University of Delaware at Diversity Task of Web Track 2010

Wei Zheng[1], Xuanhui Wang[2], and Hui Fang[1]

[1]Department of ECE, University of Delaware
[2]Yahoo!

Abstract

We report our systems and experiments in the diversity task of TREC 2010 Web track. Our goal is to evaluate the effectiveness of the proposed methods for search result diversification on the large data collection. In the diversification systems, we use the greedy algorithm to select the document with the highest diversity score on each position and return a re-ranked list of diversified documents based on the query subtopics. The system extracts different groups of semantically related terms from the original retrieved documents as the subtopics of the query. It then uses the proposed diversity retrieval functions to compute the diversity score of each document on a particular position based on the similarity between the document and each subtopic, the relevance score of the subtopic given the query and the novelty of the subtopic given the previously selected documents.

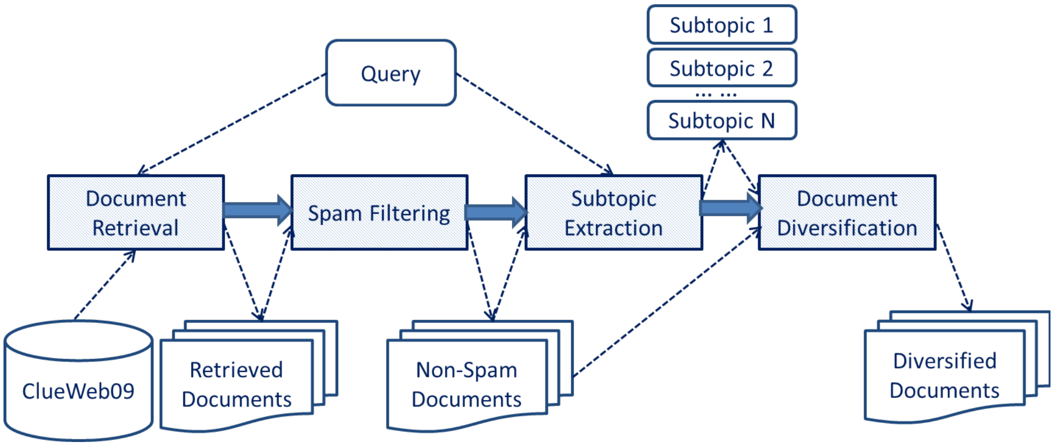
1 Introduction

The InfoLab from the ECE department at the University of Delaware participated in the diversity task of TREC 2010 web track. We study several retrieval functions for search result diversification.

Recently, many result diversification methods have been proposed and studied [8, 9, 5, 23, 7, 14, 21, 6, 18, 2, 19, 20]. The goal of search result diversification is to return documents that are not only relevant to the query but also diversified to cover different subtopics of the query. Some studies [5, 23, 7, 14, 21, 11, 19] directly compare the retrieved documents and minimize the redundant information among the documents. Other studies [6, 18, 2, 12, 4, 15, 3, 22, 16, 20, 19] model the diversity among documents based on their relation to the subtopics of the queries. The commonly used methods and evaluation measures for result diversification are based on the coverage of query subtopics [8, 9, 2].

The challenges in search result diversification are to extract the subtopics of the query and to diversify the returned results based on the subtopics. Query subtopics are the representative information needs associated with the query [8]. Most of the existing methods of subtopic extraction depends on external resources [17, 8, 19] which are difficult to represent the features of different kinds of document collections. In this paper, we propose to extract representative groups of semantically related terms from the original retrieved documents as the subtopics. Given the subtopics, the next problem is how to diversity the retrieved documents to cover different subtopics of the query. We apply the greedy algorithm [2, 19] to iteratively select the document with highest diversity score on each position and return the re-ranked list of documents. We explore two diversity functions that compute the diversity score of the document based on its relevance to the query subtopics, the importance of the subtopics given the query

Figure 1: Architecture of the search result diversification system



and the novelty of the subtopics given the previously selected documents. The TREC result shows the effectiveness of the proposed methods for search result diversification.

2 System Architecture

In this section, we describe the steps of the search result diversification system.

Figure 2 shows the architecture of our system. The system mainly has four steps.

1. **Retrieving documents for the original queries.** We use the original query to retrieve documents from the Category A collection of ClueWeb09 corpus based on Dirichlet smoothed language model retrieval function [24]. One possible solution is to directly build the index of the whole collection and use Indri to retrieve. However, it is time-consuming to build the index of the large collection and needs a lot of disk space to store it. Therefore, we need to build a smaller working set for the queries instead of using the whole category A collection.

2. **Deleting the spam documents from the retrieved documents.** ClueWeb09 collection contains a lot of spam documents [10]. The original retrieval result in the first step may also contain a lot of spam documents which would make it difficult to extract the query subtopics and diversify the results in the following steps. Therefore, we need to delete the spam from the original retrieval result and only keep the documents of high quality.

3. **Extracting subtopics for every query from the retrieved documents.** In this step, we extract the query subtopics, that will be used to diversify the result, from the retrieved documents. Intuitively, each subtopic of the query should be a group of semantically related terms that are relevant to the query and can represent the content in the relevant documents. Therefore, we extract different groups of semantically related terms from the retrieved documents as subtopics and compute their weights based on their semantic relations with the query.

4. **Re-ranking documents based on diversity.** We use the similar greedy algorithm with the methods in [2, 19]. We start from an empty document list and iteratively select one document with highest diversity score on each position. We explore two diversity

functions to compute the diversity score of each document based on their similarity with subtopics, the importance of each subtopic and the novelty of each subtopic given the previously selected documents.

3 Implementation Details

3.1 Document Retrieval

The search result diversification methods proposed in this paper re-rank the result list according to the document diversity score based on their coverage of the query subtopics. The first step of these methods is to retrieve documents using the original query. These retrieved documents will be used in the subtopic extraction and document diversification described in the following sections.

Due to the limitation of disk space for the large collection, we use the search engine for Category A collection provided by CMU [1] to retrieve documents for the queries. However, the retrieval result just contains the ranked documents but does not contain their relevance scores. Therefore, we build a small working set with these returned documents. We then use Indri to build index of these documents and compute their relevance score given the query based on the language model with Dirichlet smoothed language model retrieval function [24].

3.2 Spam Filtering

ClueWeb09 collection has a lot of spam documents. The returned documents in the first step may also contain many spam documents that would hurt the performance of the system. Cormack et al. [10] studied the spam filtering method in ClueWeb09 collection and showed that the spam filtering can significantly improve the performance of the retrieval system. Therefore, we use their method to delete the spam from the returned documents in this step. We delete the documents that are in the top 70% most likely "spam" of the corpus computed by fusion method [10]. The non-spam retrieval result will be used to extract subtopics and diversified in the following steps.

3.3 Subtopic Extraction

Intuitively, each subtopic of the query is a group of semantically related terms that can represent a piece of relevant information of the query. Therefore, we extract different groups of terms that frequently co-occur in the retrieved documents and are relevant to the query as the subtopics.

We extract the groups of terms that frequently co-occur in the retrieved documents and assume that they can represent different pieces of the relevant content of the query. Each group is assumes to be a subtopic. We set the number of subtopics to be 5. After that, we compute the relevance score of each subtopic using the following equation [13]:

$$weight(w) = \frac{\sum_{q \in Q} weight_{idf}(q) \cdot sim(q, w)}{|Q|} \tag{1}$$

where Q is a query, $|Q|$ is the query length, w is a subtopic term and $sim(q, w)$ denotes the mutual information based term similarity.

This equation computes the relevance score of the subtopic according to its relation to the whole query.

Table 1: Official results of our submitted runs

	α-nDCG@5	α-nDCG@10	α-nDCG@20	P-IA@5	P-IA@10	P-IA@10
Dir	0.104168	0.122111	0.137771	0.043148	0.046481	0.037245
Dir+SpamFilter	0.220225	0.248365	0.295044	0.119444	0.113333	0.10875
$M1$	0.224209	0.272442	0.304459	0.148889	**0.169306**	**0.153426**
$M2$	0.2019	0.255856	0.289868	0.133056	0.153056	0.150347
$M3$	**0.256228**	**0.287373**	**0.310036**	**0.163333**	0.160602	0.143241

3.4 Document Diversification

Given the query subtopics, the next step is to diversify the retrieved documents to cover different subtopics of the query. We use the similar method with the methods in [2, 19]. We first iteratively select the documents with the highest score on each ranking position and generate a re-ranked list based on the original retrieval result. The main challenge in the method is to define the diversity function that compute the diversity score of the document. The diversity function in [19] is as follows:

$$P_{M1}(d|q, S_q, T) = \sum_{s \in S_q} P(s|q) P(d|s) \prod_{d' \in T} (1 - P(d'|s)) \tag{2}$$

where $P(d|q, S_q, T)$ is the diversity score of the document d given the query q, its subtopic set S_q and the previously selected document set T. $P(s|q)$ is the relevance score of a subtopic s, $P(d|s)$ is the similarity between the document and the query and $P(d'|s)$ is the similarity between the subtopic and one previously selected document d'. We denote the diversity system with Equation 2 as $M1$. It computes the diversity score of the document based on its relevance to the subtopic, the importance of the subtopic and the novelty of each subtopic given the previously selected documents. It favors documents covering subtopics that are not only important but also not well coved by previously selected documents.

We also explore two new diversity functions which are as follows:

$$P_{M2}(d|q, S_q, T) = \sum_{s \in S_q} P(s|q) \log(1 + \frac{P(d|s)}{1 + \sum_{d' \in T} P(d'|s)}) \tag{3}$$

$$P_{M3}(d|q, S_q, T) = \sum_{s \in S_q} P(s|q) P(d|s) \times (2 - 2 \times \sum_{d' \in T} P(d'|s) - P(d|s)) \tag{4}$$

Equation 3 and 4 use different methods to measures the diversity score of the document based on the importance and novelty of subtopics that are relevant to the document.

4 Experiment Results

We submitted three runs in the diversity task of web track. All of them are based on the Category A collection of ClueWeb09 corpus. They use greedy algorithm to select documents with highest diversity score on each position. They mainly differ in the diversity function to compute the diversity score. Their functions are Equation 2-4 where $M1$ is the method proposed by [19].

Table 1 lists the results of different diversity methods. *Dir* is the retrieval result of language model with Dirichlet prior. $Dir + SpamFilter$ is the result of deleting spam documents from *Dir*. $M1$, $M2$ and $M3$ are the results of three diversification methods. The result shows that the proposed $M3$ method performs best in all the methods. Table 2 list the number of queries

Table 2: The number of queries in different categories when comparing our runs with media runs of diversity task based on α-nDCG@10

	> median	= median	< median	total
$M1$	18	1	17	36
$M2$	17	1	18	36
$M3$	20	2	14	36

where the methods perform better, worse or the same comparing to the median results of all submitted runs in diversity task. The result is based on the queries that TREC released the statistical information of all the runs submitted to diversity task and the number of these queries is 36. We can see that $M3$ performs better than the media run in more than half of the queries. When judged on these 36 queries, the α-nDCG@10 of the media runs in diversity task is 0.265, the values of α-nDCG@10 of $M1$, $M2$ and $M3$ are 0.272, 0.256 and 0.287 respectively.

References

[1] Cmu indri search engine for clueweb09 category a collection. http://boston.lti.cs.cmu.edu:8085/clueweb09/search/cata_english/lemur.cgi.

[2] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of WSDM'09*, 2009.

[3] K. Balog, M. Bron, J. He, K. Hofmann, E. Meij, M. de Rijke, M. Tsagkias, and W. Weerkamp. The university of amesterdam at trec 2009. In *Proceedings of TREC'09*, 2009.

[4] W. Bi, X. Yu, Y. Liu, F. Guan, Z. Peng, H. Xu, and X. Cheng. Ictnet at web track 2009 diversity task. In *Proceedings of TREC'09*, 2009.

[5] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR'98*, pages 335–336, 1998.

[6] B. Carterette and P. Chandar. Probabilistic models of novel document rankings for faceted topic retrieval. In *Proceedings of CIKM'09*, 2009.

[7] H. Chen and D. R. Karger. Less is more: Probabilistic models for retrieving fewer relevant documents. In *Proceedings of SIGIR'06*, 2006.

[8] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the trec 2009 web track. In *Proceedings of TREC'09*, 2009.

[9] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of SIGIR'08*, 2008.

[10] G. V. Cormack, M. D. Smucker, , and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. In *Proceedings of TREC'2010*, 2010.

[11] N. Craswell, D. Fetterly, M. Najork, S. Robertson, and E. Yilmaz. Microsoft research at trec 2009. In *Proceedings of TREC'09*, 2009.

[12] Z. Dou, K. Chen, R. Song, Y. Ma, S. Shi, and J.-R. Wen. Microsoft research asia at the web track of trec 2009. In *Proceedings of TREC'09*, 2009.

[13] H. Fang and C. Zhai. Semantic term matching in axiomatic approaches to information retrieval. In *Proceedings of SIGIR'06*, 2006.

[14] S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *Proceedings of WWW'09*, 2009.

[15] Z. Li, F. Cheng, Q. Xiang, J. Miao, Y. Xue, T. Zhu, B. Zhou, R. Cen, Y. Liu, M. Zhang, Y. Jin, and S. Ma. Thuir at trec 2009 web track: finding relevant and diverse results for large scale web search. In *Proceedings of TREC'09*, 2009.

[16] R. McCreadie, C. Macdonald, I. Ounis, J. Peng, and R. Santos. University of glasgow at trec 2009: Experiments with terrier. In *Proceedings of TREC'09*, 2009.

[17] F. Radlinsk and S. T. Dumais. Improving personalized web search using result diversification. In *Proceedings of SIGIR'06*, 2006.

[18] F. Radlinski and S. Dumais. Improving personalized web search using result diversification. In *Proceedings of SIGIR'06*, 2006.

[19] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of WWW'10*, 2010.

[20] D. Yin, Z. Xue, X. Qi, and B. D. Davison. Diversifying search results with popular subtopics. In *Proceedings of TREC'09*, 2009.

[21] Y. Yue and T. Joachims. Predicting diverse subsets using structural svms. In *Proceedings of ICML'08*, 2008.

[22] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma. Learning to cluster web search results. In *Proceedings of SIGIR'04*, 2004.

[23] C. Zhai, W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of SIGIR'03*, 2003.

[24] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR'01*, 2001.