An Evaluation and Analysis of Incorporating Term Dependency for Ad-hoc Retrieval

Abstract. Although many retrieval models incorporating term dependency have been developed, it is still unclear whether term dependency information can consistently enhance retrieval performance for *different* queries. We present a novel model that captures the main components of a topic and the relationship between those components and the power of term dependency to improve retrieval performance. Experimental results demonstrate that the power of term dependency strongly depends on the relationship between these components. Without relevance information, the model is still useful by predicting the components based on global statistical information. We show the applicability of the model for adaptively incorporating term dependency for individual queries.

1 Introduction

In most existing retrieval models, documents are scored primarily using occurrences of single query terms in documents, assuming that the query terms are independent. However, previous studies have shown that incorporating the dependency of query terms in documents into retrieval strategies can improve average retrieval effectiveness on a fixed set of queries [1][2][3]. Moreover, existing retrieval models incorporating term dependency are far from optimal. One problem of current models is that most proposed methods are uniformly applied to all the queries. In fact, we find that not all the queries benefit from taking account of term dependency. Our experimental results in section 3 show that term dependency models fail to improve retrieval performance for around 50% of queries for adhoc title-only topics in the TREC Terabyte Tracks.

Until now, little investigation has been reported on how term dependency information can enhance retrieval performance on an individual query basis. In this paper, we investigate the main features affecting the power of term dependency to enhance retrieval performance. We suggest a novel model which captures the main components of a topic and the relationship between those components and the power of term dependency. We argue, and then show experimentally, that the power of term dependency depends on the relationship between these components. In practice, we do not have the relevance information, thus we cannot compute the components of the model directly. We show that in such cases the proposed model is still useful by predicting the components based on global statistical information of the collection. Finally, we show that we can adaptively use the term dependency information on an individual query basis by making use of the proposed model.

2 A Model for the Power of Term Dependency

In this section, we propose a model for predicting the power of term dependency. In this work, the power of term dependency refers to the extent to which retrieval models incorporating term dependency can successfully improve retrieval performance for a given query compared to models based on the standard bag-of-words assumption.

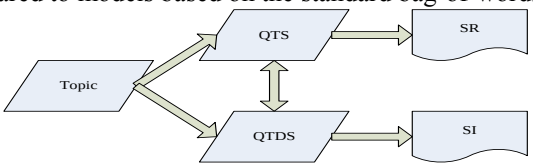

Fig. 1. A general model of a topic based on the *QTS and QTDS.*

The effectiveness of an IR model depends on its ability to distinguish relevant documents for a given query from irrelevant ones. In most existing IR models, the main features used to identify relevant documents are various kinds of term statistics such as with-document frequencies, inverse document frequencies, and document lengths. Obviously, if relevant documents have more occurrences of query terms than irrelevant ones, the query tends to achieve better results because of the high quality of the query's term statistics. In this work, the quality of term statistics (*QTS*) of a query refers to the property of the query that determines to what extent relevant documents can be identified from irrelevant ones based on the term statistics of the query.

When IR models are extended to incorporate term dependency, the quality of term dependency statistics (*QTDS*) such as the occurrences of ordered phrases and unordered phrases becomes interesting. In this work, *QTDS* refers to the property of the query that determines to what extent the relevant documents can be identified based on the term dependency statistics of the query.

Therefore, we define the primary object of the model based on the power of term dependency to be a *Topic*. A topic relates to a defined subject. The topic is comprised of two objects: *QTS* and *QTDS.* The topic is also dependent on the set of relevant documents (*SR*) and the set of irrelevant documents (SI), where *QTS* and *QTDS* are computed based on the gap in either term statistics or term dependency statistics from the relevant to irrelevant set. Thus, we denote a topic as:

$$Topic = (QTS, QTDS \mid SR, SI) \quad (1)$$

Figure 1 shows a schema of the model. The two components *QTS* and *QTDS* have high correlation with the retrieval effectiveness of a given query. When *QTS* or *QTDS* is high, term statistics or term dependency statistics tend to be good features to identify relevant documents. *QTDS* is supposed to be positively correlated with the power of term dependency, because IR models tend to benefit from term dependency when *QTDS* is high. Contrarily, *QTS* is supposed to be negatively correlated with the power. When *QTS* is high meaning that features of term statistics are good identifiers for relevant documents. Thus retrieval effectiveness based on term statistics tends to be high, which makes it harder for term dependency to improve the effectiveness. From the other prospective, when *QTS* is high and *QTDS* is low, term statistics tend to be better features than term dependency statistics. Hence in this situation, IR models

incorporating term dependency cannot achieve better results than IR models only using good term statistics features.

Now we describe our approach of computing *QTS* and *QTDS*. For *QTS*, we compute the average term frequency of query terms (*TF*) for each document in *SR* and *SI*. *QTS* is the division between the median *TF* of relevant documents and irrelevant documents. For *QTDS,* we compute the average occurrences of ordered phrases (*OF*) in a document instead. Details of the definition of ordered phrases are given [1]. *TF* and *OF* are defined as:

$$TF = \frac{\sum_{w \in Q} tf_{w,D}}{|Q|} \quad (2)$$

$$OF = \frac{\sum_{c \in O} tf_{c,D}}{|O|} \quad (3)$$

where Q is the query, w is the term of the query, $tf_{w,D}$ is the term frequency in a document, O is the set of ordered phrases, c is a kind of ordered phrase, $tf_{c,D}$ is the occurrence frequency of the ordered phrase.

In this work, *SR* refers to the set of documents judged relevant and *SI* refers to the set of documents judged irrelevant. Of course, the judged relevant and judged non-relevant documents are heavily biased because of the pooling procedure used at TREC. However, these statistics still provide valuable information.

In practice when entering a new search query, we do not have relevance information, thus we cannot compute *QTS* and *QTDS* directly. We can though predict the components based on global statistical information of the whole collection. For *QTS*, we compute the average inverse document frequency of the terms in the query (*Avg_IDF*) to predict *QTS*. When a query term has a high IDF value, meaning that the term only appears in a small fraction of documents in the collection, irrelevant documents do not have high chance to have the occurrence of the term. Thus high *Avg_IDF* indicates good quality of term statistics. For *QTDS*, we count occurrences of ordered phrases in the data collection (*OO*) and then we compute the average inverse *OO* (*Avg_IOO*) to predict *QTDS*. Occurrences of ordered phrases are strong evidence of relevance. High occurrences of ordered phrases mean that many relevant documents have ordered phrases. Thus high *OO* indicates high *QTDS*. As a result, *Avg_IOO* is supposed to be negatively correlated to *QTDS*.

3 Validating the Model

In this section, we validate our model by showing the correlation between the components of the model and the power of term dependency. We use the TREC .GOV2 Terabyte test collection, and its associated TREC 2004, 2005 and 2006 adhoc title-only topics and relevance assessment sets. 3 out of the 150 topics were removed since there are no relevant documents in the collection or the topic only has

one query term. Thus 147 topics in total were evaluated. For indexing and retrieval we use Indri[1], with Porter's stemming and stop words removal.

Retrieval was performed twice using the full independence (FI) and full dependence (FD) variants of MRF model [1] respectively. FI only uses the term statistics while FD makes use of the term dependency information. The Mean Average Precision (MAP) of FI is 0.2971 while MAP of FD is 0.3298 for the 147 queries. This indicates that incorporating term dependency can improve the average retrieval performance. However, only 87 of 147 queries actually perform better by incorporating term dependency, meaning that the dependence model fails for 41% of them.

The power of term dependency is computed by the division between the Average Precision (AP) of FD and the AP of FI for a given query. We measure the correlation between the components of our proposed model and the power of term dependency by the Spearman rank correlation test, since the power distributions are unknown. The results for correlation are shown in table 1, where bold cases indicate that the results are statistically significant at the 0.05 level.

From these results, we firstly observe that the components of the model *QTS* and *QTDS* significantly correlate with the power of term dependency. Queries of high *QTDS* tend to benefit from term dependency, while for queries with high *QTS* it becomes harder to achieve better results. Secondly, the combination of the model's two components results in higher correlation, suggesting that the two components measure different properties of a topic. Finally, it can be observed that *Avg_IDF* and *Avg_IOO* still work for the model in the absence of relevance information.

Table 1. Spearman correlation coefficients between the components of our model and the power of term dependency. Bold cases indicates that the results are statistically significant at the 0.05 level.

	QTS	*QTDS*	*Combine*	*Avg_IDF*	*Avg_IOO*
Spearman's	**-0.40**	**0.44**	**-0.528**	**-0.18**	**-0.20**

4 Uses of the Model

As shown in the above section, models incorporating term dependency can improve the average retrieval performance on a fixed set of queries. However, the dependency model fails to achieve better results for round 50% of queries, where much more computation resources are required for processing the term dependency information.

Thus, it is not beneficial to use term dependency for every query. Instead, it is advantageous to have a switch that will estimate when term dependency will improve retrieval, and when it would be detrimental to it. In the absence of relevance information, we can use the *Avg_IDF* and *Avg_IOO* of our proposed model to predict whether dependency model can work for a given query. Since both of the two features are statistically negatively correlated to the power of term dependency, queries of

[1] URL: http://www.lemurproject.org/indri/

high *Avg_IDF* and *Avg_IOO* scores tend not to benefit from term dependency. Thus we try to identify those queries for which the dependency models fail by finding queries of high *Avg_IDF* and *Avg_IOO* scores. For these identified queries, we just use the FI model, while for the other queries we use FD model instead. We name the retrieval results "Sel" in table 2 by adaptively using term dependency on a query basis.

In this work, we label queries of high *Avg_IDF* scores, when the scores are ranked in the top 20% of all the 147 queries. We label queries of high *Avg_IOO* scores in the same way. The overlap of labeled queries of high *Avg_IDF* scores and labeled queries of high *Avg_IOO* scores are identified queries for which dependency model is estimated to fail. The low threshold 20% is chosen, because we want to find those queries when term dependency would be detrimental to retrieval.

In total, 11 out of 147 queries were identified by our proposed model. 10 of the 11 identified queries indeed do not benefit from term dependency, which indicates great prediction power of our model. The retrieval results are shown in table 2. Sel has the best retrieval effectiveness among the three models under the performance measures of MAP and Geometric MAP (GAMP).

Table 2. Improvements in retrieval based on our proposed model.

	MAP	GMAP
FI	0.2971	0.2006
FD	0.3298	0.2527
Sel	0.3308	0.2531

5 Summary

This work tries to answer the question of what kind of queries can benefit from term dependency information. We describe a novel model that captures the main components of a topic and the relationship between the components to the power of term dependency. We demonstrate that the power of term dependency strongly depends on those components. Without relevance information, we can predict the components by global statistics information of the index. Finally, we demonstrate the applicability of model to adaptively using the term dependency on a query basis.

References

1. D. Metzler and W. B. Croft: A Markov Random Field Model for Term Dependencies. In Proc. of SIGIR 2005, Brazil. (2005)
2. J. Peng, C. M., B. He, V. P., I. O.: Incorporating Term Dependency in the DRF Framework. In Proc. of SIGIR 2007, The Netherlands. (2007)
3. T. Tao and C. Zhai: An Exploration of Proximity Measures in Information Retrieval. In Proc. of SIGIR 2007, The Netherlands. (2007)