Utilizing Phrase Based Semantic Information for Term Dependency

Yang Xu, Fan Ding, Bin Wang
Institute of Computing Technology, Chinese Academy of Sciences
Beijing,100190,P.R.China
{xuyang,dingfan,wangbin}@ict.ac.cn

ABSTRACT

Previous work on term dependency has not taken into account semantic information underlying query phrases. In this work, we study the impact of utilizing phrase based concepts for term dependency. We use Wikipedia to separate important and less important term dependencies, and treat them accordingly as features in a linear feature-based retrieval model. We compare our method with a Markov Random Field (MRF) model on four TREC document collections. Our experimental results show that utilizing phrase based concepts improves the retrieval effectiveness of term dependency, and reduces the size of the feature set to large extent.

Categories and Subject Descriptors

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

General Terms

Performance, Experimentation

Keywords

Information retrieval, term dependency, Markov random field model, phrase, Wikipedia

1. INTRODUCTION

Term dependency, or co-occurrences has been long of interest to researchers. Recent studies have shown that incorporating term dependency features into retrieval models can lead to significant improvements for ad-hoc retrieval tasks [1, 4]. However, most previous work on modeling term dependencies does not take the concept or ideas underlying the query phrases into account. This contradicts the intuitive understanding that information need is closely related to the meaning of query phrases.

The aim of this work is to study the impact of utilizing phrase based concepts for term dependency. A linear feature-based model provides a way of combining evidence from a wide range of textual features, we implement our idea on the basis of a Markov random field (MRF) model for information retrieval, one of the state-of-the-art linear feature-based models [3, 4].

Copyright is held by the author/owner(s).
SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

In original the MRF model, any sub-phrases of a query, either ordered or unordered, are used as features to indicate document relevance. Because the number of features is exponential in the number of query terms, it limits the application of this model to shorter queries.

To cope with the problem, we propose a method to distinguish important term dependencies from less important ones, thus to treat them accordingly for retrieval tasks. We define *semantically meaningful phrases (SMP)* and *nonsemantically meaningful phrases (NMP)*. For example, in the query *school uniform public school* (TREC topic533),*school school*, *uniform public*, *school uniform*, *public school* all are candidates for term dependency features. However, from the human judgement, documents with *school uniform* or *public school* are more relevant than documents with *school school* or *uniform public*. In this example, the latter two are *SMP*, but the first two are not.

We test our method on TREC collections of varying sizes and types. Experimental results show that our method is either significantly more effective than, or equally effective to the standard MRF model, and reduces the computing complexity by lowering the number of features.

2. RELATED WORK

Markov random fields are undirected graphical models. The MRF model for IR models the joint distribution over a query $Q = q_1, ..., q_n$ and a document D. A MRF is defined by a graph G and a set of non-negative potential functions over the cliques in G. The nodes in the graph represent the random variables and the edges define the independence semantics of the distribution. Given a graph G, a set of potentials ψ_i, and a parameter vector Λ, the joint distribution over Q and D is given by :

$$P_{G,\Lambda}(Q, D) = \frac{1}{Z_\Lambda} \prod_{c \in C(G)} \psi(c; \Lambda)$$

where Z_Λ is a normalizing constant and $C(G)$ is the set of cliques in G. The potentials are parameterized as $\psi(c; \Lambda) = exp[\lambda_i f_i(c)]$, where $f_i(c)$ is a real-valued feature function, and λ_c is the weight given to that particular feature function. Under this parameterization, documents are ranked in descending order according to $P(D|Q)$, which can be shown to be rank equivalent to :

$$P_\Lambda(D|Q) \stackrel{rank}{=} \sum_{c \in C(G)} \lambda_c f_c(c)$$

Therefore, the ranking function is a weighted linear combination of features defined over the cliques of the MRF.

In [4] the MRF model uses three clique sets including *single terms*, *ordered terms* and *unordered terms*. We will compare our method with this setting (MRF-IR) in section 4.

3. RECOGNIZE SMP

Several characteristics can be used to measure the difference between important term dependency and less important ones. In this work, we focus on whether a term dependency is associated with certain concepts or not. This can be done by recognizing phrases that are frequently used in daily language. Note that these meaningful phrases might not be necessarily formally defined in dictionaries; thus we use Wikipedia [6]. Wikipedia is a free online encyclopedia edited collaboratively by large numbers of volunteers (web users). The exponential growth and the reliability of Wikipedia make it a suitable tool for this task. Each article in Wikipedia is uniquely identified by its title, which is the most common name for the content described in the article. We refer to these titles as *Wikipedia Concept*.

We downloaded the English Wikipedia dump of January 7th, 2008. In addition to the "proper" encyclopedia pages we also index redirect pages and various log-pages. We have 6,202,531 pages (and also titles). All the experiments make use of the Indri search engine [2]. Both Wikipedia titles and queries are in lower case. Stopping is done for each query. The method of recognizing *SMP* is as follows:

is_wiki_SMP(p) : p represents a subset of query terms that appear ordered in the query. When p is submitted to Wikipedia, if there is a *Wikipedia Concept* containing p, then p is *SMP*, otherwise p is *NMP*.

Our method recognizes more than one *SMP* for 468 queries (72%). 321 queries(50%) have only one *SMP*. This can be explained by the fact that the average query length of the test query set is 2.67.

We use three types of features for retrieval. Firstly, *single terms*(T),*SMP*(S) and *NMP*(N). Our idea for *SMP* is straightforward, that the sequence of terms in *SMP* should appear in the document exactly as in the query. For *NMP*, although no specific idea might be associated with it, its appearance can provide some evidence of relevance. Thus we require that *NMP* should appear within a given proximity of each other, ether ordered or unordered. We choose the scope of proximity to be four times the number of terms, as suggested by several previous work. Therefore, our ranking function is of the form:

$$P_\Lambda(D|Q) \stackrel{rank}{=} \sum_{c\in T} \lambda_T f_T(c) + \sum_{c\in S} \lambda_S f_S(c) + \sum_{c\in N} \lambda_N f_N(c)$$

4. EXPERIMENTAL RESULTS

We employ our method on four TREC document collections, and use 650 queries (only the title field) to test the retrieval effectiveness. We compare against two methods: (unigram) language model(U-LM)[5], and MRF model for IR (MRF-IR). U-LM is based on the bag-of-words assumption which does not use any term dependencies. The setting for MRF-IR is the same as that of FD in [4]. We present two methods, WIKI-SMP and WIKI-SMP-NMP, we set (λ_T,λ_S,λ_N) to be (0.9,0.1,0.0) and (0.8,0.1,0.1) respectively. Note that all the four methods are based on the linear feature model.

Table 2 shows that utilizing phrase based concepts in queries improves the retrieval effectiveness, as compared with the U-LM model. Moreover, our methods get comparable results with MRF-IR, but use much fewer features. We observed similar results in terms of metrics P@5 and P@10. Our experiments demonstrate that separating important and less important term dependencies and treating them accordingly is effective and efficient for document retrieval.

Name	Size	#Docs	Topics
AP88-90	730MB	242,918	51-200
ROBUST2004	1.9GB	528,155	301-450,601-700
WT10g	11GB	1,692,096	451-550
GOV2	427GB	25,205,179	701-850

Table 1: Overview of TREC collections and topics

Method/Corpus	AP	Robust	WT10g	GOV2
U-LM	0.1376	0.2219	0.1831	0.2803
MRF-IR	0.1393*	0.2294*	0.1945*	0.3090*
WIKI-SMP	0.1435*	0.2292*	0.1927*	0.2950*
WIKI-SMP-NMP	0.1427*	0.2335*	0.1970*	0.3113*

Table 2: Comparison of effectiveness in terms of MAP. * indicates a significant improvement over the first method (p<0.05 with a one-tailed paired t-test).

Method	# of query terms					
	2	3	4	5	6	7
MRF-IR	3	7	17	36	73	143
WIKI-SMP	1	1	2	2	2	3
WIKI-SMP-NMP	3	5	9	14	19	25

Table 3: Comparison of average size of features set, for queries with different number of terms.

Our explanation for the results is that, important term dependencies are associated with specific concepts, and these concepts help predict contents related to user need. Thus when used as features, these types of dependency should be kept exactly the same as the pattern appearing in the query. As for Wikipedia, it is a dynamic and quickly growing resource, and we expect that better results using *SMP* recognition can be obtained when more resources are available, and we will investigate whether better quality of *SMP* recognition will lead to further improvement.

5. REFERENCES

[1] J. Gao, J.-Y. Nie, G. Wu, and G. Cao. Dependence language model for information retrieval. In *SIGIR*, pages 170–177, 2004.

[2] Indri. *http://www.lemurproject.org/indri/*.

[3] D. Metzler. Automatic feature selection in the markov random field model for information retrieval. In *CIKM '07*, pages 253–262, 2007.

[4] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *SIGIR*, pages 472–479, 2005.

[5] F. Song and W. B. Croft. A general language model for information retrieval. In *CIKM '99*, pages 316–321, New York, NY, USA, 1999. ACM.

[6] Wikipedia. *http://www.wikipedia.org*.