ICTNET at Web Track 2010 Diversity Task

Yuanhai Xue[12], Zeying Peng[12], Xiaoming Yu[1], Yue Liu[1], 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

In this paper, our team – "ICTNET", participated in the diversity task of Web Track of TREC 2010. The full Category A dataset was used. The same settings as the ad-hoc task were adopted for retrieval. Different clustering methods which were then applied on different fields are elaborated. Query expansion techniques are presented next.

1. Introduction

TREC Web Track [1] is designed to explore and evaluate Web retrieval technologies over the billion-page ClueWeb09 Dataset [2]. ClueWeb09 Dataset was crawled by the Language Technologies Institute at Carnegie Mellon University during January and February 2009, including roughly 1 billion pages in 10 languages. Last year in this track, Category B of ClueWeb09, the scale of which is rather small, was used. This year, Category A is used to fully explore the whole dataset.

Diversity task of TREC Web Track aims at diversifying the search results, which means to find the various aspects of the original query implies. Since ad-hoc task focuses on retrieval relevant, diversity task focuses on predicting the varieties of user intentions and reorganize the results to meet different requirements.

In section 2, a short description of the system settings and query model inherited from ad-hoc task is given. In section 3 the methods for clustering different fields are presented. Section 4 enumerates the details on query refinement. In section 5, re-ranking methods are presented. Section 6 examines the results of evaluation, and then section 7 concludes this report.

2. Data Preparation

In this task, we used the search results from the ad-hoc task with the same method and settings. Firtex [3] was deployed over ten servers, which is an open source high performance search platform. The Category A dataset was dispatched onto each machine equally.

The sliding window BM25 with adaptive window's size is used as the query model. The results of each query were anti-spammed with the Waterloo Spam Rankings with a threshold at 50 percentile.

This query model was used for both the original fifty queries and the query expansions. When the results of the original queries were returned, the content of each page was also fetched for clustering. For query expansions, only the TREC-ids were needed for re-ranking.

3. Clustering

In order to diversify the search results, subtopics of each topic should be mined. Clustering is one of the ways to partition a given set into disjoint subsets. There are a large amount of methods to cluster documents, such as K-means, PAM, Hierarchy Clustering, OPTICS and so on.

In this task, the developed K-means algorithm which is called Bisecting K-means [4] was applied. This algorithm starts with a single cluster of all the documents. Then at each iteration step, choose a cluster to split into two sub-clusters using the basic K-means algorithm. Repeat that until the desired number of clusters is reached. The largest cluster is picked to split where the size of a cluster is defined by the linear combination of convergence and diameter measures as following:

$$Size_c = k_1 \times Convergence(C) + k_2 \times Diameter(C) \,,$$

where *Size*$_c$ denotes the size of cluster C; *Convergence* is the average distance between every two documents in the cluster:

$$Convergence(C) = average[D(d_i, d_j)];$$

and *Diameter* is the longest distance among the documents within the cluster:

$$Diameter(C) = \max D(d_i, d_j)$$

There are several methods to calculate vectors' distance. The cosine distance was used in our algorithm. For each cluster, all the documents within it were sorted by the distance between document and the center of the cluster in ascending order. At last, all the clusters were sorted by the number of documents within them in descending order.

Assuming that content, keyword and anchor text provide the most useful information, these three fields were distilled from the result documents returned by ad-hoc task. The first two fields, content and keyword, are separately clustered using the Bisecting K-means algorithm.

For anchor text, the following method was used to cluster them. The importance of all the words which appeared in the result documents measured by their term frequency – inverse document frequency weights as usual were first calculated. Then each word was treated as a subtopic and all the documents in which the word appears are viewed as covering this subtopic.

4. Query Refinements

1) Query Expansion by Commercial Search Engine.

Query expansion performs an important role for IR in both research and practice. In diversity task or commercial searching actions, queries provided are relatively short, usually no more than three words. This means that it is difficult to predict the actual intentions of users who give the queries. Query expansion can be used to address this problem by providing additional key words to the search engine. With the extra information, users' intensions would be specified, and ambiguities are avoided.

One common way to get query expansion is relevant feedback [5]. But since relevant feedback needs excessive human resources, plus the difficulty to guess the diverse aspects of queries, it is not suitable in diversity task's scenario.

Another way for query expansion is query log learning. Through analysis of user behavior patterns, related query words could be found. But this method requires a large set of logs, which we did not have at hand. Even if we do have other source of query logs, they might not correspond to the Clueweb09 dataset in this task. Thus, this method is not suitable either.

Finally, we resorted to commercial search engines. Recently, most search engines give search suggestions in the results pages, which are commonly high-quality expansions for the original queries. In this task, all the queries were fed into Google, and collected all query suggestions in the results pages for each query correspondingly.

These query expansions were used to fetch 1000 results from our Firtex retriever.

2) Query Expansion by search results clustering

Clustering search results is a frequently used method to reveal the cohesion and diversity of the results' content. After the results are clustered, salient phrases can be extracted from each cluster, which representing the aspect of each cluster [6]. These salient phrases are quite different from each other, so that they imply diversity.

The clustering results from two search engines: Cuil [7]* and Clusty [8] were utilized. Both of the two web sites cluster the search results and provide salient phrases for every cluster. These salient phrases were extracted from the result pages for every query.

Since most of the salient phrases do not include words of the original queries, the original query words and the salient phrases were combined to form the expansion for each query. After that, these expansions were used to fetch 1000 results from our Firtex search platform.

5. Re-ranking

In this section, the search result diversification algorithm introduced in [9] was applied to re-rank the result documents returned by ad-hoc task. Assuming that a good ranking should cover as many relevant subtopics as possible, the algorithm is a greedy one to iteratively select the best document from the remaining documents. The Σ function was adopted to combine subtopics from six diversity dimensions mentioned above, which are clustering results of content, keyword, anchor text and search results of query expansion by Google, Cuil, Clusty. Since the keywords usually represent the intention of the document better, we increase the weight of the keyword dimension in the diversification algorithm.

6. Results

Three runs were submitted in this year's diversity task. In the first run, clustering was used for web page content field. The second run applied clustering also on the keyword field. The third run used anchor texts in addition with the clustering method described above

The results of the runs submitted are listed in Table 1.

Run	*ERR-IA*@20	*α-nDCG*@20	*NRBP*	*MAP-IA*	*P-IA*@20
ICTNETDV10R1	0.3008	0.4390	0.2516	0.0368	0.1394
ICTNETDV10R2	0.3292	0.4605	0.2853	0.0407	0.1456
ICTNETDV10R3	0.3250	0.4556	0.2798	0.0395	0.1409

Table 1: Results of submitted runs

7. Conclusion

In this task, different clustering methods were applied on different fields extracted from the search result pages. An improvement could be seen after the clustering output from the keyword field was added. But when the clustering method was applied on in-link anchor texts, the evaluation result dropped slightly. This is in accord with the observation that the quality of in-link anchor texts is not satisfying in the ClueWeb09 dataset. Although the effect of using anchor texts was not promising, we still believe that anchor texts are worth of future explore.

Acknowledgements

We thank all the organizers of TREC Web Track and NIST. We appreciate the efforts of all assessors for judging the runs. This work is supported by NSF of China Grants No. 60933005, No. 60903139 and No. 60873245, and also by "863" Program of China Grant No. 2006AA010105-02.

References:

[1] http://plg.uwaterloo.ca/~trecweb/2010.html

[2] http://boston.lti.cs.cmu.edu/Data/clueweb09/

[3] http://sourceforge.net/projects/firtex/

[4] Steinbach, M.; Karypis, G. & Kumar, V. A Comparison of Document Clustering Techniques KDD Workshop on Text Mining, 2000, 34, 35

[5] Manning, C.D. and Raghavan, P. and H Schütze, Introduction to Information Retrieval, Cambridge University Press; 1st edition, July 7, 2008

[6] Zeng, H.; He, Q.; Chen, Z.; Ma, W. & Ma, J. Learning to cluster web search results, Proceedings of the 27th Annual International ACM SIGIR Conference 2004, 217

[7] http://www.cuil.com/

[8] http://search.yippy.com/

[9] Zhicheng Dou, Kun Cheny, Ruihua Song, Yunxiao Ma, Shuming Shi, and Ji-Rong Wen, Microsoft Research Asia at the Web Track of TREC 2009, in Proceedings of TREC 2009, 2009