuristic is not entirely novel, as a previous work [1] has reported that a larger b value is often needed for longer queries in the vector space model based on analyzing several small document collections. We revisit this heuristic on BM25 using several larger TREC collections. We first do a similar binning analysis, and plot the optimal b values w.r.t. the corresponding query lengths using BM25, which shows that the optimal b correlates with the query length well (and the optimal b also increases sub-linearly with the query length). We also apply the same linear regression method to compute a query-length aware $b = \alpha' \log |Q| + \beta'$.

5. EXPERIMENTS

5.1 Testing Collections and Evaluation

We use three representative TREC collections: WT10G, Terabyte and Robust04, which represent different sizes and genre of text collections. WT10G and Terabyte are medium and large Web collections respectively. Robust04 is a representative news dataset. Our queries are taken from *both the title and the description fields* of the TREC topics. For all the datasets, the preprocessing of documents and queries is minimum, involving only Porter's stemming. We do not remove any stopwords. An overview of the involved query topics and document collections are shown in Table 1.

The top-ranked 1000 documents for each run are compared in terms of their mean average precisions (MAP), which also serves as the objective function for parameter training. In addition, the precision at top-10 documents (P@10) is also considered.

[1] We do a linear search of the optimal k_1 (that optimizes AP) for each query in range $[0, 5]$ with a step 0.1.

	WT10G		Terabyte		Robust04	
	MAP	P@10	MAP	P@10	MAP	P@10
BM25	0.1768	0.3051	0.2403	0.5443	0.2352	0.4163
BM25$_Q$	0.1668	0.2904	0.2427	0.5107	0.2335	0.4153
BM25-adpt [5]	0.1765	0.3040	-	-	0.2306	0.4044
BM25QL (b)	0.1800^{2}	0.3045	0.2452^{12}	0.5389	0.2413^{123}	0.4207
BM25QL (k_1)	0.1787^{2}	0.3056	0.2562^{12}	0.5490	0.2371^{13}	0.4155
BM25QL	$\mathbf{0.1817}^{123}$	0.3035	$\mathbf{0.2599}^{12}$	0.5426	$\mathbf{0.2422}^{123}$	0.4211
UpperBound	0.2059	0.3263	0.2723	0.5589	0.2503	0.4295

Table 2: Performance comparison. ^{1}, ^{2} and ^{3} indicate significant MAP improvements over BM25, BM25$_Q$ and BM25-adpt respectively at the 0.01 level using the Wilcoxon non-directional test.

We first employ a two-fold cross-validation for parameter tuning, where the query topics are split into even and odd numbered topics as the two folds. We compare the proposed query-length aware BM25 (labeled as "BM25QL') with two baselines: (1) a standard BM25 for which both b and k_1 are trained (labeled as "BM25"), and (2) a query-length specific BM25 (labeled as "BM25$_Q$") that does a two-fold cross-validation among queries of the same length (if any), and degenerates to "BM25" if there is no other query of the same length. The comparison results are summarized in Table 2. We can see that BM25QL outperforms both BM25 and BM25$_Q$ significantly in terms of MAP. BM25$_Q$ is sometimes even worse than BM25. These observations suggest that a "hard" way to train a query-length specific retrieval model (like BM25$_Q$) may not work well due to the reduction of training data, but a "soft" way, like BM25QL, that uses all the training data to train a query-length aware retrieval function, works effectively. In addition, we can see that although BM25QL improves MAP significantly over the baselines, its P@10 scores are similar to those baselines without any significant differences, probably because BM25QL is trained to optimize MAP rather than P@10.

We also compare BM25QL with our previous work on adaptive BM25, namely BM25-adpt [5], on WT10G and Robust04 in Table 2 [2]. We use an enhanced version of BM25-adpt [5], where we add a free parameter α into the objective function of BM25-adpt to control the range of k_1, formally

$$\arg\min_{k_1} \sum_{i=0}^{T} \left(\alpha \cdot \frac{IG_q^t}{IG_q^1} - \frac{(k_1+1)\cdot t}{k_1+t} \right)^2 \qquad (7)$$

Obviously, the new object function takes the original one [5] as its special case when $\alpha = 1$. In our work, α is also optimized using cross validation. We can see that the proposed BM25QL works better. One possible reason is that, although BM25-adpt works well on short keyword queries [5], it does not work effectively on verbose queries that are being used in this work. But the proposed method works better to adapt to different query lengths.

To understand the performance of query-length aware k_1 and query-length aware b separately, we train two retrieval functions "BM25QL (k_1)" and "BM25QL (b)", respectively. We use a query-length aware k_1 and a query-length unaware b [3] in the former, while using a query-length aware b and a query-length unaware k_1 in the latter. We can see that both methods work more effectively than BM25 and BM25$_Q$ in terms of MAP, suggesting that both parts contribute to the performance of BM25QL.

6. CONCLUSIONS

In this paper, we revealed that query length affects TF normalization, and explored an idea of query-length aware retrieval models. Specifically, we proposed a desirable formal constraint to capture the heuristic of query length in retrieval functions, diagnosed BM25 and other retrieval functions to show that they cannot satisfy the constraint, and developed a query-length aware TF normalization methodology. We applied the proposed techniques on BM25, which was shown to improve the standard BM25 significantly.

7. ACKNOWLEDGMENTS

I thank Dr. ChengXiang Zhai for his helpful discussions, and three anonymous reviewers for their useful comments.

8. REFERENCES

[1] T. L. Chung, R. W. P. Luk, K. F. Wong, K. L. Kwok, and D. L. Lee. Adapting pivoted document-length normalization for query size: Experiments in chinese and english. *ACM Trans. Asian Lang. Inf. Proc.*, 5(3):245–263, Sept. 2006.

[2] R. Cummins and C. O'Riordan. A constraint to automatically regulate document-length normalisation. In *CIKM '12*, pages 2443–2446, 2012.

[3] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *SIGIR '04*, pages 49–56, 2004.

[4] H. Fang and C. Zhai. An exploration of axiomatic approaches to information retrieval. In *SIGIR '05*, pages 480–487, 2005.

[5] Y. Lv and C. Zhai. Adaptive term frequency normalization for bm25. In *CIKM '11*, pages 1985–1988, 2011.

[6] Y. Lv and C. Zhai. Lower-bounding term frequency normalization. In *CIKM '11*, pages 7–16, 2011.

[7] Y. Lv and C. Zhai. A log-logistic model-based interpretation of tf normalization of bm25. In *ECIR'12*, pages 244–255, 2012.

[8] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94*, pages 232–241, 1994.

[9] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *TREC '94*, pages 109–126, 1994.

[10] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *SIGIR '96*, pages 21–29, 1996.

[11] C. Zhai and J. D. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR '01*, pages 334–342, 2001.

[2] BM25-adpt was implemented based on the Lemur toolkit (http://www.lemurproject.org/lemur.php) that is not scalable to Terabyte data.

[3] From the standard cross validation run "BM25"