ICTNET at Web Track 2012 Diversity Task

Zilong Feng[1, 2], Yuanhai Xue[1, 2], Xiaoming Yu[1], Hongbo Xu[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

Abstract

In this paper, we report our experiments at Diversity task, Web Track 2012. In this year, we attempt to use query expansion and topic model such as LDA[5] to get subtopics. And an model based on xQuAD[10] was used to re-rank the ad-hoc search results.

1. Introduction

The TREC Web Track explores and evaluates web retrieval technology over large collections of web data. As an inherently indistinct representative of more complex or ambiguous information needs, keywords submitted to a web search engine are often ambiguous. Such a query may cover many different aspects. Traditional IR systems use document-query relevance as the only measure of relevance to rank the web pages. Excessive redundancy web Pages of same aspects may be ranked higher. 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.

2. Data Preparation

In this task, we used the search results from the ad-hoc task with the same method and settings. Then we cluster the search results and re-rank them according to our clustering results.

The search result for a single query from the ad-hoc task is a list of structured data; each contains a web TREC-ID and the extracted main body of content. Since the extracting work has filtered most spam, the main body is still raw. So before clustering, we did some usual preprocessing on our web content. First we tokenize the text and remove all the punctuation, digits and tokens whose length is no more than 2. Then, we remove all the stop-words according a stop-word list. At last, we stem the words on the content using a tool called lib-stemmer library [1].

3. Clustering

Clustering is a usual and simple way to get the aspects of the original topic explicitly from the search results themselves. There are many text clustering methods can be applied, such as K-means, PAM, Hierarchy Clustering, OPTICS and so on. The year before last year, we applied a developed K-means algorithm which is called Bisecting K-means [2]. Last year, we found an obvious drawback of bisecting k-means. It's a hard clustering method which is usually not true in the real scene. In last year's diversity task, we use a soft clustering called fuzzy c-means [3] to re-cluster the documents based on the result of bisecting k-means. To our strange, as the result, it only improved the result slightly in our experiments. This year, we find an interpretation. In our diversity model, what we consider about is the probability of document with every individual aspect of origin topic. In clustering, it is the distance between the document and every individual clustering center. We don't care about which clustering center a document belong to, but the similarity to every clustering center. So a soft re-clustering not improving remarkably is reasonable.

In this year, we abandon the traditional clustering ways and try the topic model. A topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. An early topic model was described by Papadimitriou, Raghavan, Tamaki and Vempala in 1998[4]. In our diversity task, we use the most common topic model currently, Latent Dirichlet allocation (LDA). LDA was first presented as a graphical model for topic discovery by David Blei, Andrew Ng, and Michael

Jordan in 2002[5]. Using LDA, given the result documents from a single query, we can easily get subtopic-document distributions, i.e., P (topic|document) by setting number of sub-topics.

There are many open source implementations of LDA. Such as "Latent Dirichlet Allocation in C"[6], "GibbsLDA"[7], "JGibbLDA"[8] and so on. Finally we choose JGibbLDA, A Java Implementation of Latent Dirichlet Allocation using Gibbs Sampling for Parameter Estimation and Inference.

4. Query Expansion

Query expansion (QE) is the process of reformulating a seed query to improve retrieval performance in information retrieval operations [9]. In diversity task, we can consider each query expansion as an aspect or sub-topic of the origin query. According to our experience in TREC 2009, TREC 2010 and TREC 2011, query expansion is effective to improve the result. So we also try this method in TREC 2012.

We expanded the queries by commercial search engines Google. Actually, we put each query into the search engine and extracted the items of "Related Searches" on the result page. Usually there are total 8 items. We considered each item as one sub-topic. We treat these expansions as new queries to retrieve documents using the identical model in the ad-hoc task. For each result document by the originating query from ad-hoc task, we try to find it in every query expansion's retrieval results, and use the relevance score as the coverage of this document to the sub-topic.

But this year, maybe we didn't get good query expansions, or maybe there is something wrong with the experiment. We get many zero values when compute the coverage of this document to the sub-topic. And the final result is not as good as in former years. So at last, we didn't use any query expansion result in our submitted run.

5. Re-ranking Model

In the diversity task, our goal is re-ranking the documents which make them look diversified. The sub-topics from LDA or query expansion just provide guidance and supports when re-ranking. For how to re-ranking, our model is based on xQuAD[10] proposed by Santos in 2010, the same as last year. In last year's task, we changed the original probability formula and gained a new formula. This year, we use both formulas in our experiment. The result is shown in the next section. We also try some modification, but none of them showed a remarkable outperformance.

6. Result

In TREC 2011, we submitted three runs for diversity task. We do the parameter training on the data from the past three years (TREC2009, TREC2010, and TREC2011). All the three runs have the same model and framework. But there is still some difference in the details. In run 1, we didn't do any preprocessing before clustering using LDA, no stop-words removing, stemming or normalization. We didn't try any modification on the original model, just naive LDA and xQuAD. In run 2, we added stop-words removing, stemming and normalization, using our modified formula last year instead of the original formula in xQuAD. In run 3, we abandon word stemming before clustering because we found it express better in the past three years without stemming. And we apply the original formula back.

RUN	ERR-IA@20	nERR-IA@20	α-DCG@20	α-nDCG@20
ICTNET11DVR1	**0.3257**	**0.3466**	0.4001	0.4223
ICTNET12DVR2	0.3183	0.3396	0.3950	0.4175
ICTNET12DVR3	0.3243	0.3448	**0.4012**	**0.4239**

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

The results are listed in Table 1. We can observe that run 1 and run 3 have a better expression than run 2. Run 1 performs slightly better than run 3 in ERR-IA@20 and nERR-IA@20. But run 3 is better in α-DCG@20 and α-nDCG@20.

7. Conclusion and Future Work

We describe our methods and experiment of the diversity task in this report above. This year, we apply a new clustering method which is LDA model to cluster the documents and used the xQuAD model to re-rank them. From the results we can see: 1. From the experiment result of the 4 years data, LDA and xQuAD model has effectively performed in document diversity; 2. The change in the probability formula of xQuAD doesn't influence the perform much; 3. Without stemming before LDA clustering outperform with stemming. To the third point, we think it may be the reason that we set the same sub-topics number in the LDA model for every query. But in real scenes, for a different query, the number of aspects (sub-topics) may vary greatly.

Usually we all don't know the number of aspects of a query, using LDA may face a risk: how to set the topic number before training? In the future we will try to find some methods to deal with the problems. And we will also attempt to improve the present diversity model in our future experiment.

8. Acknowledgements

Thank all the organizers of TREC 2012 and NIST. Thank all the participants and assessors. We really appreciate of your efforts for judging the runs. This work is sponsored by NSF of China Grants No. 60933005, No. 61100083 and No.61173064, and by 242 Program of China Grants No.2011F65.

References

[1] http://snowball.tartarus.org/texts/stemmersoverview.html

[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 Algorithms. Springer, 1 edition, July 31, 1981.

[4] Papadimitriou, Christos; Raghavan, Prabhakar; Tamaki, Hisao; Vempala, Santosh (1998). "Latent Semantic Indexing: A probabilistic analysis". Proceedings of ACM PODS.

[5] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John. Ed. "Latent Dirichlet allocation". Journal of Machine Learning Research 3 (4–5): pp. 993–1022. doi:10.1162/jmlr.2003.3.4-5.993

[6] http://www.cs.princeton.edu/~blei/lda-c/

[7] http://gibbslda.sourceforge.net/

[8] http://jgibblda.sourceforge.net/

[9] D. Abberley, D. Kirby, S. Renals, and T. Robinson, The THISL broadcast news retrieval system. In Proc. ESCA ETRW Workshop Accessing Information in Spoken Audio, (Cambridge), pp. 14–19, 1999. Section on Query Expansion - Concise, mathematical overview.

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