Modeling Subset Distributions for Verbose Queries

Xiaobing Xue W. Bruce Croft
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts, Amherst, MA,
01003, USA
{xuexb,croft}@cs.umass.edu

ABSTRACT

Improving verbose (or long) queries poses a new challenge for search systems. Previous techniques mainly focused on two aspects, weighting the important words or phrases and selecting the best subset query. The former does not consider how words and phrases are used in actual subset queries, while the latter ignores alternative subset queries. Recently, a novel reformulation framework has been proposed to transform the original query as a distribution of reformulated queries, which overcomes the disadvantages of previous techniques. In this paper, we apply this framework to verbose queries, where a reformulated query is specified as a subset query. Experiments on TREC collections show that the query distribution based framework outperforms the state-of-the-art techniques.

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

General Terms
Algorithms, Experimentation, Performance

Keywords
Verbose Query, Query Distribution, Query Reformulation

1. INTRODUCTION

Verbose (or long) queries have attracted much attention recently, since they allow users to express their information need using natural languages. However, current search systems can not deal with the verbose query well due to its complexity. Thus, improving verbose queries poses a new challenge for the development of search systems.

Previous techniques on improving verbose queries can be roughly divided into two categories. The first category emphasizes important words and phrases of the original query [6][3][2]. The methods from this category do not consider how those important words and phrases are used together to form actual subset queries therefore missing important relationships between words and phrases. The second category attempts to select the best subset query [5][1]. The methods of this category indeed consider a subset query as a whole, but they mainly focus on picking the best subset query and ignore alternative subset queries.

Recently, a general reformulation framework has been proposed to transform the original query into a distribution of reformulated queries [9]. This framework addresses the disadvantages of previous techniques mentioned above. On one hand, a reformulated query is explicitly modeled in this framework, which helps capture the dependencies between words and phrases that are imposed by actual queries. On the other hand, this framework considers not only the best reformulated query but also other alternatives. In this paper, we apply this general framework to improve verbose queries, where the original verbose query is reformulated as a distribution of subset queries.

Modeling the subset distribution for a verbose query has been first studied by Xue et al [10]. However, they used a fixed parameter to combine the original query with the generated subset distribution. Thus, a principled method that can effectively incorporate the original query and its subset queries within the same distribution is still missing.

2. MODELING SUBSET DISTRIBUTION

In this section, we describe how to model the subset distribution for a verbose query. Formally, given the original verbose query Q, we first construct a set of subset queries $V_{Q_s} = \{Q_s\}$, where Q_s is a subset query extracted from Q. Note that Q also belongs to V_{Q_s}. Then, Q is reformulated as a distribution over V_{Q_s}, i.e. $\mathbf{P}_{Q_s} = \{(P(Q_s|Q)\ Q_s)\}$, where $P(Q_s|Q)$ is the probability assigned to Q_s in the distribution $\mathbf{P}_{Q_s}$. We will describe the generation of V_{Q_s} and the estimation of $P(Q_s|Q)$, respectively.

Following Kumaran and Carvalho [5], only subset queries with the length between three to six words are considered. In addition, we also consider two special subset queries, one is the original query and the other is the key concept discovered in previous work [2]. Furthermore, since the subset queries generated will be finally used for retrieval, it is necessary to indicate the retrieval model used. In this paper, we consider two types of retrieval models, the Query Likelihood Model (QL) [8] and the Sequential Dependency Model (DM) [7].

In order to estimate the probability for each Q_s, we assume $P(Q_s|Q)$ is a linear combination of a variety of query features. These features characterize a subset query as a whole therefore capturing the relationships between words and phrases. Examples of features include various query quality predictors, the number of passages where a subset query has been observed in target corpus and the language model probability returned by Microsoft Web NGram Service. In order to learn the combination parameter for each query feature, we generate the corresponding retrieval feature by calculating the sum of the retrieval scores of using all subset queries weighted by their query feature values. Then, a learning to rank method is used to learn the parameters

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

Table 1: Example of the subset distribution. For "subset query (n)", n indicates the length.

Original Query (Q): remedies treatments given lessen stop effects ovarian cancer		AP: 13.53
QDist-QL		AP: 21.02
$P(Q_s\|Q)$	Subset Query(Q_s)	Subset Query Type
0.153	remedies treatments given lessen stop effects ovarian cancer	original query
0.099	remedies treatments ovarian cancer	subset query (4) & key concept
0.046	ovarian cancer	key concept
0.007	remedies treatments stop ovarian cancer	subset query (5)

Table 2: Results of different models. ^{q,d,s,k} denotes significantly different with QL (^{q}), DM (^{d}), SRank (^{s}) and KeyConcept (^{k}), respectively.

	Gov2		Robust04	
	MAP	P@10	MAP	P@10
QL	25.43	52.21	25.49	43.13
DM	27.85	54.03	26.83	44.94
SRank	24.99	50.74	24.78	41.57
KeyConcept	27.52	53.83	25.97	41.65
QL+SubQL	26.76	53.15	26.20	43.21
DM+SubQL	28.70	55.37	27.37	**45.14**
QDist-QL	27.41^{qs}	53.42^{s}	26.07^{s}	42.69
QDist-DM	$\mathbf{29.59}^{qdsk}$	$\mathbf{55.84}^{qs}$	$\mathbf{27.55}^{qsk}$	44.94^{qsk}

of these generated retrieval features and these parameters finally serve as the combination parameters of query features. In this paper, a ListNet [4] is used as the learning method.

3. EXPERIMENTS

Two TREC collections (Gov2 and Robust04) are used for experiments. For each collection, the index is built using Indri with Porter Stemmer. For each topic, the description part is used as the query after stopword removal. Mean average precision (MAP) and precision at 10 (P@10) are used to measure the retrieval performance. The two tailed t-test measures significance. The query set is split into a training set and a test set. Ten-fold cross validation is conducted. Two types of subset query distributions are implemented, one uses QL as the retrieval model (QDist-QL) and the other uses DM (QDist-DM). Table 1 shows an example of the subset distribution learned for QDist-QL, which significantly outperforms the original query.

The baselines used include QL, DM, SRank [5], KeyConcept [2] and two methods from Xue et al [10] (QL+SubQL and DM+SubQL). The retrieval performance is displayed in Table 2. The best performance is bolded.

Table 2 shows that QDist-QL is comparable with DM, SRank and KeyConcept, i.e., the state-of-the-art techniques on improving verbose queries. Moreover, QDist-DM significantly outperforms most of the baseline methods and achieves the best performance on both collections. Then, QL+SubQL and DM+SubQL are compared with QDist-QL and QDist-DM, respectively. The difference is that the former uses a fixed parameter to combine the original query with the generated subset distribution, while the latter learns a unified distribution including both the original query and the subset queries. The latter method outperforms the former one on Gov2 and they are comparable on Robust04.

The current subset distribution consists of queries with the mixed length from three to six. It is interesting to explore the fixed-length subset distribution. Table 3 shows the performance of using the fixed-length subset distribution from three to six. "mix" denotes the distribution mixing queries with different lengths.

Table 3 shows that the performance of the fixed-length subset distributions is close to and sometimes even better than the mixed-length distribution and the length itself does not have much influence. Thus, we can replace the mixed-length subset distribution with the fixed-length distribution, which significantly reduces the number of subset queries in the distribution therefore improving efficiency.

Table 3: Retrieval performance of fixed-length subset query distribution.

	Gov2		Robust04	
length	MAP	P@10	MAP	P@10
3	27.39	53.56	26.10	42.65
4	27.33	53.96	25.97	42.33
5	27.09	53.36	25.97	42.09
6	27.21	53.62	26.01	42.09
mix	27.41	53.42	26.07	42.69

4. CONCLUSION

Modeling the subset distribution for a verbose query helps overcome the disadvantages of previous techniques. A recently proposed framework is used in this paper to learn a unified subset distribution. Experiments on TREC collections show the effectiveness of this method.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by ARRA NSF IIS-9014442. Any opinions, findings and conclusions or recommendations expressed in this material are the author's and do not necessarily reflect those of the sponsor.

5. REFERENCES

[1] N. Balasubramanian, G. Kumaran, and V. R. Carvalho. Exploring reductions for long web queries. In *SIGIR10*, pages 571–578.

[2] M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. In *SIGIR08*, pages 491–498.

[3] M. Bendersky, D. Metzler, and W. B. Croft. Learning concept importance using a weighted dependence model. In *WSDM10*.

[4] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *ICML07*, pages 129–136.

[5] G. Kumaran and V. R. Carvalho. Reducing long queries using query quality predictors. In *SIGIR09*, pages 564–571.

[6] M. Lease, J. Allan, and W. B. Croft. Regression rank: learning to meet the oppotunity of descriptive queries. In *ECIR09*, pages 472–479.

[7] D. Metzler and W. B. Croft. A Markov random field model for term dependencies. In *SIGIR05*, pages 472–479.

[8] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR98*, pages 275–281.

[9] X. Xue and W. B. Croft. Representing queries as distributions. In *SIGIR10 Workshop on Query Representation and Understanding*, pages 9–12.

[10] X. Xue, S. Huston, and W. B. Croft. Improving verbose queries using subset distribution. In *CIKM10*, pages 1059–1068.