ICTNET at Web Track 2009 Diversity task

Wenjing Bi[1,2], Xiaoming Yu[1], Yue Liu[1], Feng Guan[1,2], Zeying Peng[1,2], Hongbo Xu[1], Xueqi Cheng[1]
1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190
2. Graduate School of Chinese Academy of Sciences, Beijing 100190

ABSTRACT
We (ICTNET team) participated in Web Track of TREC2009, and in this paper, we summarize our work on Diversity task of Web Track, which is new in this year. The goal of the diversity task is to return a ranked list of pages that together provide complete coverage for a query, while avoiding excessive redundancy in the result list. For this task, we cluster the results of ad hoc task, and rerank the results depend on subtopics docs covers. Besides, we introduce two methods which tried to find the implicit subtopic by using the docs returned from commerce search engine.

1. INTRODUCTION
The goal of Web Track is to explore and evaluate Web retrieval technologies over billion pages. The web track has two tasks in this year. One is a traditional adhoc task , whose objective is to retrieve the most relevant documents for each topic, and the other is a new diversity task, we will talk about it explicitly in Section 2 .What's more, a new web corpus ClueWeb09 [1]was introduced in this year, which is crawled by the Language Technologies Institute at Carnegie Mellon University during January and February 2009.

In the older web track , several models have been proposed to combat the adhoc task. We follows the traditional methods and combines several characters distilled from documents and queries together to find the most relevant documents, such as, the term frequency, inverse document frequency, document length, the pagerank value of documents , and so on.

For diversity task, we adopts two main methods to solve this problem. The one is clustering the results of ad hoc task, and reranking the results depend on subtopics docs covers, which is traditional, Besides, we tried to find the implicit subtopic by using the knowledge distilled from the internet.

The report is organized as follows. Section 2 we talk about web track explicitly. In Section3, we discuss the cluster method, and the explicit subtopic identification was introduced in Section 4. Conclusion is given in section 5.

2. Diversity task
Different from the older web track, besides a traditional adhoc retrieval task , a new diversity task was introduced. The goal of this new diversity task is to return a ranked list of pages that together provide complete coverage for a query, while avoiding excessive redundancy in the result list.

Traditional retrieval models assume that the relevance of a document is dependent of the relevance of other documents, which makes it possible to retrieval the relevant documents separately. In reality, however this independent assumption rarely holds. For example, a relevant document may be useless to a user if the user has already seen another document with the same content. Another example is when the user's information need is best satisfied with several documents working together; in this case, all the documents should be return, and the rank of each document is dependent on other documents[2]. Previous Web tracks have explored specific aspects of Web retrieval, including named page finding, topic distillation, and traditional adhoc retrieval, and this year diversity task combines aspects of all these older tasks by concerning the diversity and dependency.

For diversity task, TREC2009 give same 50 topics as adhoc task, which were developed from information extracted from the logs of a commercial Web search engine. Each topic was structured as a representative set of subtopics relating to different users' need. And the subtopic will not released until team runs were submitted.

The judging process and evaluation measures are also different with adhoc task. For this task, TREC2009 introduced two new evaluation methods, which are α-nDCG measure[3] proposed by

Clarke et al and intent aware MAP[4] (MAP-IA) proposed by Agrawal et al. Documents were judged with respect to the subtopics. For each subtopic, a binary judgment was made as to whether or not the document satisfies the information need associated with the subtopic. the probability of relevance of a document is conditioned on the documents that appear before it in the result list.

3. Cluster

The major goal of clustering is partitioning a given data set into disjoint subsets (clusters). In the diversity task, clustering is a good method to catch the different document relevant to different subtopics. There are a lot of method to cluster documents, such as K-means, PAM, Hierarchy Agglo Clustering, OPTICS and so on. In our task, we developed K-means algorithm by getting the cluster num dynamically. There are several methods to calculate vectors' distance, such as cosine distance and Euclidean distance. We use both cosine distance and Euclidean distance in our method.

We cluster the result returned by adhoc task, using the developed K-means algorithms. Each doc cluster represent a subtopic. And then, we rerank the documents depend on the subtopics the doc covers. For example, the larger cluster is the more documents will be appeared in the final results.

In TREC2009 diversity task, our run based on cluster gets the third highest α-nDCG (Category B), and its IA-P@10 is also very high, and you can see more details in Table 1.

4. Finding the Implicit Subtopics

As the one topic is represented by several implicit subtopics, if the subtopics was given, the relevance of documents will be set easier. In this section, we tried to find the implicit subtopics with global information coming from the World Wide Web. We adopt two main method to get the implicit subtopics. The one is selective query expansion and the other is implicit subtopic finding.

4.1 Query Expansion

Query expansion is a technology to match additional documents by expanding the original search query. In diversity task, topics were represented by short queries most of which has no more than three words. It will be very hard to predict subtopics only use those short queries. In order to reach the information of the topics, query expansion will be a good method, and it also can be used to rerank the docs returned from the adhoc task. In our method, the diversity of topics was represented by the weight of expansion words.

People can use pseudo-relevant documents or other resources (such as Wikipedia) as the basis of query expansion. For example, the anchor texts in the article titled by the query are used to expand the original query. In our system, we rely on the documents returned by commerce search engine. As the topics were developed from information extracted from the logs of a commercial Web search engine, we thought using search engine results as the pseudo-relevant documents would be useful and reasonable. Usually, only the top-ranked documents are considered to be pseudo-relevant documents. The terms occurred in the pseudo-relevant documents are weighted and ranked. In our work, we choose two kinds of term weighting models. The one, known as Bo1, is based on Bose-Einstein statistics. The other is based on the Kullback Leibler (KL) divergence[5] between the pseudo-relevant documents and the collection. People can select the top-ranked terms as candidate for query expansion. We use the default setting of the parameters for query expansion, and choose the top 10 terms from the top 10 ranked documents, suggested by Amati in[5]. The term weights were used to distinguish subtopics.

Table 1 describes the results of our three runs for document search task. We can see that the run based on Query Expansion retrieval model gets the highest IA-P@10 and its α-nDCG @10 is also acceptable.

Table 1. Results for the diversity task

Methods	α-nDCG@10	IA-P@10
Cluster	0.272	0.095
Query Expansion	0.212	0.098
Implicit Subtopic	0.061	0.026

4.2 Implicit Subtopic

In "Implicit Subtopic" method, the search engine results were also used. As we know, different subtopics can be represented by several documents. We obtain the subtopics by catching the representative documents, however, getting those documents is a challenging work. In our "implicit subtopic" run, we tried to use the information on the internet. Theoretically, finding the implicit subtopics can improve the diversity task system's performance.

On the World Wide Web, there exist many documents which represents several implicit subtopics. We used commerce search engines to gather those documents. In this task, our work can be divided into five steps. First, we collect documents returned by commerce search engines, and considered those documents can satisfy different users' need. Second, as clustering is a good method to partition a mess data set into disjoint subsets (clusters), we partition the search engine results into different subsets, each set presents a subtopic. Third, We obtain the signature words for each subtopic from the clustered document sets. Four, we searched the ClueWeb09 corpus by those distilled signature words, and the subtopic relevant documents were obtained. Finally, we rerank those results by an greedy algorithm introduced by Zhai[2]. Although the evaluate result is not as high as we forecasted, we still believe subtopic implicating will catch more researchers' interests.

5. Conclusion and Future work

This paper reports the experiments of our team on Diversity Task of Web Track 2009. For this task, we cluster the results of ad hoc task, and rerank the results depend on subtopics docs covers. Besides, we introduce two methods which tried to find the implicit subtopics by using the docs returned from commerce search engine.

In the future, we will devote to improve cluster algorithms in our task and explore methods to find the implicit subtopics.

6. ACKNOWLEDGMENTS

Thank to the organizers of TREC 2009 Web track, the NIST assessors and the other track participants for judging the documents. Thank to the members of Information Retrieval Group in the Institute of Computing Technology, Chinese Academy of Sciences.. In addition, our work was supported by Key Program of NSFC 60933005， 863 Program of China 2007AA01Z441， 973 Program of China 2007CB311100.

7. REFERENCES

[1]Jamie C, Mark Hoy, Changkuk Yoo, and Le Zhao. The web09-bst Dataset [J]. 2002:

[2]Zhai C, Cohen W, Lafferty J. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval[A]. ACM New York, NY, USA, 2003:10-17.

[3]Clarke C, Kolla M, Cormack G, et al. Novelty and diversity in information retrieval evaluation[A]. ACM New York, NY, USA, 2008:659-666.

[4]Agrawal R, Gollapudi S, Halverson A, et al. Diversifying search results[A]. ACM New York, NY, USA, 2009:5-14.

[5]AMATI G, VAN RIJSBERGEN C. Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness[J]. ACM Transactions on Information Systems, 2002,20(4): 357-389