A Diagnostic Study of Search Result Diversification Methods

Wei Zheng
Department of Electrical and Computer Engineering
University of Delaware
Newark, DE USA
zwei@udel.edu

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware
Newark, DE USA
hfang@udel.edu

ABSTRACT

Search result diversification aims to maximize the coverage of different pieces of relevant information in the search results. Many diversification methods have been proposed and studied. However, the advantage and disadvantage of each method still remain unclear. In this paper, we conduct a diagnostic study over two state of the art diversification methods with the goal of identifying the weaknesses of these methods to further improve the performance. Specifically, we design a set of perturbation tests that isolate individual factors, i.e., relevance and diversity, which affect the diversification performance. The test results are expected to provide insights on how well each method deals with these factors in the diversification process. Experimental results suggest that some methods perform better in queries whose originally retrieved documents are more relevant to the query while other methods perform better when the documents are more diversified. We therefore propose methods to combine these existing methods based on the predicted factor of the query. The experimental results show that the combined methods can outperform individual methods on TREC collections.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Search process

General Terms: Algorithms, Performance, Experimentation

Keywords: Diversification, diagnostic, perturbation

1. INTRODUCTION

The goal of search result diversification is to diversify top-ranked documents so that they can cover different pieces of relevant information [1]. Various diversification methods have been proposed and studied [3, 6, 8–11]. Some methods focus on balancing the relevance and diversity of the documents [8, 9, 11], while others are more aggressive in diversifying documents [3, 6]. Although all of these methods are effective in diversifying search results, there is no clear winner. In order to further improve the diversification performance, it would be necessary to identify the strengths and weaknesses of each method and then study how to combine the strengths of different diversification methods.

In this paper, we apply the diagnostic evaluation [4] and design perturbation tests in order to better understand the strengths and weakness of two state of the art diversification methods. Since diversity and relevance are the two most important factors in diversification process, we design two perturbation tests to examine how the change of each factor could affect the performance of a diversification method. We then use existing methods to diversify documents in these perturbed collections, and observe the relationships between the properties of perturbed collections and the performances of different diversification methods. Our experimental results show that the perturbation tests can provide insights on these methods. In particular, we find that methods aggressively diversifying documents perform better when the originally retrieved documents are more relevant but less diversified, while the methods balancing the relevance and diversity would perform better when documents are less relevant. Based on these observations, we also proposed a method to predict the diversity and combine the two methods. The experimental results show that the combined methods can outperform individual methods on TREC collections. This shows the potential of combining different methods to improve the diversification performance.

2. RELATED WORK

We now briefly summarize a few related papers.

Fang et al. [4] compared traditional retrieval methods in perturbed collections. They adjusted document lengths, the number of query terms and the number of noisy terms to test the impact of different components. However, they focused on traditional retrieval methods instead of diversification methods. He et al. [5] divided queries into easy and difficult queries based on the performance of original retrieval results. They then compared performances of diversification methods in different queries. In our previous study [10], we gradually changed the quality of original retrieval results and compared performances of diversification methods on these results. We found that some methods, e.g., xQuAD, have larger gain when retrieval results are worse while other methods perform better in better retrieval results. However, we did not isolate important properties, i.e., relevance and diversity of the results, in that paper. Finally, our work

ICTIR '13, September 29 - October 02 2013, Copenhagen, Denmark
Copyright 2013 ACM 978-1-4503-2107-5/13/09 $15.00.
http://dx.doi.org/10.1145/2499178.2499190

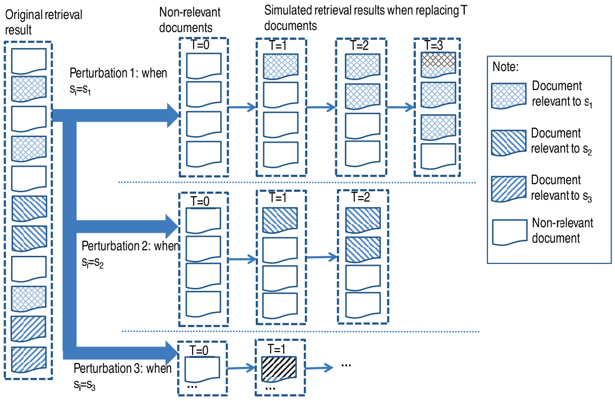
Original retrieval result
Non-relevant documents
Simulated retrieval results when replacing T documents
T=0
T=1
T=2
T=3
Perturbation 1: when s_i=s_1
Perturbation 2: when s_i=s_2
Perturbation 3: when s_i=s_3
...
Note:
Document relevant to s_1
Document relevant to s_2
Document relevant to s_3
Non-relevant document

Figure 1: Perturbation example of test 1 when using different subtopics as s_i.

is also related to selective diversification method [9], which used the machine learning method to dynamically balanced the relevance and diversity based on the query properties. However, they focused on a single diversification and tried to adjust the parameter dynamically. On the contrary, we tried to combine multiple methods to leverage the strength.

3. PERTURBATION TESTS

In this section, we perturb the original collections and focus on the two important factors in diversification process, i.e., relevance and diversity. We independently increase one factor while keeping the other unchanged. After that, we would be able to observe which methods perform better in queries with higher relevance and which perform better in queries with higher diversity.

3.1 Test 1: Increasing Relevance

We first separate these documents to two lists, i.e., a list of non-relevant documents and a list of relevant documents based on the judgment file. In each query, we start from a list of documents which replaces one document in the non-relevant list with a document that is relevant to only one subtopic s_i. In each step, we replace one non-relevant document with a document that is only relevant to s_i. We therefore get a simulated list of retrieved documents at each replacing step for each s_i. s_i can be any subtopic. Figure 1 shows this process when using different subtopics as s_i and replacing non-relevant documents from the top of the list. The total number of replacing steps is the maximum number of relevant documents in the original retrieval results that are relevant to only one subtopic. When using different subtopics as s_i in a query q, we have $|S(q)|$ perturbed lists at each replacing step, where $|S(q)|$ is the number of subtopics with relevant documents in q. Therefore, the results totally have $\sum_q |S(q)|$ simulated lists over all queries at each replacing step T.

The relevance of these lists gradually increases when the number of relevant documents, i.e., T, increases. However, in the whole process, the diversity of the list in the same query does not change since all relevant documents are covering the same subtopic. We will then use these lists of documents as simulated retrieval results and compare different methods by their performance in diversifying these lists of documents. This allows us to compare different methods in lists of documents with increased relevance.

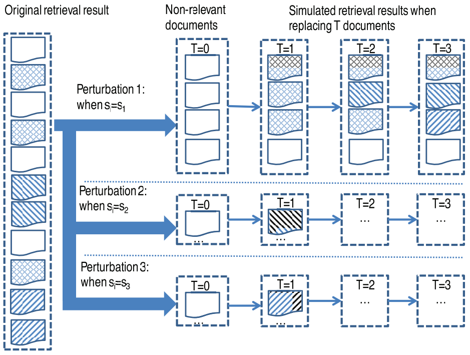
Original retrieval result
Non-relevant documents
Simulated retrieval results when replacing T documents
T=0
T=1
T=2
T=3
Perturbation 1: when s_i=s_1
Perturbation 2: when s_i=s_2
Perturbation 3: when s_i=s_3
...

Figure 2: Perturbation example of test 2 when using different subtopics as s_i.

3.2 Test 2: Increasing Diversity

For each query, we start from the list replacing $|S(q)|$ non-relevant documents with $|S(q)|$ documents which are all relevant to a subtopic s_i. s_i can be any subtopic in $S(q)$. We then iteratively go to each subtopic s_j, where $s_j \neq s_i$, and replace one relevant document of s_i in the list with a document relevant to s_j. Figure 2 shows perturbed lists when using different subtopics as s_i. The results also have $\sum_q |S(q)|$ simulated retrieval lists in all queries at each step. However, there are only $max_q(|S(q)|)$ replacing steps which is the maximum number of subtopics with relevant documents in the queries. This number is much smaller than the number of steps in increasing relevance.

The diversity of the lists gradually increases while the relevance does not change. The reason is that we always have the same number of relevant documents, i.e., $|S(q)|$, in the same query although we cover more subtopics.

4. PERTURBATION RESULTS

4.1 Experiment Design

We conduct the perturbation tests over the TREC 2009 diversity task collection [1] to diagnose two state of the art diversification methods, i.e., $xQuAD$ [9] and $PM2$ [3]. To isolate the effect of subtopic quality in the diversification process, we use the real subtopics from the relevance judgements.

At each perturbation step T, we use these methods to diversify each list of documents and evaluate their performance by 0.5-$nDCG$@20. We aggregate the results of all lists on the same step T since they have the same relevance in test 1 and diversity in test 2. Therefore, each replacing step will be represented as one data point. As described earlier, the number of points in test 2 is very small, which could make it difficult to observe patterns with very smaller number of points. We instead use a new set of queries that merges N original queries. Each new query contains the subtopics of N original queries and their retrieved documents. The relevance of documents with regard to the subtopic keeps the same. Therefore, each new query has N times subtopics comparing to the original query. We randomly pick 100 new subtopics from the set of all possible

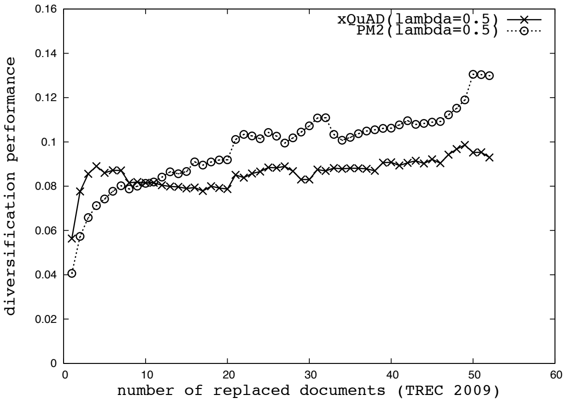

Figure 3: Results of test 1 (increasing relevance)

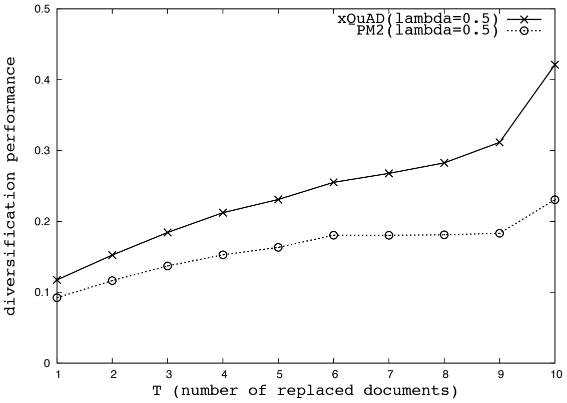

Figure 4: Results of test 2 (increasing diversity)

.

new queries in order to reduce the time complexity. We therefore do the perturbation test on these new queries. We only report the results when we set N to 4. The results when using different values for N are similar.

4.2 Experimental Results

Figures 3 shows the result of increasing relevance. It is clear that $PM2$ performs worse than $xQuAD$ when T is small and performs better when T is large. It implies that $PM2$ performs better in queries whose original retrieval results have higher relevance while $xQuAD$ performs better in queries whose original retrieval results have lower relevance. Therefore, when we combine these two methods, we should use $PM2$ when we know that original retrieval result has higher relevance. Otherwise, we use $xQuAD$.

Figure 4 shows the results of increasing diversity. $xQuAD$ outperforms $PM2$ more when T is larger.It indicates that $xQuAD$ performs better than $PM2$ in queries whose original retrieval results have higher diversity.

5. COMBINATION OF DIVERSIFICATION METHODS

5.1 Combination Methods

In this section, we propose methods to combine $PM2$ and $xQuAD$ to diversify original queries. We propose two methods that combine them based on the observed patterns in the previous section.

- Relevance-based combination. In each query, we use $xQuAD$ if the relevance of the original retrieval result is smaller than a threshold γ. Otherwise, we use $PM2$. The threshold γ is a parameter that will be tuned in the experiments.
- Diversity-based combination. We use $PM2$ if the diversity of the original retrieval result in the query is smaller than γ. Otherwise, we use $xQuAD$.

One concern in diversity-based combination is that whether we should use $PM2$ when diversity is small, since $PM2$ always perform worse than $xQuAD$ in increasing diversity test on perturbed collections. We think that we still need to use it since the perturbed collections are quite different from the original collection. Actually, the experimental results also show the effectiveness of $PM2$ in diversity-based combination.

The only unsolved problem in the combination methods is how to know the relevance and diversity of the original retrieval result of the query. There have been a lot of studies in predicting relevance of retrieval results in the traditional information retrieval systems [2]. We therefore focus on predicting the diversity of the original retrieval result which is the unique property in search result diversification systems. We predict the diversity using the distance between subtopics. Since we do not know the subtopics, we can use document clusters as subtopics. We first use PLSA [7] to cluster originally retrieved documents of the query and terms in these documents. We set the number of clusters to 15. We then compute the similarity between each pair of clusters using their terms and term weights. We propose two methods to predict diversity based on the similarities between clusters.

- Average similarity between clusters. This method computes the diversity of the query based on the average similarity between clusters as follows:

$$Div_{avg}(q) = \frac{1}{avgSim(CP)} = \frac{|CP|}{\sum_{\{C_i,C_j\}\in CP} sim(C_i,C_j)}, \quad (1)$$

  where CP contains all pairs of PLSA clusters in the query, C_i and C_j are two different clusters, $sim(C_i, C_j)$ is their similarity which is cosine similarity of their terms.

- Minimum cluster similarity. The diversity is computed based on the minimum similarity between clusters as follows:

$$Div_{min}(q) = \frac{1}{minSim(CP)} = \frac{1}{\min_{\{C_i,C_j\}\in CP} sim(C_i,C_j)}. \quad (2)$$

We use $PM2$ when $Div_{avg}(q)$ or $Div_{min}(q)$ is smaller than γ. Otherwise, we use $xQuAD$ in the queries.

5.2 Experimental Results

We first use the judgment file to test whether the observed patterns still work in original queries on TREC 2009 collection. We use 0-$nDCG$@20 and 1-$nDCG$@20 to measure the relevance and diversity, respectively. Table 1 showed the optimal performance of relevance-based combination method,

Table 1: Optimal performance (0.5-$nDCG$@20) of combination in original queries with judgment file.

	$PM2$	$xQuAD$	$RCombine$	$DCombine$
TREC09	0.3147	0.3141	0.3207	**0.3347**

Table 2: Optimal performance (0.5-$nDCG$@20) of diversity-based combination.

	TREC09	TREC10	TREC11
$PM2$	0.3147	0.3455	0.5335
$xQuAD$	0.3141	0.3651	0.5464
$AvgCombine$	0.3241	**0.3739**	**0.5540**
$MinCombine$	**0.3324**	0.3710	0.5516

i.e., *RCombine*, and diversity-based combination, i.e., *DCombine*. We can see that both combination methods can outperform $PM2$ and $xQuAD$.

We then combine methods based on proposed prediction methods. Table 2 shows the optimal performances of diversity-based combination. *AvgCombine* and *MinCombine* are the combination methods based on average similarity and minimum similarity between clusters, respectively, as described in Section 5.1. We can have two interesting observations. (1) Both combined methods outperform individual methods of $PM2$ and $xQuAD$. It shows the potential of combined methods in improving performances of retrieval systems. (2) *MinCombine* performs better than *AvgCombine* on TREC09, while *AvgCombine* performs a little bit better than *MinCombine* on TREC10 and TREC11.

The optimal performances of combination methods are better than $PM2$ and $xQuAD$. However, the combination method cannot outperform $PM2$ and $xQuAD$ when we test them on TREC10 and TREC11 collection with parameters tuned on TREC09 collection. To investigate the reason, we show the results for sensitivity of the threshold γ. Figure 5 shows the performance of *MinCombine* with different values of γ on all collections. We can see that the trend on TREC09 is very different from trends of TREC10 and TREC11. This is the reason why parameters tuned on TREC09 cannot perform well on the other collections.

We might get better testing performance when tuning γ on different collections. For example, if we tune γ on TREC10, the testing performances are 0.3214, 0.3710 and 0.5481 on TREC09, TREC10 and TREC11 collections, respectively. All of them outperform corresponding $PM2$ and $xQuAD$. An interesting study would be to compare different collections, and dynamically adjust the combination methods when applying them to different collections. We leave this to our future work.

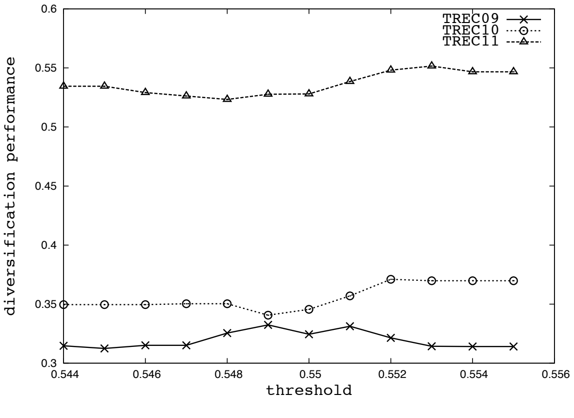

Figure 5: Performance of $MinCombine$ with different values of γ.

6. CONCLUSION AND FUTURE WORK

In this paper, we first studied the relationships between query properties, i.e., relevance and diversity, and the performances of different methods. We found that $PM2$ performed better than other methods in the perturbed queries whose originally retrieved documents were more relevant, while $xQuAD$ performs better when originally retrieved documents were more diversified. We applied observed patterns to combine $PM2$ and $xQuAD$ based on the predicted diversity in the query. The experimental results showed that the combined methods outperformed the individual methods.

There are several interesting directions for the future work. (1) We will study more properties of the query itself besides the properties of the originally retrieved documents. (2) It would be interesting to test more combination methods. We can use more sophisticated methods such as machine learning to combine individual methods.

7. REFERENCES

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

[2] S. Cronen-Townsend and W. B. Croft. Quantifying query ambiguity. In *Proceedings of HLT'02*, 2002.

[3] V. Dang and W. B. Croft. Diversity by proportionality: An election-based approach to search result diversification. In *Proceedings of SIGIR'12*, 2012.

[4] H. Fang, T. Tao, and C. Zhai. Diagnostic evaluation of information retrieval models. *ACM Transactions on Information Systems*, 29:1–42, 2011.

[5] J. He, M. Bron, and M. D. Rijke. A query performance analysis for result diversification. In *Proceedings of 3rd International Conference on the Theory of Information Retrieval*, 2011.

[6] J. He, E. Meij, and M. de Rijke. Result diversification based on query-specific cluster ranking. *Journal of The American Society for Information Science and Technology*, 62(3):550–571, 2011.

[7] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of UAI'99*, 1999.

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

[9] R. L. T. Santos, C. Macdonald, and I. Ounis. Selectively diversifying web search results. In *Proceedings of CIKM'10*, 2010.

[10] W. Zheng and H. Fang. A comparative study of search result diversification methods. In *Proceedings of DDR'11*, 2011.

[11] W. Zheng, X. Wang, H. Fang, and H. Cheng. Coverage-based search result diversification. *Information Retrieval*, 15:433–457, 2012.