A Comparative Study of Search Result Diversification Methods

Wei Zheng and Hui Fang

University of Delaware, Newark DE 19716, USA
zwei@udel.edu, hfang@ece.udel.edu

Abstract. Top-ranked documents returned by traditional retrieval functions may cover the same piece of relevant information and cannot satisfy different user needs. Search result diversification solves this problem by diversifying results to cover more information needs, i.e., query subtopics, in top-ranked documents. Many diversification methods have been proposed and studied, and most of them re-rank original retrieved documents according to both relevance and diversity functions in a probabilistic framework. Although official TREC results make it possible to compare the effectiveness of different diversification systems, it remains unclear whether the better performance of a system comes from better diversification methods or component estimation methods. In this paper, we conduct a systematic study on comparing three representative diversification methods which can be implemented using probabilistic methods. We not only analytically compare the methods but also conduct empirical studies and evaluate the effectiveness of these methods in a controlled manner.

1 Introduction

Traditional retrieval functions ignore the relations among returned documents. As a result, top ranked documents may contain relevant yet redundant information. In order to maximize the satisfaction of different search users, it is necessary to diversify search results.

Many diversification methods have been proposed. For example, Carbonell and Goldstein [2] proposed the maximal marginal relevance (**MMR**) ranking strategy to balance the relevance and the redundancy among the returned documents. Yin et. al. [7] derived a diversification method using language modeling approach, i.e., **WUME**. Santos et. al. [6] proposed a probabilistic framework, i.e., **xQuAD**, that estimates the diversity based on the relevance of documents to query subtopics and the importance of query subtopics. The first method is a classical method and has been widely cited, but none of the top-ranked diversity systems from TREC used this method. The last two methods were implemented in the systems participating in TREC 2009 Web track. Although the evaluation results of these two methods are quite different according to the official TREC results [3], it is unclear whether the performance differences are caused by the underlying diversification methods or the ways of estimating the component functions in the methods.

In this paper, we conduct a systematic study to compare the above three representative diversification methods both analytically and empirically. Specifically, we first analyze the methods and summarize their commonalities and differences. These methods mainly differ in *diversity modeling*, i.e., whether the diversity is implicitly modeled through document similarities or explicitly modeled through the coverage of query subtopics, and *document dependency*, i.e., whether the diversity score of a document is related to other documents or not. To make a more meaningful empirical comparison, we modify the three methods under the same framework and use the variants of the three methods in this paper. All of the variants of the methods re-rank the original retrieved documents based on a linear combination of relevance and diversity scores. The variants also use the same methods to estimate components in their functions. This would allow us to focus on the differences in the diversification methods. Moreover, following the idea of diagnostic evaluation [5], we conduct four sets of experiments using simulated collections. Our goal is to not only compare different diversification methods but also study how the performance of a diversification method can be impacted by different factors, i.e., the quality of relevant functions, the tradeoff between relevance and diversity, and the number of query subtopics.

Experiment results show that *diversity modeling* has a large impact on the effectiveness of a diversification method. Explicitly modeling the diversity with query subtopics is more effective than implicitly modeling the diversity through document similarities. As an example, MMR performs worse than the other two methods consistently. Moreover, *document dependency* has a smaller impact on the diversity performance. Although computing the diversity score of a document based on other documents is intuitively desirable, the empirical performance gain is small. Finally, we can also make the following interesting observations.

- The effectiveness of a diversification method is closely related to the effectiveness of its relevance function. In particular, the performance improvement of the diversification method decreases as the performance of the relevance function increases.
- The number of query subtopics affects the diversity performance of the methods that explicitly model the diversity based on subtopics. However, they may still achieve reasonably good performance when the quality of subtopics is good and the number of missed subtopics is small.

2 Analytical Comparisons of Diversification Methods

Most of existing diversification methods first retrieve a set of documents based on only their relevance scores, and then re-rank the documents so that the top-ranked documents are diversified to cover more query subtopics [2–4, 6, 7]. Since the problem of finding an optimum set of diversified documents is NP-hard [1], a greedy algorithm is often used to iteratively select the diversified document.

In this paper, we focus on three representative diversification methods discussed in the previous section.

- *MMR* [2]: It maximizes the margin relevance of the documents and iteratively select the document that is not only relevant to the query but also dissimilar to the previously selected documents.
- *WUME* [7]: It maximizes the probability that the document meets the user needs. Its diversification function iteratively selects the document that covers both the query and the important subtopics of the query.
- *xQuAD* [6]: It uses the probability model to maximize the combination of the likelihood of a document is observed given the query and the likelihood of the document while not the previously selected documents is observed given the query. It iteratively selects the document that is not only relevant to the query but also covers the subtopics that have not be well covered by previously selected documents.

All these three methods iteratively select the document that is not only relevant to the query but also diversified to cover more query subtopics, explicitly or implicitly. Therefore, all of them fit into a general framework that iteratively selects the document with the highest relevance and diversity scores [2, 6, 1]:

$$d^* = \arg\max_{d \in D \setminus D'} (\lambda \times (Rel(d,q) + (1-\lambda) \times Div(d,q,D'))) \tag{1}$$

where D is a set of documents that need to be re-ranked, D' is the set of previously selected documents, λ is a parameter that balances the relevance score of the document i.e., $Rel(d,q)$, and the diversity score $Div(d,q,D')$.

We then implement the variants of these methods under the framework and they are referred to as MMR^*, $WUME^*$ and $xQuAD^*$:

1. Maximal marginal relevance (*MMR*) variant method [2]:

$$Div_{MMR^*}(d,q,D') = -\max_{d' \in D'} p(d|d') \tag{2}$$

2. *WUME* variant method [7]:

$$Div_{WUME^*}(d,q,D') = \sum_{s \in S(q)} p(s|q)p(d|s) \tag{3}$$

3. Explicit query aspect diversification (*xQuAD*) variant method [6]:

$$Div_{xQuAD^*}(d,q,D') = \sum_{s \in S(q)} p(s|q)p(d|s) \prod_{d' \in D'} (1 - p(d'|s)) \tag{4}$$

$S(q)$ is the subtopic set of query q. $p(d|d')$ measures the similarity between current document and selected document, $p(d|s)$ measures the similarity between the document and the subtopic, $p(s|q)$ measures the importance of the subtopic in the query and $\prod_{d' \in D'} (1 - p(d'|s))$ is the subtopic importance penalization component that penalizes the importance of the subtopic that has been covered in previously selected documents.

In $WUME^*$, we split the probability of the document given both the query and subtopics existing in $WUME$, in order to make it comparable with the other methods. We consider the probability of the document given the query in $Rel(d, q)$ and the probability of the document given the subtopics in $Div_{WUME^*}(d, q, D')$. Other main differences between the original diversification methods and these variants are how to estimate the component functions in the methods and how to find query subtopics. Since we focus on comparing different diversity functions, we use query subtopics given in the judgment file as $S(q)$ and use the same method to estimate the components.

Comparing these three diversity functions, we can see that they mainly differ in two aspects. The first aspect is the *diversity modeling*. MMR^* implicitly models the diversity through document similarities and ignores the information about query subtopics. On the contrary, the other two methods explicitly model the diversity through the coverage of query subtopics. The second aspect is the *document dependency*. $WUME^*$ assumes that the diversity score of a document is independent of other documents while the other two methods assume that the diversity score depends on the previously selected documents.

Intuitively, it is more reasonable to explicit use subtopics to model diversity and assume that the documents are dependent of each other. Therefore, $xQuAD^*$ should perform best and the performance of $WUME^*$ would be the second best. However, it is unclear whether both explicit subtopics and document dependence have big effects on the diversification results, and whether the difference between the diversification methods is significantly. We will compare their performances in the following section.

3 Experiments

In our experiment, we use the TREC09 and TREC10 collections [3], each of which has 50 queries, and the Category B of ClueWeb09 collection that contains 428 million documents. We use α-nDCG@100, together with α-nDCG@20 used in TREC, as the measures to evaluate the diversification results. The reason is that we want to observe the performance of a longer document ranking list. We use the Dirichlet retrieval function [9] to retrieve the original results and compute the probabilities in Equation (1)-(4). We use the real subtopics given in the judgment file for diversification in explicit subtopic based methods. We then design the experiments to study the following questions: (1) the optimum performances of diversification methods; (2) the impact of retrieval performance of the original ranking on diversification results; (3) the impact of parameters, i.e., tradeoff between diversity and relevance, and number of subtopics.

3.1 Comparison of diversification methods

In this section, we test whether using explicit subtopics and document dependence can significantly perform better. Table 1 shows the optimum performances

	TREC09 result		TREC10 result	
	α-nDCG@20	α-nDCG@100	α-nDCG@20	α-nDCG@100
MMR^*	0.365	0.427	0.344	0.415
$WUME^*$	0.479	0.546	0.579	0.630
$xQuAD^*$	0.482	0.550	0.588	0.636

Table 1. The performances of diversification methods when using all real subtopics for diversification

of the diversification methods both on the original TREC09 and TREC10 collections. All the parameters in each method are set to the optimum values. We can see that both $xQuAD^*$ and $WUME^*$ perform significantly better than MMR^*. It shows that using explicit subtopics in diversification is better than implicit subtopics, which is consistent with the observation in [6]. However, the performances of $xQuAD^*$ and $WUME^*$ are not significantly different. It tells that the component of subtopic importance penalization in Equation 4 of $xQuAD^*$ needs to be modified to further improve the performance. We leave this study for our future work.

3.2 Impact of original retrieval result quality

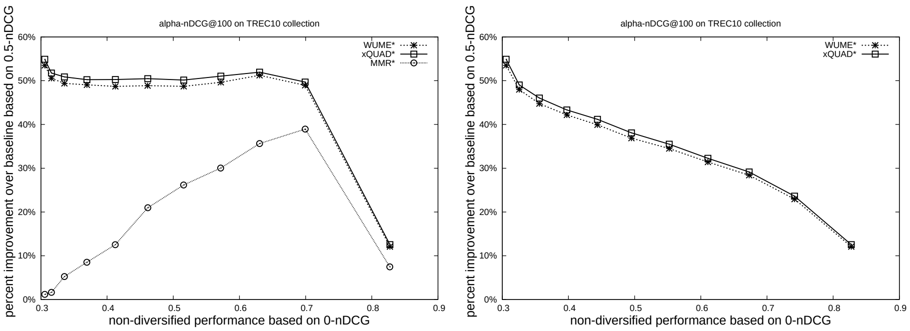

Fig. 1. The percentage improvements of diversification methods over non-diversified methods that combined all relevant documents with non-relevant documents selected from the top results (left) or random selected (right) from the original retrieval result.

We now test the impact of original retrieval result quality on diversification results. Due to the space limitation, we only show results of TREC2010 while ignore TREC2009 that has similar trend in the following experiments. We simulate the original retrieval results with different relevance qualities, evaluated by $0 - nDCG$. We combine all the relevant documents in the judgment file with N non-relevant documents selected from the top documents in the original retrieval result in each query. We then re-compute the relevance scores of all these documents given the query. The simulated retrieval result only contains the relevant documents when N is 0 and is the same as the original retrieval result when N

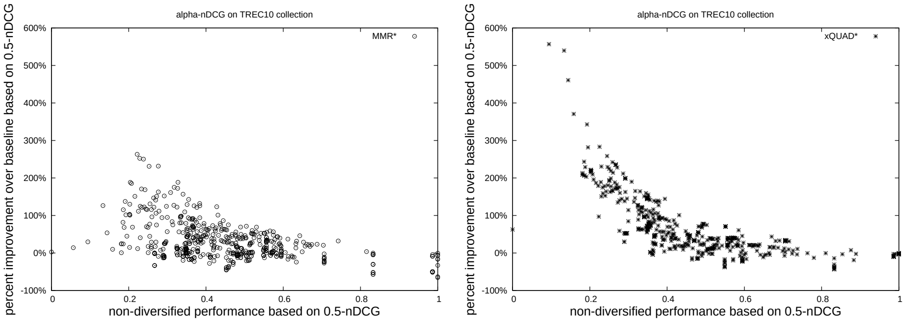

Fig. 2. The improvements of diversity methods in each query

is 100. The left plot of Figure 1 shows the performance improvements of different diversification methods in diversifying these simulated retrieval results. The values of N corresponding to points from left to right on each line are 100, 90, ..., 10 and 0, respectively.

There are three interesting observations in the plot. (1) $xQuAD^*$ and $WUME^*$ can consistently outperform MMR^*. What's more, the difference between MMR^* and the other two methods is bigger when the original retrieval result is worse. MMR^* aggressively selects the document that are most different from the previously selected document. This helps diversify the relevant documents but also selects more non-relevant documents when the original result is worse. The reason is that many non-relevant documents are less similar to relevant documents [8] and the non-relevant documents themselves are also different. (2) The performance differences between $WUME^*$ and $xQuAD^*$ are always small.

We also use the other method to select the N non-relevant documents and compare these two methods on the new stimulate retrieval results. We randomly select these non-relevant documents 10 times from the original retrieval result for each value of N. We then diversify each 10 results corresponding to the same value of N and use their average relevance performance to represent the performance of that value of N. The right plot of Figure 1 shows the performances of $WUME^*$ and $xQuAD^*$. Their performances are still similar. It again shows that a new method to penalize the subtopic importance is needed to further improve the performance. (3) The worse the non-diversified method performs, the larger the improvement of diversification is. The reason is that these methods can use the subtopics to not only diversify relevant documents but also rank non-relevant documents lower when the quality of non-diversified result is poor. It is also interesting to study the improvement trend of diversification methods with different diversity performances of the original retrieval method, evaluated by $0.5-nDCG$. Figure 2 shows the improvements of $xQuAD^*$ and MMR^* over baseline in each query with simulated retrieval results. We can also see that $xQuAD^*$ performance has larger gain when the diversity quality of the query is worse.

3.3 Impact of parameters in diversification

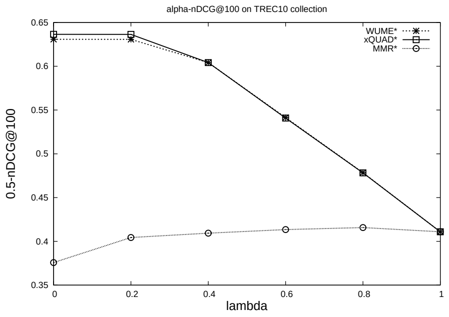

Fig. 3. Impact of λ on the diversification performance

There are two parameters in the diversification methods. One is λ that balances relevance score and diversity score in Equation 1. The other is the number of subtopics in explicit subtopic based methods. We use the original retrieval result and the real subtopics in the judgement to tune these parameters. Figure 3 shows the impact of λ on different methods when using all real subtopics. The smaller the value of λ is, the more the methods are focusing on diversity. The optimum value of λ in methods based on explicit subtopics is 0. The subtopics used in these methods are the real subtopics in judgment file and therefore they can achieve optimum performance without considering the document relevance with the original query. However, the optimum value of λ may not be 0 if they do not use real subtopics and use other methods to extract subtopics from the collection.

In the above experiment, we use all real subtopics in diversification. However, the extracted subtopics in methods based on explicit subtopics may be incomplete. Therefore, we study the impact of the number of subtopics while using the optimum value of λ in each method. We randomly select $n\%$ real subtopics for diversification in each query. We extract each possible combination of real subtopics for each value of n. For each value of n, we evaluate the diversification performance of using its subtopic sets and use the average performance to represent the diversification performance corresponding to that value of n.

Figure 4 shows the diversification performance using incomplete subtopic set. The improvements of $WUME^*$ and $xQuAD^*$ decrease when the percentage of missed real subtopics decreases, but they can still outperform MMR^*. What's more, their performance decrease is not significant when the percentage of missed real subtopics is small, i.e., 20%. The right plot in Figure 4 shows the percentage of queries in different categories when comparing the diversification using $n\%$ of real subtopics and that using all real subtopics. When n is greater or equal to 80, the result of using these incomplete subtopics is very close to the result using all real subtopics, which shows that the explicit subtopic modeling methods are robust to the quality of subtopics and can still achieve reasonably good performance when their extracted subtopics do not contain all real subtopics.

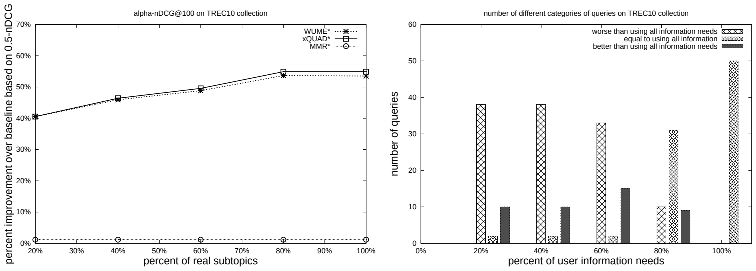

Fig. 4. Performance improvement (left) and query comparison (right) when using $n\%$ of real subtopics for diversification

4 Conclusion

In this paper, we revisited the existing diversification methods based on the language model and systematically compared their diversity functions. We compared the diversity modeling and document dependency strategies used in diversification functions. The experiment result shows that the explicit subtopic modeling and subtopic importance penalization strategies perform better but the effect of the penalization is small. It is also interesting to find that the explicit subtopic based methods are robust to the number of subtopics and can still achieve reasonable good performance when missing a small number of real subtopics.

References

1. R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of WSDM'09*, 2009.
2. 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.
3. C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the trec 2009 web track. In *Proceedings of TREC'09*, 2009.
4. M. Drosou and E. Pitoura. Search result diversification. In *Proceedings of SIGMOD'2010*, 2010.
5. H. Fang, T. Tao, and C. Zhai. Diagnotic evaluation of information retrieval models. *ACM Transactions of Information Systems*, To Appear.
6. R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of WWW'10*, 2010.
7. D. Yin, Z. Xue, X. Qi, and B. D. Davison. Diversifying search results with popular subtopics. In *Proceedings of TREC'09*, 2009.
8. C. Zhai, W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of SIGIR'03*, 2003.
9. 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.