ICTNET at Web Track 2011 Diversity Task

Shengxian Wan[12], Yuanhai Xue[12], Xiaoming Yu[1], Feng Guan[1], Yue Liu[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

1. Introduction

Traditional IR systems use document-query relevance as the main measure for ranking and consider the relevance is independent of the other documents. However, the only measure of relevance cannot deal with redundant information and can fail to reflect the broad range of user needs that can underlie a query. IR systems need to consider the diversity and novelty of the result list. Diversity task is trying to solve this problem. The goal of 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. This year, the primary evaluation measure is intent aware expected reciprocal rank (ERR-IA) [1], some other evaluation measures such as α-nDCG [1] have also been evaluated.

This paper is organized as follows. Section 2 describes different clustering methods used for text clustering. Section 3 describes our query expansion method. Section 4 describes the diversity model used for re-ranking. Section 5 analyzes the result and in section 6 we conclude our work.

2. Clustering

We obtained all relevant documents of a query by techniques used in ad hoc task. The first run of this year's ad hoc task was used as the input of diversity task. In order to diversify the result list, we tried to obtain the sub-topic information of each query and documents relevance of each sub-topic. Clustering is a naturally method to cope with this problem, it can divide documents into different set by document similarity. There are different clustering methods can be used, such as k -means, PAM, Hierarchy Clustering, OPTICS and so on. In last year's diversity task, we used bisecting k-means [2] to cluster the search result and considered that each cluster represented one subtopic.

There is one obvious drawback of bisecting k-means. It is a hard clustering method and that is to say, each document can belong to only one cluster. This is not true in real scene. One document can contain more than one sub-topic of a query. For example, one *wiki* page may contain almost every popular sub-topic of a query. In order to solve this problem, we tried to use fuzzy clustering instead of hard clustering and selected fuzzy c-means [3] as our main clustering method. One problem of using fuzzy c-means is how to determine the argument K just like k-means. Our solution is using bisecting k-means first and then using fuzzy c-means to re-cluster the documents based on the result of bisecting k-means.

We used vector space model (VSM) to represent documents and used *cosine* distance between document vectors as the distance between documents. The dimension of document vector was very high if using all the words appeared in the document set. There was too much noise in the high dimension vector and it made document distance inaccurate. We found the words appeared in the *keyword* region of each document are representative and we used all the *keywords* of the whole document set as universal word set while ignoring other words. We still considered each cluster as one subtopic, and got document-sub-topic relevance by degree of membership between document and cluster.

3. Query expansion

Query expansion is an effective method to boost recall and precision in IR systems. It reformulates a seed query and compensates the problem of insufficient information of the query. In diversity task, the queries are commonly very short and the real user intent is not obvious. We tried to use query expansion to cover as more probable user intents as possible. According to our experience in TREC 2009 and TREC 2010, query expansion is effective to improve the result and we also used this method in TREC 2011.

We expanded the queries by commercial search engines Bing and Google. For that *wiki* has little information about most of the queries, we did not use it. We parsed query expansions from Bing's and Google's websites and considered each expansion as one subtopic. We used these expansions as new queries to retrieve documents from the whole corpus. By the result ranking list, we got the relevance between each document and each subtopic.

4. Diversity model

Many scholars have modeled the diversity and novelty of information retrieval based on different perspectives. Our model is based on xQuAD [4] proposed by Santos in 2010. xQuAD explicitly models an ambiguous query as a set of sub-queries. It greedily selects one document every time based on the following probability formula:

$$(1-\lambda)\mathrm{P}(\mathrm{d}|\mathrm{q})+\lambda\sum_{q_i\in Q}[P(q_i|q)P(d|q_i)\prod_{d_j\in S}(1-P(d_j|q_i)] \qquad (1)$$

Where q is the initial query, Q is the sub-topic set of q， q_i is a sub-topic in Q, S is documents selected and d represents a document, $P(d|q)$ is the likelihood of document d being observed given the initial query q, $P(q_i|q)$ can be seen as a measure of the relative *importance* of the sub-topic q_i, $P(d|q_i)$ is a measure of the *coverage* of document d with respect to the sub- topic q_i. We can regard the first part of the formula as modeling the *relevance* and the last part as modeling the *diversity*. The coefficient λ measures the weight between *relevance* and *diversity*.

We changed this model in our work. First, we used more than one method to mining the subtopic. Different methods may have different contributions for our problem and thus we tried to mix them all. We divided the *diversity* part of (1) and gave a weight to each method. Second, we considered $\prod_{d_j\in S}(1-P(d_j|q_i))$ of (1) is to penalize redundant for each sub-topic and gave different subtopic different speed of weight decreasing. There are two reasons, first, if one subtopic is more popular, or more important, more pages may be needed to satisfy people's information needs and thus the speed of weight decreasing should be lower. The second reason is from practical point of view. We may get noise sub-topics in practice, if they are treated the same as correct sub-topics, according to (1), the result list will include many document of the noise sub-topic. We used $P(q_i|q)$, the measure of the subtopic's relative importance, to change the speed of weight decreasing. Besides, the diversity part of (1) will decrease exponentially and make little function quickly. In our work we normalize the weight of sub-topic after redundant punishment each time. Our formula is as follows:

$$(1-\lambda)\mathrm{P}(\mathrm{d}|\mathrm{q})+\lambda\sum_{m_k\in M}w_k\left\{\sum_{q_i\in Q}\left[P(d|q_i)Norm_{q_i\in Q}\left(P(q_i|q)\prod_{d_j\in S}(1-P(d_j|q_i))^{P(q_i|q)}\right)\right]\right\} \qquad (2)$$

Where M represents all methods we used in practice and w_k is the weight of the kth method.

5. Result

In TREC 2011, we submitted three runs for diversity task. In run 1, we used both bisecting *k*-means clustering and query expansion. The method is similar with which we used in TREC 2010 diversity task but we modified the method in details. We considered this run as a baseline. In run 2, we used bisecting *k*-means first and then used fuzzy *c*-means to re-cluster the result as we described in Section 2, we used the diversity model described in Section 4 to re-rank the result. In run 3, we used the same clustering method and diversity model of run 2 but we combined the query expansion method in order to see how significant the effective of query expansion is.

RUN	ERR-IA@20	nERR-IA@20	α-DCG	α-nDCG
ICTNET11DVR1	0.4716	0.5102	0.5442	0.5795
ICTNET11DVR2	0.4603	0.5001	0.5274	0.5638
ICTNET11DVR3	0.4764	0.5150	0.5458	0.5818

Table 1. Performance of our runs in TREC 2011 diversity task

The result is listed in Table 1. We can see that run 3 is slightly better than run 1.The fuzzy clustering method boosts the result but the effect is not as good as we supposed. According to run 2 and run 3, we can see that query expansion plays an important role in our method.

6. Conclusion and Future work

We describe our experiment of diversity task in this report. This year, we applied new clustering method to cluster the documents and used new diversity model to re-rank them. Though fuzzy clustering seems more suitable for diversity task naturally but it only improved the result slightly in our experiments. Yet we still believe it is worth for future research. Besides, we find that query expansion improved the result significantly.

Acknowledgements

Thank all the organizers of TREC 2011. Thank all the participants and assessors. We really appreciate of your efforts. This work is sponsored by NSF of China Grants No. 60903139 and No. 60873243, and by 863 Program of China Grants No. 2010AA012502 and No. 2010AA012503.

References

[1] C. L. A. Clarke, N. Craswell, I. Soboroff & A. Ashkan. A Comparative Analysis of Cascade Measures for Novelty and Diversity. In *Proc. of WSDM*, pages 75-84, 2011.

[2] M. Steinbach, G. Karypis & V. Kumar. A Comparison of Document Clustering Techniques. In *KDD Workshop on Text Mining* , pages 34-35, 2000.

[3] J. C. Bezdek. Pattern Recognition with Fuzzy Objective Function Algoritms. Springer, 1 edition, July 31, 1981.

[4] R. L. T. Santos, C. Macdonald & I. Ounis. Exploiting Query Reformulations for Web Search Result Diversification. In *Proc. of WWW*, pages 881-890, 2010.