On Setting the Hyper-Parameters of Term Frequency Normalization for Information Retrieval

BEN HE and IADH OUNIS
University of Glasgow

The setting of the term frequency normalization hyper-parameter suffers from the query dependence and collection dependence problems, which remarkably hurt the robustness of the retrieval performance. Our study in this article investigates three term frequency normalization methods, namely normalization 2, BM25's normalization and the Dirichlet Priors normalization. We tackle the query dependence problem by modifying the query term weight using a Divergence From Randomness term weighting model, and tackle the collection dependence problem by measuring the correlation of the normalized term frequency with the document length. Our research hypotheses for the two problems, as well as an automatic hyper-parameter setting methodology, are extensively validated and evaluated on four Text REtrieval Conference (TREC) collections.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*

General Terms: Algorithms, Experimentation, Theory

Additional Key Words and Phrases: Query-dependence, collection-dependence, information retrieval models, relevance feedback, term frequency normalization, TREC experimentation

ACM Reference Format:

He, B. and Ounis, I. 2007. On setting the hyper-parameters of term frequency normalization for information retrieval. ACM Trans. Inform. Syst. 25, 3, Article 13 (July 2007), 24 pages. DOI = 10.1145/1247715.1247719 http://doi.acm.org/10.1145/1247715.1247719

1. INTRODUCTION

In Information Retrieval (IR), the dependence between the document length and the term frequency (*tf*) has an important impact on assigning relevance

This work is funded by the Leverhulme Trust, grant number F/00179/S. The project funds the development of the Smooth project, which investigates term frequency normalization (URL: http://ir.dcs.gla.ac.uk/smooth).
The experiments in this article were conducted using Terrier [Ounis et al. 2006] (URL: http://ir.dcs.gla.ac.uk/terrier).

Authors' address: Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, United Kingdom. Corresponding email: ben@dcs.gla.ac.uk.

scores to the returned documents. The assignment of the relevance scores aims to provide an adequate document ranking for the given query [Van Rijsbergen 1979]. In this article, the document length refers to the number of tokens in a document. In order to smooth this dependence, a technique called *term frequency normalization* is usually applied.

Many normalization methods have been developed in the past, including the so-called normalization 2 [Amati and Van Rijsbergen 2002] and BM25's normalization component [Robertson et al. 1995]. Moreover, as proposed in Amati [2003] and He and Ounis [2005], the Dirichlet Priors, which have been applied to the IR language modeling approach [Zhai and Lafferty 2001], can also be applied to *tf* normalization. In this article, we denote *tf* normalization using the Dirichlet Priors as the *Dirichlet Priors normalization*. We will introduce these normalization methods in Section 2.

All the above three normalization methods can be seen as a *tf* density estimation of the document length. In previous work, the hyper-parameters of the normalization methods are empirically set [Amati and Van Rijsbergen 2002; Amati 2003; Robertson et al. 1995; Zhai and Lafferty 2001]. (In Zhai and Lafferty [2001], the Dirichlet Priors are applied in the context of Zhai and Lafferty's language modeling approach.) However, this empirical strategy suffers from the following two problems:

(1) *The Query Dependence Problem.* The optimal setting depends on the query type. Indeed, as an illustration, for the TREC ad-hoc tasks, the optimal settings for short and long queries are widely different [Amati and Van Rijsbergen 2002; Amati 2003]. In this article, the optimal setting refers to the setting that results in the highest mean average precision.
(2) *The Collection Dependence Problem.* The optimal setting varies across diverse collections.

As a consequence, the use of the empirical setting strategy for the hyper-parameters of *tf* normalization methods can significantly hurt the robustness of an IR system and results in unstable retrieval performance. For example, for the topics in the TREC-9 [Hawking 2000] and TREC-10 [Hawking and Craswell 2001] ad-hoc tasks, using all the query fields in the topics, the PL2 weighting model [Amati and Van Rijsbergen 2002] provides a mean average precision (MAP) of 0.2327 with the default hyper-parameter setting, while it provides an MAP of 0.2535 with the actual optimal hyper-parameter setting. The p-value is 3.948e-4 according to the Wilcoxon test, which indicates a significant difference in terms of retrieval performance.

In this article, we aim to conduct a comprehensive study of *tf* normalization that provides a better understanding of the above two problems. In our hypothesis, we suggest that the query dependence problem is due to an inadequate original query term weighting so that the informative query terms do not stand out from the noninformative ones. In our proposed study, we use a Divergence From Randomness (DFR) term weighting model to reweigh the query terms in order to refine the query term weights and overcome the query dependence problem. We will introduce the applied DFR term weighting model in Section 2. Next, we base our study of the collection dependence problem on the query reweighing. We hypothesize that the optimal settings of the reweighed queries on diverse collections provide similar correlations between the normalized term frequency and the document length. Using the above two hypotheses (the latter hypothesis is based on the former), we propose an automatic hyper-parameter setting methodology for *tf* normalization. The proposed methodology tackles both the query dependence and collection dependence problems, and it is applicable to the three different *tf* normalization methods used. Using four TREC test collections, our research hypotheses and the proposed automatic hyper-parameter setting methodology are extensively validated and evaluated in experiments with three normalization methods, that is, normalization 2, BM25's normalization and the Dirichlet Priors normalization. The results show that our research hypotheses hold for the three involved normalization methods, and that the proposed methodology is effective and stable across various test collections. We believe that the research hypotheses are general enough to cope with *tf* normalization, and the derived methodology provides a solution for automatically setting the hyper-parameters of term frequency normalization.

The remainder of this article is organized as follows: we introduce the related work, that is, normalization 2, BM25's normalization and the Dirichlet Priors for the *tf* normalization, as well as the DFR weighting models, in Section 2. We study the query dependence problem and validate our hypothesis for this problem in Section 3. In Section 4, we study the collection dependence problem. We propose an automatic hyper-parameter setting methodology that tackles the two problems. The hypothesis, as well as the proposed hyper-parameter setting methodology are validated and evaluated, respectively, in the experiments. Experiments studying how query expansion affects the hyper-parameter setting are presented in Section 5. Finally, we conclude the study in Section 6.

2. RELATED WORK

In Section 2.1, we introduce the normalization methods that are studied in this article. In Section 2.2, we introduce the DFR term weighting models for pseudo-relevance feedback.

2.1 The Normalization Methods

As one of the most established weighting models, Okapi's *BM25* computes the relevance score of a document d for a query Q by the following formula [Robertson et al. 1995]:

$$score(d, Q) = \sum_{t \in Q} w^{(1)} \frac{(k_1 + 1)tf}{K + tf} qtw \tag{1}$$

where

—*tf* is the within document frequency of the given term t.

—k_1 is a parameter. Its default setting is $k_1 = 1.2$ [Robertson et al. 1995].

—$w^{(1)}$ is the *idf* factor, which is given by:

$$w^{(1)} = \log_2 \frac{N - N_t + 0.5}{N_t + 0.5},$$

where N is the number of documents in the whole collection. N_t is the document frequency of term t.

—qtw is the query term weight that is given by

$$\frac{(k_3 + 1)qtf}{k_3 + qtf},$$

where qtf is the number of occurrences of the given term in the query. k_3 is a parameter. Its default setting is $k_3 = 1000$ [Robertson et al. 1995].

—K is:

$$k_1 \left((1 - b) + b \frac{l}{avg\_l} \right),$$

where l and avg_l are the document length and the average document length in the collection, respectively. The document length refers to the number of tokens in a document. b is a hyper-parameter. The default setting is $b = 0.75$ [Robertson et al. 1995].

In the following derivation, we show how the BM25 model employs a *tf* normalization component.

Let $tfn = \frac{tf}{(1-b)+b\cdot\frac{l}{avg\_l}}$, where *tfn* denotes the normalized term frequency. We obtain:

$$\begin{aligned} score(d, Q) &= \sum_{t \in Q} w^{(1)} \frac{k_1 + 1}{\frac{k_1}{tfn} + 1} \frac{(k_3 + 1)qtf}{k_3 + qtf} \\ &= \sum_{t \in Q} w^{(1)} \frac{(k_1 + 1)tfn}{k_1 + tfn} \frac{(k_3 + 1)qtf}{k_3 + qtf}. \end{aligned} \tag{2}$$

Hence, the term frequency normalization component of the BM25 formula is:

$$tfn = \frac{tf}{(1 - b) + b \cdot \frac{l}{avg\_l}}. \tag{3}$$

The *pivoted normalization* for normalizing the $tf \cdot idf$ weight [Singhal et al. 1996] can be seen as a generalization of the above BM25's normalization method. Singhal et al. [1996] studied a set of collections and proposed a heuristic normalization method by "pivoting" a normalization factor to fit the relevance judgmen t information.

PL2 is one of the Divergence From Randomness (DFR) document weighting models [Amati and Van Rijsbergen 2002]. The idea of the DFR models is to infer the importance of a query term in a document by measuring the divergence of the term's distribution in the document from randomness. In the PL2 model, the randomness is modelled by an approximation to the Poisson distribution with the use of the Laplace succession to normalise the relevance score. Using the PL2 model, the relevance score of a document d for a query Q is given by:

$$score(d, Q) = \sum_{t \in Q} qtw \cdot \frac{1}{tfn + 1} \left(tfn \cdot \log_2 \frac{tfn}{\lambda} + (\lambda - tfn) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tfn) \right) \quad (4)$$

where λ is the mean and variance of a Poisson distribution. It is given by $\lambda = F/N$. F is the frequency of the query term in the collection and N is the number of documents in the collection. The query term weight qtw is given by qtf/qtf_{max}. qtf is the query term frequency. qtf_{max} is the maximum query term frequency among the query terms.

The normalized term frequency *tfn* is given by the so-called *normalization 2*:

$$tfn = tf \cdot \log_2 \left(1 + c \cdot \frac{avg\_l}{l} \right), (c > 0) \quad (5)$$

where l is the document length and avg_l is the average document length in the whole collection. *tf* is the original term frequency. c is the hyper-parameter of normalization 2. Its default setting is $c = 7$ for short queries and $c = 1$ for long queries [Amati and Van Rijsbergen 2002].

The Dirichlet Priors stand for the priors in a Dirichlet distribution, which is a generalization of the Beta distribution in a multinomial case. Zhai and Lafferty [2001] have applied the Dirichlet Priors in defining their language model for IR. The Dirichlet Priors can also be used for *tf* normalization [Amati 2003; He and Ounis 2005]:

$$tfn = \frac{tf + \mu \cdot \frac{F}{l_c}}{l + \mu} \cdot \mu \quad (6)$$

where *tfn* is the normalized term frequency. l is the document length. F is the frequency of the given query term in the collection. l_c is the number of tokens in the whole collection. μ is the hyper-parameter of the Dirichlet Priors normalization.

We define the model PL3 as Eq. (4), where the normalized term frequency *tfn* is given by the above Dirichlet Priors normalization, That is Eq. (6). Amati [2003] denotes the Dirichlet Priors normalization as normalization 3. In this article, we follow his notation and denote a Dirichlet Priors-based weighting model M as M3.

2.2 Divergence from Randomness Term Weighting Models

The Divergence from Randomness (DFR) term weighting models have been applied for pseudo-relevance feedback. In these models, the importance of a query term is measured by the divergence of its distribution in a pseudo-relevance document set from a random distribution [Amati 2003].

One of the best-performing and most robust DFR weighting models is the Bo1 model [Amati 2003], which is based on the Bose–Einstein statistics. Using this model, the weight of a term t in the exp_doc top-ranked documents is given by:

$$w(t) = tf_x \cdot \log_2 \frac{1 + P_n}{P_n} + \log_2(1 + P_n), \tag{7}$$

where tf_x is the frequency of the query term in the *exp_doc* top-ranked documents. *exp_doc* usually ranges from 3 to 10 [Amati 2003]. P_n is given by $\frac{F}{N}$, where F is the term frequency of the query term in the collection and N is the number of documents in the collection.

The modified query term weight qtw_m is given by the following parameter-free formula [Amati et al. 2004]:

$$qtw_m = qtw + \frac{w(t)}{M} \tag{8}$$

where qtw is the original query term weight. $w(t)$ is the weight of the query term that is given by the Bo1 model. M is the theoretically upper bound of the weight of a term in the *exp_doc* top-ranked documents. It is computed by:

$$\begin{aligned} M &= \lim_{F \longrightarrow F_{max}} w(t) \\ &= F_{max} \log_2 \frac{1 + P_{max}}{P_{max}} + \log_2(1 + P_{max}), \end{aligned} \tag{9}$$

where F is the frequency of the query term in the whole collection, and F_{max} is the frequency of the query term with the highest $w(t)$ weight in the top-ranked documents. P_{max} is given by F_{max}/N. Equation (8) is based on the same idea of Rocchio's query term reweighing formula [Rocchio 1971]. It is parameter-free in the sense that the parameter in the latter query term reweighing formula has been omitted [Amati et al. 2004].

3. TACKLING THE QUERY DEPENDENCE PROBLEM

As introduced in Section 1, in this article, we suggest that reweighing the query terms can help smooth the query dependence. Our solution for the query dependence problem allows for the retrieval performance, obtained using the same parameter settings, to be correlated for different types of queries. Based on the query reweighing, we also hypothesize that the correlation of the normalized term frequency with the document length can indicate the optimal hyper-parameter setting, which smooths the collection dependence. Our solution for the latter problem leads to a methodology that automatically sets the hyper-parameters. Therefore, we start our study with the former problem.

In this section, we explain the query dependence problem by studying the effect of query length, that is, the number of unique non-stopwords in a query, on the *tf* normalization hyper-parameter setting. Our study involves normalization 2, BM25's normalization and the Dirichlet Priors normalization. We discuss the effect of query length on the setting of the hyper-parameters in Section 3.1, and apply a query reweighing method that deals with the query dependence problem in Section 3.2. In Section 3.3, we evaluate the query reweighing method through extensive experiments. In Section 3.4, we further explain our approach by conducting additional experiments.

3.1 The Effect of Query Length

The optimal hyper-parameter settings vary with different query types. For example, when using normalization 2, in general, $c = 7$ provides the optimal performance for short queries, while $c = 1$ provides the optimal performance for long queries [Amati and Van Rijsbergen 2002; Amati 2003]. As we can see from its formula, that is, Eq. (5), compared with $c = 1$, $c = 7$ results in less variation of the term frequency, since tfn/tf is closer to 1. Our explanation is that a short query usually consists of highly informative query terms. When the query becomes longer, it is “contaminated” by the noninformative query terms. Although the query term weight is meant to reflect the importance of a term in the query, it accounts only for the query term frequency and it is still not adequate to differentiate informative query terms from other query terms. As a consequence, compared with a short query, a long query requires more variation in its query term frequencies, to reduce the effect of the noninformative query terms on the document ranking.

Therefore, the change in the optimal setting for short and long queries is due to the fact that the original query term weight cannot adequately reflect the importance of a term in the query. However, we suggest that if we reweigh the terms in both short and long queries so that we achieve an adequate query term weighting, then both short and long queries would require a uniform normalization.

In the next section, we describe our query reweighing method that is based on the Divergence From Randomness (DFR) term weighting models.

3.2 Query Reweighing Using Divergence from Randomness Term Weighting Models

As suggested above, there is a need to reweigh the query terms. A popular approach is to apply a pseudo-relevance feedback technique [Buckley et al. 1992; Efthimiadis 1996]. A pseudo-relevance feedback technique reweighs the query terms, and it refines the query by considering the distribution of the query terms in a relevance document set. The refined query term weights can represent better the importance of the query terms than the original query term weights.

Although we could use any term weighting method for query reweighing, in this article, we use the Bo1 Divergence From Randomness (DFR) term weighting model to reweigh query terms. The reason for using Bo1 is that it is one of the best-performing and stable DFR term weighting models [Amati 2003]. Given a query, if we do a first-pass retrieval, the Bo1 model measures the divergence of the term’s distribution in the top-ranked documents, that is, the assumed relevance set, from its distribution in the whole collection. This divergence measure reflects the degree to which a term is related to the topic. The larger the divergence is, the closer the term is related to the topic. Using the Bo1 model, our query reweighing method is described as follows:

—Given a query, we do a first-pass retrieval for the query. If the query is short, the composing query terms are usually very informative. Therefore, in the first-pass retrieval, we take into account all the query terms of a short query. On the other hand, if the query is long, only a few informative terms are closely related to the topic, while the other query terms are not as important as those informative terms. In this case, we rank the documents with respect to the most informative query terms. The reason for taking only the most informative terms in the first-pass retrieval is that the retrieval performance of the second-pass retrieval is correlated with that of the first-pass retrieval. Using the noninformative terms in the first-pass retrieval, the reweighed query would still suffer from the query dependence brought by those noninformative terms used in the first-pass retrieval (see Section 3.4). In this article, we consider the query terms with the lowest document frequencies as the most informative ones, and ignore other query terms in the first-pass retrieval. The lowest document frequencies indicate the highest inverse document frequencies. Therefore, if the query has more than *qls* unique terms, we do the document ranking for the *qls* terms with the lowest document frequencies; otherwise, we rank the documents with respect to all the query terms. *qls* is a constant discriminating short and long queries. It can be set to any small integer and is robust in terms of smoothing the query dependence.[1] Since a short query usually consists of no more than 5 unique terms (for instance, 83.45% of the title-only queries that are used in our study consist of no more than 5 unique terms.), *qls* is arbitrarily fixed to 5 in this article.

—We modify the original query term weights of all of the query terms using the Bo1 term weighting model as introduced in Section 2.2. Then, we perform the second-pass retrieval with *all* of the query terms.

The above query term reweighing method is based on the pseudo-relevance feedback mechanism described in Section 2.2. The only difference is that in our reweighing method, the first-pass retrieval is done with the most informative terms in a query, while it is done with all the query terms in the pseudo-relevance feedback mechanism.

In Section 3.3, we run experiments to validate whether the optimal setting is indeed consistent for both the short and long reweighed queries. In Section 3.4, for long queries, we conduct additional experiments using all of the query terms in the first-pass retrieval in order to test the impact of the noninformative terms in the first-pass retrieval on the second-pass retrieval in terms of query dependence.

3.3 Experiments

In this section, we experiment on four TREC collections. The aim of the experiments is not to set the hyper-parameters of the normalization methods, but to validate whether query reweighing allows the retrieval performance, obtained using the same parameter settings, to be correlated for different types of reweighed queries.

[1] We have tested different small *qls* values in our experiments. Changing *qls* has little impact on smoothing the query dependence. In this paper, we report only the experiments using $qls = 5$ since experiments using other *qls* values do not provide additional insight.

Table I. Details of the Four TREC Collections and Their Associated Queries Used in Our Experiments

	disk1&2	disk4&5	WT2G	WT10G
Topics	51–200	301–450 and 601–700	401–450	451–550
$avql_s$	5.35	2.76	2.48	4.22
$avql_l$	102.29	60.75	53.96	43.04

The second row gives the number of topics associated to each collection. $avql_s$ and $avql_l$ are the average query lengths (i.e. the number of non-stopwords) of short and long queries, respectively.

3.3.1 *Experimental Environment.* Our experiments are conducted using Terrier [Ounis et al. 2006]. Terrier is a modular platform for the rapid development of large-scale Information Retrieval applications, providing indexing and retrieval functionalities. It is developed at the University of Glasgow.[2]

The four collections used are the disk1&2, disk4&5 (minus the Congressional Record on disk4) of the classical TREC collections,[3] and the WT2G [Hawking et al. 1999] and WT10G [Hawking 2000] Web collections. The test queries are TREC topics that are numbered from 51 to 200 for the disk1&2, from 301 to 450 and from 601–700 for the disk4&5, from 401 to 450 for the WT2G, and from 451 to 550 for the WT10G, respectively (see Table I).

Each TREC topic consists of three fields, that is, title, description and narrative. In this article, we experiment with two types of queries with respect to the use of different topic fields, in order to check if our hypothesis for the query dependence problem stands. The two types of queries are:

- **Short queries**: Only the titles are used.
- **Long queries**: All of the three fields (title, description and narrative) are used.

We could have used other combinations of the topic fields. However, the two query types used represent the two extremes of query length, which allow to show the query dependence problem.

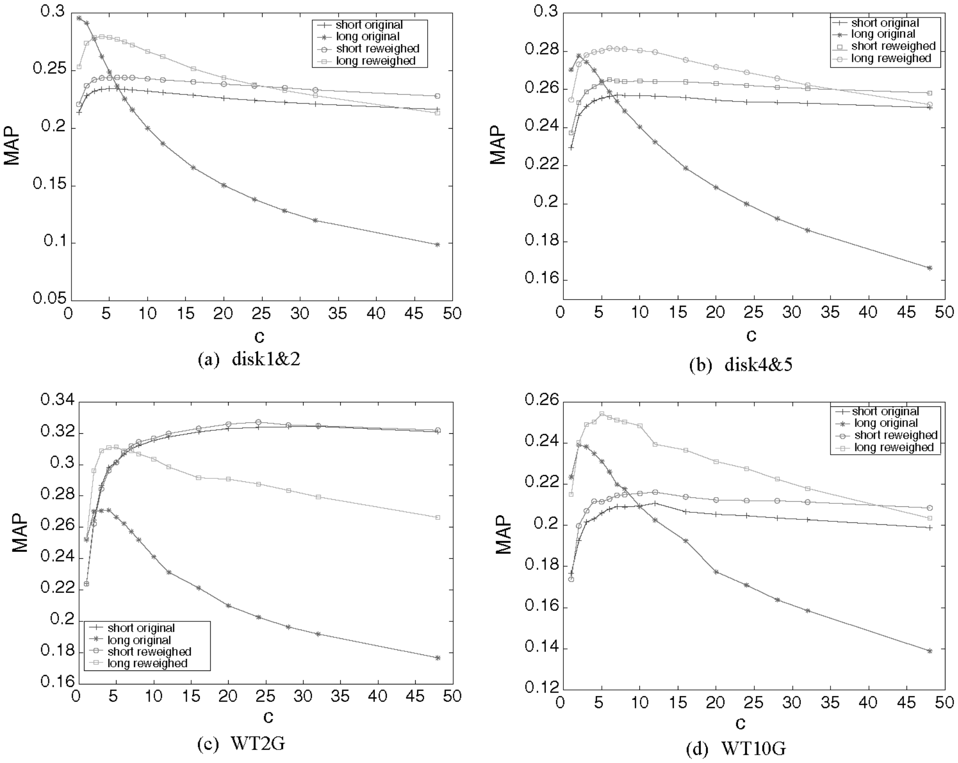
We reweigh the query terms using the *exp_doc* top-ranked documents. *exp_doc* usually ranges from 3 to 10 [Amati 2003]. In this section, we arbitrarily fix *exp_doc* to 5. Changing *exp_doc* will change the absolute retrieval performance, For example, the mean average precision, but has little impact on the effectiveness in smoothing the query dependence.

Table I lists the TREC topics used, and the average query length of each topic set of the two different query types. As shown in Table I, on average, the long queries have many more non-stopwords than the short queries, and the two query types are indeed very different in terms of query length.

By smoothing the query dependence, there should be a high positive correlation between the retrieval performance obtained for short and long reweighed queries. Therefore, for a given collection with a set of queries, we run the retrieval process for a reasonably large range of hyper-parameter values and

[2] More information about Terrier can be found at the following URL: http://ir.dcs.gla.ac.uk/terrier/.

[3] Related information of disk1&2 and disk4&5 of the TREC collections can be found from the following URL: http://trec.nist.gov/data/docs_eng.html.



(a) disk1&2

(b) disk4&5

(c) WT2G

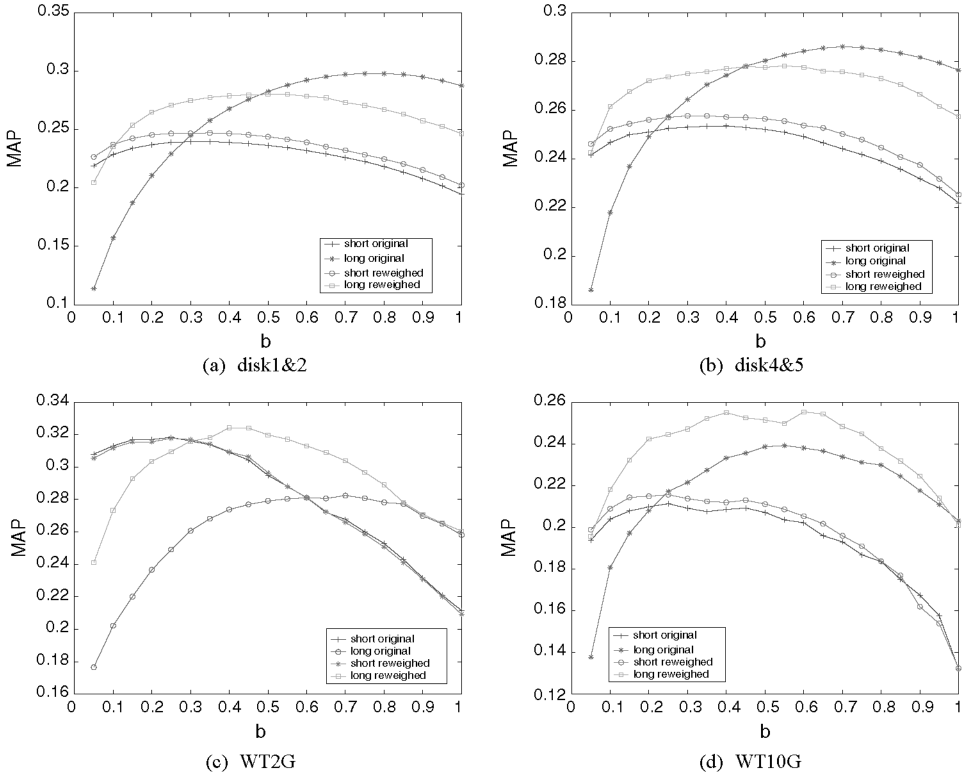
(d) WT10G

Fig. 1. Mean average precision (MAP) against the *tf* normalization hyper-parameter for both original and reweighed queries on the four TREC collections. Results are obtained using *PL2*. On WT2G, the two curves for short queries are almost undistinguishable.

obtain the mean average precision measures, which represent the retrieval performance, using relevance assessment. For normalization 2, the various different hyper-parameter values applied range from 0 (larger than 0) to 48. For BM25's normalization, the applied hyper-parameter values range from 0 (larger than 0) to 1 with an interval of 0.05. For normalization 3, the various different hyper-parameter values applied range from 0 (larger than 0) to 10000. Then, we compute the Spearman's correlation between the MAP measures obtained for short and long queries. The correlations are computed for the original and the reweighed queries, respectively. A high positive correlation for the reweighed queries indicates that our query reweighing method overcomes successfully the query dependence. The experiments are done for PL2, BM25 and PL3, respectively, so that the query reweighing method is tested with all of the three normalization methods which are the normalization 2, BM25's normalization and the Dirichlet Priors normalization.

In all our experiments, stopword removal and the Porter's stemmer are applied. In the next section, we present the results.

3.3.2 *Results.* As shown in Figures 1, 2 and 3, we plot the mean average precision against the hyper-parameter setting. We can see that the curves of



(a) disk1&2

(b) disk4&5

(c) WT2G

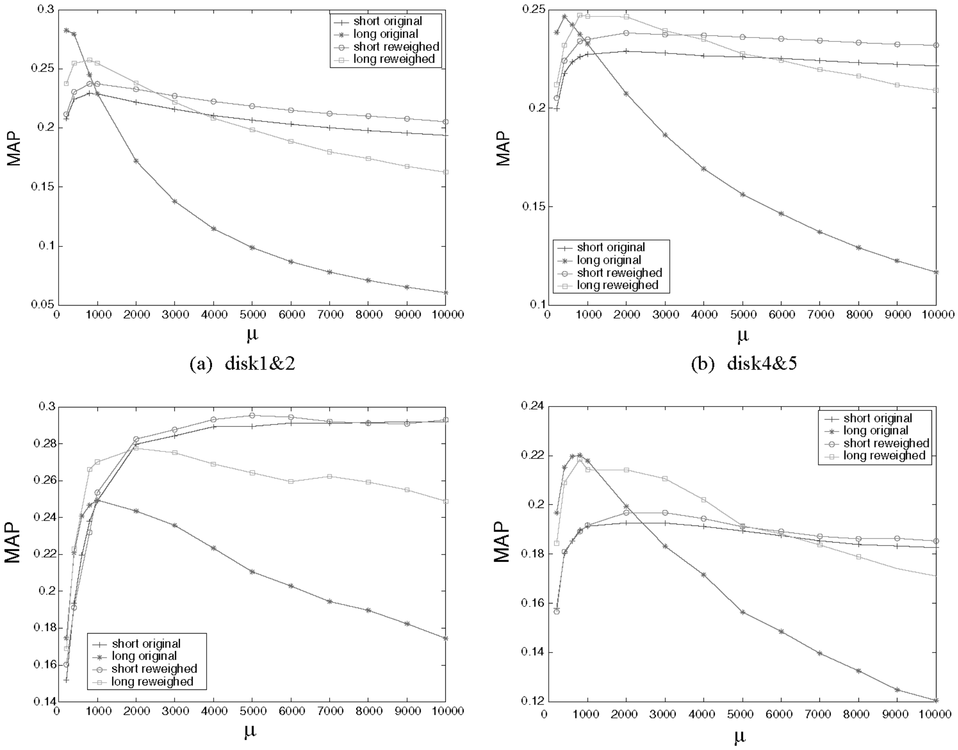
(d) WT10G

Fig. 2. Mean average precision (MAP) against the *tf* normalization hyper-parameter for both original and reweighed queries on the four TREC collections. Results are obtained using *BM25*. On WT2G, the two curves for short queries are almost undistinguishable.

the short and long queries behave very differently for the original queries, while they behave similarly when the query reweighing is applied. Moreover, as shown in the figures, the optimal settings of the original queries depend on both collections and queries in that the peak points of the curves are different for the two different query types and on the four collections used. This illustrates the two problems of setting the *tf* normalization hyper-parameters.

To do a quantitative analysis, we compute the Spearman's correlation between the curves of short and long queries, with and without applying the query reweighing process, respectively. A high positive correlation indicates a consistent retrieval performance obtained using a particular hyper-parameter setting for both short and long queries, and a low positive or a negative correlation indicates the contrary. Table II contains the results for the three applied weighting models. The obtained Spearman's correlation measures allow for the following observations to be made:

—In general, the query reweighing leads to a high correlation (see the ρ_r values in Table II) between the mean average precision obtained for short and long queries, which indicates a more stable retrieval performance with respect



(a) disk1&2

(b) disk4&5

(c) WT2G

(d) WT10G

Fig. 3. Mean average precision (MAP) against the *tf* normalization hyper-parameter for both original and reweighed queries on the four TREC collections. Results are obtained using *PL3*.

Table II. The Spearman's Correlation Measures Between the Mean Average Precision for Short Original and Long Original Queries (ρ_o), and Between the Mean Average Precision for Short Reweighed and Long Reweighed Queries (ρ_r)

	PL2		BM25		PL3	
coll.	ρ_o	ρ_r	ρ_o	ρ_r	ρ_o	ρ_r
disk1&2	0.4537	0.7846	−0.4868	0.6124	0.8681	0.9986
disk4&5	−0.07647	0.8257	−0.3429	0.6436	0.05824	0.5934
WT2G	−0.8779	−0.4588	−0.4767	0.4981	−0.5242	0.1264
WT10G	−0.08529	0.3551	0.07368	0.4286	0.3275	0.6126

to the hyper-parameter setting for the reweighed queries than the original queries (see the ρ_o values in Table II).

—The query reweighing works particularly well on the disk1&2, disk4&5 and WT10G collections, providing high positive correlations between the mean average precision for short and long queries (see the ρ_r values for the three collections in Table II). On the WT2G collection, for PL2 and PL3, although the correlation values for the reweighed queries are not high positive values (which are -0.4588 and 0.1264, respectively), compared to the original queries, the reweighed queries still have much higher correlation values.

Table III. Optimal Settings and the Obtained Mean Average Precision (MAP) Measures for Both Original and Reweighed Queries. Results are Given by *PL2*

	Original Query				Reweighed Query			
	Short		Long		Short		Long	
coll.	c	MAP	c	MAP	c	MAP	c	MAP
disk1&2	7	0.2338	1	0.2958	6	0.2445	4	0.2919
disk4&5	7	0.2570	2	0.2777	6	0.2652	6	0.2907
WT2G	32	0.3282	4	0.2709	32	0.3248	7	0.3197
WT10G	12	0.2108	2.2	0.2404	12	0.2178	5	0.2595

Table IV. Optimal Settings and the Obtained Mean Average Precision (MAP) Measures for Both Original and Reweighed Queries. Results are Given by *BM25*

	Original Query				Reweighed Query			
	Short		Long		Short		Long	
coll.	b	MAP	b	MAP	b	MAP	b	MAP
disk1&2	0.35	0.2398	0.75	0.2472	0.35	0.2473	0.55	0.2908
disk4&5	0.40	0.2535	0.70	0.2577	0.35	0.2581	0.55	0.2876
WT2G	0.25	0.3183	0.70	0.3177	0.25	0.3180	0.40	0.3281
WT10G	0.25	0.2113	0.55	0.2156	0.25	0.2156	0.60	0.2562

Table V. Optimal Settings and the Obtained Mean Average Precision (MAP) Measures for Both Original and Reweighed Queries. Results are Given by *PL3*

	Original Query				Reweighed Query			
	Short		Long		Short		Long	
coll.	μ	MAP	μ	MAP	μ	MAP	μ	MAP
disk1&2	800	0.2292	200	0.2825	800	0.2373	800	0.2671
disk4&5	1500	0.2297	400	0.2465	2000	0.2382	800	0.2511
WT2G	10000	0.2919	1000	0.2495	5000	0.2954	2000	0.2831
WT10G	2000	0.1927	800	0.2201	3000	0.1976	800	0.2217

—The obtained correlation values across the four collections are more stable when using BM25 than when using PL2 and PL3.

The purpose of the experiments in this section is to validate whether we can tackle the query dependence problem using the query reweighing method. However, we still report the manually obtained optimal settings and the corresponding mean average precision using PL2, BM25 and PL3 in Tables III, IV and V, respectively. For the reweighed queries, although the optimal hyper-parameter settings for short and long queries are not identical, we have observed high positive correlations between the retrieval performance of the reweighed short and long queries in most cases (see Table II), which validates our hypothesis.

To summarize, in this section, we have studied the query dependence problem of term frequency normalization. The query dependence problem is due to the diversity of query length, which leads to an inadequate query term weighting. By reweighing the query terms, using the same parameter settings, we can achieve a correlated retrieval performance for different types of queries. Overall, the query reweighing process overcomes the query dependence problem of the term frequency normalization hyper-parameter setting.

Table VI. The Spearman's Correlation Measures Between the Mean Average Precision for Short Original and Long Original Queries (ρ_o), and Between the Mean Average Precision for Short Reweighed and Long Reweighed Queries Using the Most Informative Terms (ρ_r) and all of the Query Terms (ρ_a), Respectively, in the First-Pass Retrieval

coll.	ρ_o	ρ_r	ρ_a
		PL2	
disk1&2	0.4537	0.7846	0.2794
disk4&5	−0.07647	0.8257	0.2588
WT2G	−0.8779	−0.4588	−0.5794
WT10G	−0.08529	0.3551	−0.04779
		BM25	
disk1&2	−0.4868	0.6124	0.3128
disk4&5	−0.3429	0.6436	0.2861
WT2G	−0.4767	−0.02707	0.4981
WT10G	0.07368	0.4286	0.4485
		PL3	
disk1&2	0.8681	0.9986	0.6000
disk4&5	0.05824	0.5934	−0.189
WT2G	−0.5242	0.1264	0.2571
WT10G	0.3275	0.6126	0.6055

In the next section, for long queries, we conduct experiments using all the query terms in the first-pass retrieval to test the impact of the noninformative terms used in the first-pass retrieval on the second-pass retrieval.

3.4 Discussion

In addition to the experiments that use the most informative query terms in the first-pass retrieval, in this section, we run additional experiments using all query terms in the first-pass retrieval. As mentioned in Section 3.2, the use of the noninformative terms in the first-pass retrieval can affect the second-pass retrieval in terms of query dependence. Therefore, we expect lower correlations between the MAP measures obtained for short and long reweighed queries than the ones obtained using the most informative query terms in the first-pass retrieval.

In Table VI, we provide the correlations between the MAP values for short and long queries using all query terms in the first-pass retrieval. Table VI provides again the ρ_o and ρ_r values, taken from Table II, for an easy comparison of the correlation values. From Table VI, we observe that in most cases, using all query terms in the first-pass retrieval, the obtained correlations are higher than those of the original queries, but lower than those of the reweighed queries using the most informative query terms. This observation is expected since we have suggested that the reweighed queries would still suffer from the query dependence brought by the noninformative query terms used in the first-pass retrieval.

Note that using only the most informative query terms in the first-pass retrieval, the reweighed queries still suffer from the collection dependence problem. For example, using PL3, the optimal parameter setting for the short reweighed queries has a large variation over the four collections, which are 800, 2000, 5000, 3000, respectively (see column μ of the short reweighed queries in Table V). Using the optimal setting on the disk1&2 collection (i.e., $\mu = 800$) for the WT2G collection, the obtained MAP is 0.2331, which is statistically significantly lower than the optimised MAP (i.e., 0.2954) with a p-value of 4.538e-06 according to the Wilcoxon test.

In the next section, we extend our study to the collection dependence problem of term frequency normalization.

4. TACKLING THE COLLECTION DEPENDENCE PROBLEM

In this section, we deal with the collection dependence problem, based on measuring the correlation (ρ) of the normalized term frequency (*tfn*) with the document length (l) for a given query term. The collection dependence problem refers to the variation of the optimal term frequency normalization hyper-parameter setting on diverse collections. As a consequence, using the optimal hyper-parameter obtained on one collection on other collections is not a reliable approach.

We provide a statistical understanding of the collection dependence problem, and propose a hypothesis in Section 4.1, on which the proposed automatic hyper-parameter setting methodology in Section 4.2 is based. The proposed setting methodology tackles both query dependence and collection dependence problems. Our hypothesis for the collection dependence problem is built on the query reweighing process. We conduct experiments to validate our hypothesis for the collection dependence problem and evaluate the proposed automatic setting methodology in Section 4.3. A further discussion of our approach and its hypothesis is provided in Section 4.4.

4.1 Correlation of the Normalized Term Frequency with Document Length

As mentioned in Section 1, the aim of term frequency normalization is to smooth the dependence between the document length and the term frequency, where the dependence can be inferred by the correlation of the two variables.

In this article, following the definition of correlation in DeGroot [1989], the correlation of the normalized term frequency with document length is given by:

$$\rho(tfn, l) = \frac{COV(tfn, l)}{\sigma(tfn)\sigma(l)}, \tag{10}$$

where COV stands for covariance. $\sigma(l)$ is the standard deviation of the length of the documents containing the given query term. $\sigma(tfn)$ is the standard deviation of the normalized term frequency of the given query term in all the documents containing the term.

tf normalization can be seen as the *tf* density estimation of the document length. On different collections, the ideal *tf* density function should result in similar correlation measures of *tfn* with the document length. Since the document length distribution depends on the collection, and the optimal hyper-parameter setting depends on the document length distribution, different collections have different optimal hyper-parameter settings. Therefore, for the collection dependence problem, we assume the following hypothesis:

> Hypothesis:
> On different collections, the optimal term frequency normalization hyper-parameter setting provides similar average correlation of the normalized term frequency with the document length for a given set of reweighed query terms.

We base the above hypothesis on the query reweighing method in order to avoid the query dependence problem.

In this article, for a set of given queries, we apply a particular hyper-parameter setting for all the query terms, therefore, we use the mean of the $\rho(tfn, l)$ of the query terms to represent the dependence between document length and the normalized term frequency for the set of queries. Ideally, we can tune the hyper-parameter on a query term-basis. However, we believe that the use of the mean $\rho(tfn, l)$ value is adequate because of the following reason: since a query term normally follows a random distribution in the documents containing itself, on a particular collection, the mean $\rho(tfn, l)$ value of a set of query terms should be stable.

Based on the above proposed hypothesis for the collection dependence problem, we propose an automatic hyper-parameter setting methodology in the next section.

4.2 Automatic Hyper-Parameter Setting

We have hypothesized that on different collections, the optimal hyper-parameter settings provide similar $\rho(tfn, l)$ values. Therefore, we can use the $\rho(tfn, l)$ to indicate the optimal hyper-parameter setting. Based on this hypothesis and our hypothesis for the query dependence problem, we propose an automatic hyper-parameter setting methodology as follows:

(1) We assume a constant optimal mean $\rho(tfn, l)$ value, given by the optimal *tf* normalization hyper-parameter settings, over diverse collections.

(2) We obtain the constant, that is, the optimal mean $\rho(tfn, l)$ value, on a training collection using relevance assessment.

(3) For a given new collection with a set of queries, we reweigh the queries using the approach described in Section 3.2.

(4) We apply such a hyper-parameter setting that it gives the optimal mean $\rho(tfn, l)$ for all the terms in the given set of queries.

In other words, for the application of the above parameter setting methodology, we need a training collection to obtain the optimal mean $\rho(tfn, l)$. The training process needs to be done only once. For a given new collection, we need an appropriate set of queries, though no associated relevance judgement. Finally, we apply the parameter setting that results in the optimal mean $\rho(tfn, l)$ on the given query set, instead of optimizing the hyper-parameter by maximizing the retrieval performance using relevance judgment. Note that our proposed approach does not remove the hyper-parameter but merely substitutes it with the mean $\rho(tfn, l)$.

As we suggested in the previous section, *tf* normalization can be seen as the *tf* density estimation of the document length. The optimal hyper-parameter settings should result in an “ideal” correlation of *tf* with document length. Therefore, it is expected that the optimal mean $\rho(tfn, l)$ is independent of different *tf* normalization methods. The obtained experimental results will confirm it.

In the next section, we evaluate the proposed automatic hyper-parameter setting methodology, as well as its underlying hypothesis of Section 4.1.

4.3 Experiments

The experimental environment in this section, including the collections and queries used, weighting models, stopword removal, and stemming algorithm, is the same as the one we used in Section 3.3. In particular, we conduct a 3-fold holdout evaluation to validate our hypothesis for the collection dependence problem and evaluate the proposed automatic hyper-parameter setting methodology.

In each fold, we use one collection for training and the other three for evaluation. The three training collections used are the disk1&2, disk4&5 and WT10G collections. The WT2G collection is not used as a training collection due to its relatively small number of available associated queries. However, we still use this collection for the evaluation. On each collection, we compare the average of the mean average precision (MAP) obtained in each fold of the holdout evaluation with the best manually obtained MAP by data sweeping. If the difference in the retrieval performance between the two hyper-parameter settings is insignificant, our hypothesis for the collection dependence problem is shown to stand, and the proposed hyper-parameter setting methodology is successful.

As mentioned in the previous section, different *tf* normalization methods are expected to have similar optimal mean $\rho(tfn, l)$ values. We will confirm this in the experiments later on. Therefore, for the three weighting models used, which apply three different normalization methods, we only compute the optimal mean $\rho(tfn, l)$ value for the PL2 model. The obtained optimal mean $\rho(tfn, l)$ value for PL2 is supposed to be the same as those for BM25 and PL3. On a given new collection, we use the optimal mean $\rho(tfn, l)$ obtained for PL2 to estimate the hyper-parameter setting of all the three weighting models used. An insignificant difference in the retrieval performance between the estimated setting and the best manually obtained setting would indicate that our methodology is independent of the three *tf* normalization methods used.

Table VII contains the optimal mean $\rho(tfn, l)$ values, corresponding to the optimal hyper-parameter settings, on each collection using three different weighting models. These optimal hyper-parameter settings were obtained in Section 3.3.2. In general, the optimal mean $\rho(tfn, l)$ values are similar across various collections and weighting models.

Tables VIII, IX and X present the results of the 3-fold holdout evaluation for the three applied weighting models, that is, PL2, BM25 and PL3, respectively. In the tables, AvgMAP and OptMAP are the averages of the mean average

Table VII. The Optimal Mean $\rho(tfn, l)$ Values on Each Collection Using Different Weighting Models. The Correlation Values in *italic* are Used for the Hyper-Parameter Estimation

	PL2	BM25	PL3
disk1&2	*−0.0884*	−0.0840	−0.0852
disk4&5	*−0.0957*	−0.0874	−0.0823
WT2G	−0.1314	−0.0834	−0.0626
WT10G	*−0.1011*	−0.0972	−0.0908

Table VIII. Results Obtained Using *PL2* on the Four TREC Collections

Collection	disk1&2	disk4&5	WT2G	WT10G
	Short			
AvgMAP	0.2440	0.2637	0.3209	0.2158
OptMAP	0.2445	0.2652	0.3248	0.2178
diff	0.20%	0.94%	1.20%	0.92%
p-value	0.0975	0.0003*	0.4840	0.4809
AvgDiff	0.82%			
	Long			
AvgMAP	0.2915	0.2897	0.3148	0.2451
OptMAP	0.2919	0.2907	0.3197	0.2595
diff	0.14%	0.34%	1.28%	5.55%
p-value	0.0302*	0.0484*	0.7944	0.4809
AvgDiff	1.83%			

Table IX. Results Obtained Using *BM25* on the Four TREC Collections

Collection	disk1&2	disk4&5	WT2G	WT10G
	Short			
AvgMAP	0.2473	0.2578	0.3179	0.2147
OptMAP	0.2473	0.2581	0.3180	0.2156
diff	0	0.12%	≈0	0.42%
p-value	0.2532	0.9661	0.1923	0.8664
AvgDiff	0.14%			
	Long			
AvgMAP	0.2900	0.2837	0.3173	0.2544
OptMAP	0.2908	0.2876	0.3281	0.2562
diff	0.28%	1.36%	3.29%	0.70%
p-value	0.4501	0.2133	0.1476	0.8505
AvgDiff	1.41%			

precision (MAP) obtained in the holdout evaluation process, and the best manually obtained MAP, respectively. diff is the difference between the two MAP measures in percentage. Because OptMAP is the upper bound of the retrieval performance, diff is always positive.

AvgDiff is the average diff over the four collections. A p-value marked with a star indicates a statistically significant difference at 0.05 level according to the Wilcoxon test. The Wilcoxon test is done over two lists of average precision of each query. The first one is the average precision of each query when the

Table X. Results Obtained Using *PL3* on the Four TREC Collections

Collection	disk1&2	disk4&5	WT2G	WT10G
	Short			
AvgMAP	0.2339	0.2329	0.2909	0.1925
OptMAP	0.2373	0.2382	0.2954	0.1976
diff	1.43%	2.23%	1.52%	2.58%
p-value	0.2196	0.1179	0.1812	0.0049*
AvgDiff	1.49%			
	Long			
AvgMAP	0.2587	0.2507	0.2794	0.2038
OptMAP	0.2671	0.2511	0.2831	0.2217
diff	3.14%	0.16%	1.31%	8.07%
p-value	2.603e-8*	0.0528	0.1236	0.0628
AvgDiff	3.17%			

parameter setting is optimal in data sweeping. The second one is the mean of the average precisions of each query in the hold-out process. An insignificant difference indicates a success of the automatic hyper-parameter setting methodology, since its performance does not statistically differ from the best manually obtained setting. As shown by the results in the three tables, on average, the difference between the MAP given by the estimated settings and the manually obtained optimal settings is minor in most cases, indicating a consistent optimal hyper-parameter setting over diverse collections. This insignificant difference also indicates that the optimal mean $\rho(tfn, l)$ value is independent of the three *tf* normalization methods used, which was expected in Section 4.2.

Overall, according to the results of the 3-fold holdout evaluation process, our hypothesis for the collection dependence problem stands for the three normalized methods. The experimental results show that with the use of the query reweighing method, for a given collection, we can apply the hyper-parameter setting such that it gives an optimal mean $\rho(tfn, l)$. Therefore, we have smoothed both the query dependence and collection dependence of term frequency normalization. In addition, the results indicate that the involved estimation process is not only robust, but also practical in the sense that we only need to run the training process on a collection once, and estimate the hyper-parameter settings on other collections without further relevance judgment. Therefore, it has a low overhead. Finally, as expected, the results show that the optimal settings of the three normalization methods lead to similar mean $\rho(tfn, l)$ values. Hence, the proposed setting methodology is independent of the *tf* normalization methods used.

In the next section, we discuss the correlation between the normalized term frequency and the document length and investigate if we can improve our proposed approach by suggesting a new hypothesis.

4.4 Discussion

In this section, first, we look at the $\rho(tfn, l)$ values and relate our findings to a previous work. Second, we study a zero-correlation hypothesis based on

Table XI. The Mean Correlation $\rho(tf, l)$ Between the Raw Term Frequency and the Document Length on the Four Collections

Collection	disk1&2	disk4&5	WT2G	WT10G
$\rho(tf, l)$	0.2738	0.1844	0.1688	0.1204

the obtained experimental results in the previous section. Next, we test this hypothesis through experiments.

In Table XI, we provide the correlation values between the raw term frequency and the document length on the four collections used. From Table XI, we can see that the term frequency has a moderately positive correlation with the document length when *tf* normalization is not applied. Moreover, in our experiments in the previous section, the optimal correlation values on different collections are shown to be small negative (see Table VII). From this, we infer that in the ad-hoc tasks used, when the normalization parameters are optimized, the normalized term frequency does not necessarily increase, and even decrease a little, with the increment of the document length. Our findings can be linked to the work in Singhal et al. [1996], in which Singhal et al. conclude that the document score does increase with document length, but not proportionally. Since the normalized term frequency is positively correlated with the term weight, having a small negative correlation between the normalized term frequency and the document length may imply some positive correlation of the document score with the document length, because a longer document is likely to contain more query terms.

We have mentioned that the purpose of *tf* normalization is to smooth the dependence between the term frequency and the document length. Since in our experiments, the obtained optimal correlation values are small negative and close to zero, we investigate the following zero-correlation hypothesis:

> The Zero-correlation Hypothesis:
> The term frequency normalization parameters are optimal when the correlation between the normalized term frequency and the document length is zero.

The above hypothesis implies that when the *tf* normalization parameters are optimal, there is no correlation between the normalized term frequency and the document length. Compared with our hypothesis in Section 4.1, an advantage of the zero-correlation hypothesis is that its resulting tuning process does not require a training process to obtain the optimal mean correlation, which is assumed to be zero by the zero-correlation hypothesis.

Next, we conduct experiments to validate the above zero-correlation hypothesis. The collections and queries used are the same as in the previous sections. According to the zero-correlation hypothesis, we do not train the optimal mean $\rho(tfn, l)$, but fix it to 0. For a given collection, the estimated parameter values result in the zero correlation between the normalized term frequency and the document length. Next, we compare the mean average precision (MAP) obtained by the estimated parameter value with the best manually obtained MAP using data sweeping. If the difference in the retrieval performance between the two

Table XII. Results Obtained Using *PL2* on the Four TREC Collections for Testing the Zero-Probability Hypothesis

Collection	disk1&2	disk4&5	WT2G	WT10G
	Short			
MAP	0.2239	0.2473	0.3176	0.1878
OptMAP	0.2445	0.2652	0.3248	0.2178
diff	8.42%	6.75%	2.22%	13.77%
p-value	5.599e-07*	2.843e-4	4.892e-2*	1.602e-07*
	Long			
MAP	0.2662	0.2372	0.2686	0.1541
OptMAP	0.2919	0.2907	0.3197	0.2595
diff	8.80%	18.40%	15.98%	40.62%
p-value	2.045e-3*	2.108e-20	5.241e-05*	1.912e-12*

Table XIII. Results Obtained Using *BM25* on the Four TREC Collections for Testing the Zero-Probability Hypothesis

Collection	disk1&2	disk4&5	WT2G	WT10G
	Short			
MAP	0.2362	0.2492	0.3105	0.1926
OptMAP	0.2473	0.2581	0.3180	0.2156
diff	4.49%	3.45%	2.36%	10.67%
p-value	2.23e-4*	3.181e-3*	0.3693	6.961e-05*
	Long			
MAP	0.2507	0.2610	0.2853	0.1900
OptMAP	0.2908	0.2876	0.3281	0.2562
diff	13.79%	9.25%	13.04%	25.84%
p-value	3.469e-09*	1.369e-08*	2.516e-3*	1.191e-4*

hyper-parameter settings is statistically significant, then the zero-correlation hypothesis is not reliable. Otherwise, the zero-correlation hypothesis holds, and we can simplify our tuning methodology as the zero-correlation hypothesis does not require a training process.

Tables XII, XIII and XIV present the experimental results using PL2, BM25 and PL3, respectively. In the three tables, MAP and OptMAP are the mean average precision obtained by the zero-correlation hypothesis and by data sweeping, respectively. diff is the difference between MAP and OptMAP. A p-value marked with a star indicates a significant difference at the 0.05 level according to the Wilcoxon test. From the three tables, we observe that in most of the cases, the zero-correlation hypothesis provides a retrieval performance that is statistically significantly lower than the optimal retrieval performance obtained using data sweeping. Therefore, we conclude that the zero-correlation hypothesis is not as reliable as our hypothesis in Section 4.1. This may suggest that the optimal correlation value is task-oriented. Different search tasks require different ideal correlation values. The training process allows the search task to be taken into account. Overall, we conclude that the purpose of *tf* normalization is to achieve an optimal correlation between the normalized term frequency and the document length for a search task, not to remove the dependence between the two variables.

Table XIV. Results Obtained Using *PL3* on the Four TREC Collections for Testing the Zero-Probability Hypothesis

Collection	disk1&2	disk4&5	WT2G	WT10G
	Short			
MAP	0.2020	0.2318	0.2331	0.1738
OptMAP	0.2373	0.2382	0.2954	0.1976
diff	14.88%	2.69%	21.09%	12.04%
p-value	4.396e-09	0.08177	4.538e-06*	2.886e-07*
	Long			
MAP	0.1799	0.2255	0.2676	0.1552
OptMAP	0.2671	0.2511	0.2831	0.2217
diff	25.16%	10.20%	5.48%	30.00%
p-value	9.633e-19	1.867e-07*	0.02543*	3.642e-07*

5. EXPERIMENTS WITH QUERY EXPANSION

In the previous section, our experiments were conducted using the original query terms—we reweigh the original query terms but without adding new terms into the query. In this section, we test our hypothesis for the collection dependence problem with the use of query expansion. For a given short query, we reweigh the *exp_terms* most informative terms, that is, the terms with the highest weights according to the Bo1 model, from the top *exp_doc* documents. In this section, *exp_doc* and *exp_terms* are set to 3 and 10, respectively, which is the recommended setting for query expansion in Amati [2003]. The expanded query consists of the original query terms and the 10 reweighed terms. If an original query term appears in the 10 reweighed terms, its query term weight is modified using Eqs. (8) and (9). We do not run experiments for long queries because using all the three topic fields, the queries are lengthy enough to render query expansion equivalent to a query reweighing process.

Note that our hypothesis for the collection dependence problem assumes an optimal mean $\rho(tfn, l)$ over collections for the reweighed queries. In this section, we test if this hypothesis still stands for the expanded queries. The experimental methodology is the same as the one we used in Section 4.3. We do a 3-fold holdout evaluation to test our hypothesis for the collection dependence problem with the expanded queries. On each collection, the estimated hyper-parameter setting is such a value that gives the optimal mean $\rho(tfn, l)$. In our experiments in Section 3.3, the optimal mean $\rho(tfn, l)$ is obtained using the reweighed queries. In this section, instead of computing the optimal mean $\rho(tfn, l)$ for the expanded queries, we reuse the values for the reweighed queries (see Table VII). The use of the two kinds of optimal mean $\rho(tfn, l)$ values are equivalent because the terms in both reweighed and expanded queries are reweighed. In accordance with our hypothesis for the collection dependence problem, the optimal mean $\rho(tfn, l)$ should be stable for the reweighed query terms.

In our experiments, we conduct the same comparison as in Section 4.3: we compare the retrieval performance obtained by the automatic hyper-parameter setting methodology with that of the best manual setting. We consider our approach successful if we observe a statistically insignificant difference between their performance.

Table XV. Results Obtained with the Use of Query Expansion

Collection	disk1&2	disk4&5	WT2G	WT10G
	PL2			
AvgMAP	0.2721	0.2919	0.3252	0.2312
OptMAP	0.2734	0.2931	0.3298	0.2369
diff	0.48%	0.41%	1.03%	1.82%
p-value	0.0023*	0.1147	0.6397	0.9985
	BM25			
AvgMAP	0.2770	0.2883	0.3249	0.2396
OptMAP	0.2785	0.2902	0.3327	0.2446
diff	0.54%	0.65%	2.34%	2.04%
p-value	0.2226	0.1483	0.0506	0.3513
	PL3			
AvgMAP	0.2614	0.2591	0.2964	0.2021
OptMAP	0.2617	0.2660	0.3002	0.2126
diff	0.11%	2.59%	1.27%	4.94%
p-value	0.1420	1.486e-5*	0.7647	0.0465*

In Table XV, AvgMAP is the average mean average precision obtained in the 3-fold holdout evaluation and OptMAP is the mean average precision obtained by the best manual setting. diff is the difference between the two MAP measures in percentage. A p-value marked with a star indicates a statistically significant difference at 0.05 level according to the Wilcoxon test.

As shown in Table XV, in general, the retrieval performance of the estimated setting is very close to the performance of the best manually obtained setting. We observe an insignificant difference between the two MAP measures in most cases (see the p-values in Table XV). Therefore, we obtain consistent results with and without the use of query expansion. This observation is expected since the query terms are reweighed in both the query reweighing method and query expansion.

6. CONCLUSIONS

In this article, we have studied the issue of setting the hyper-parameters of three term frequency normalization methods, which are normalization 2, BM25's normalization and the Dirichlet Priors normalization. Our aim was not only to provide a better understanding of the query dependence and collection dependence problems of the hyper-parameter setting, but also to propose an automatic methodology that tackles these two problems.

We have shown that the *tf* normalization hyper-parameter setting reflects the importance of a query term, and that there is a need for smoothing the query term weight. Therefore, we apply a divergence from randomness term weighting model to tackle the query dependence problem.

We have also shown that the dependence between the document length and the term frequency can be inferred by the correlation between the two variables. The optimal hyper-parameter settings provide similar correlation values over diverse collections. This hypothesis for the collection dependence problem is shown to stand according to a 3-fold holdout evaluation. In particular, it is applicable to the three studied normalization methods in this article. Moreover, the experiments also show that the hyper-parameter setting methodology is robust and effective. Finally, we have observed the same conclusions with and without using query expansion.

ACKNOWLEDGMENTS

We would like to thank Gianni Amati and Keith Van Rijsbergen for their advice on this work. We would also like to thank Christina Lioma for reading various versions of this article. Finally, we would like to thank the anonymous reviewers for their useful and constructive comments.

REFERENCES

AMATI, G. 2003. Probabilistic models for information retrieval based on divergence from randomness. Ph.D. dissertation, Department of Computing Science, University of Glasgow.

AMATI, G., CARPINETO, C., AND ROMANO, G. 2004. Fondazione Ugo Bordoni at TREC 2004. In *Proceedings of the 13th Text REtrieval Conference (TREC 2004)* (Gaithersburg, MD).

AMATI, G. AND VAN RIJSBERGEN, C. J. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. In *ACM Trans. Inf. Syst. (TOIS) 20*, 4.

BUCKLEY, C., SALTON, G., AND ALLAN, J. 1992. Automatic retrieval with locality information using Smart. In *Proceedings of the 1st Text REtrieval Conference (TREC-1)* (Gaithersburg, MD).

DEGROOT, M. 1989. *Probability and Statistics*, 2nd edition ed. Addison Wesley, Reading, MA.

EFTHIMIADIS, N. E. 1996. Query expansion. In *Ann. Rev. Inf. Syst. Tech. 31*.

HAWKING, D. 2000. Overview of the TREC-9 Web Track. In *Proceedings of the 9th Text REtrieval Conference (TREC-9)* (Gaithersburg, MD).

HAWKING, D. AND CRASWELL, N. 2001. Overview of the TREC 2001 Web Track. In *Proceedings of the 10th Text REtrieval Conference (TREC-10)* (Gaithersburg, MD).

HAWKING, D., VOORHEES, E., CRASWELL, N., AND BAILEY, P. 1999. Overview of the TREC-8 Web Track. In *Proceedings of the 9th Text REtrieval Conference (TREC-8)* (Gaithersburg, MD).

HE, B. AND OUNIS, I. 2005. A study of the Dirichlet Priors for term frequency normalization. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Salvador, Brazil). ACM, New York.

OUNIS, I., AMATI, G., PLACHOURAS, V., HE, B., MACDONALD, C., AND LIONA, C. 2006. Terrier: A high performance and scalable Information Retrieval platform. In *Proceedings of ACM SIGIR OSIR Workshop 2006* (Seattle, WA).

ROBERTSON, S.E., WALKER, S., BEAULIEU, M. M., GATFORD, M., AND PAYNE, A. 1995. Okapi at TREC-4. In *NIST Special Publication 500-236: The 4th Text REtrieval Conference (TREC-4)* (Gaithersburg, MD).

ROCCHIO, J. 1971. Relevance feedback in information retrieval. Prentice-Hall, Englewood Cliffs, NJ.

SINGHAL, A., BUCKLEY, C., AND MITRA, M. 1996. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Zurich, Switzerland). ACM, New York.

VAN RIJSBERGEN, C. J. 1979. *Information Retrieval, 2nd edition*. Department of Computer Science, Univ. Glasgow.

ZHAI, C. AND LAFFERTY, J. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, LA). ACM, New York.

Received October 2005; revised May 2006, September 2006; accepted January 2007