Diversifying search results with popular subtopics

Dawei Yin, Zhenzhen Xue, Xiaoguang Qi, Brian D. Davison
Department of Computer Science & Engineering, Lehigh University
Bethlehem, PA 18015 USA
{day207,zhx208,xiq204,davison}@cse.lehigh.edu

25 October 2009

Abstract

This paper describes the method we use in the diversity task of web track in TREC 2009.

The problem we aim to solve is the diversification of search results for ambiguous web queries. We present a model based on knowledge of the diversity of query subtopics to generate a diversified ranking for retrieved documents. We expand the original query into several related queries, assuming that query expansions expose subtopics of the original query. Moreover, each query expansion is given a weight which reflects the likelihood of the interpretation (the fraction of users who issued this query given the general query topic). We issue all those expanded queries including the original query to a standard BM25 search engine, then re-rank the retrieved documents to generate the final ranking. Our method can detect possible subtopics of a given query and provide a reasonable ranking that satisfies both relevancy and diversity metrics. The TREC evaluations show our method is effective on the diversity task.

1 Introduction

Ambiguous queries are those having more than one interpretation. When ranking documents relevant to an ambiguous query, a search engine should not only consider document relevancy, but provide documents that satisfy different interpretations of the query. A good ranking for an ambiguous query should maximize the satisfaction of average users by covering a variety of subtopics (more specific topics) in which searchers could be interested. In order to generate such a ranking, our system makes use of prior knowledge on subtopics of a query and statistical information of user intent on these subtopics.

For most of the current web search engines, there are two main problems in the diversity task. The first is the granularity of diversity (that is, the level of diversity). The second is how to rank the retrieved documents. Previous research usually models the diversity task as needing to pick a diverse subset from all subtopics, but not to consider the ranking order among these retrieved documents.

In this paper, we introduce a model which is based on understanding the diversity of subtopics to achieve an optimal ranking for retrieved documents. Our model tries to optimize the whole retrieval system. It issues multiple queries to obtain diverse results. It does not only try to return a set of documents which covers all the subtopics of a given query, but also uses statistical information of users' intention on different subtopics to rerank the retrieved documents.

2 Related Work

There have been some previous results on diversity task. Carbonell and Goldstein [2] firstly introduced a preliminary model for diversity based reranking–maximal marginal relevance (MMR). Their model does not only focus on the relevance of the documents but also maximizes dissimilarity among the retrieved documents. Vee et al. [11] formalized diversity in the structured search. They proposed algorithms to return diversified results by using B+ tree. However, the algorithms are not suitable for unstructured search such as web search. In recommendation, diversity also is a problem which is very similar to the retrieval system. Ziegler et al. [14] improved recommendation system through involving topic diversification. They reviewed the main metrics of evaluation in recommendation systems and designed their metrics which involve diversification. For each product candidate, their algorithm measures dissimilarity between the item and the rest of product candidates and combines this dissimilarity to the original relevance order. Yu et al. [12] introduced an explanation-based diversity algorithm for recommendation systems and formalized the diversification problem as a compromise between accuracy and diversity. The algorithm tries to maximize the diversity under relevance constraints. Diversity task is also closely related to the redundancy detection. Chen et al. [3] considered that users are satisfied with some limited number of relevant documents, rather than needing all relevant documents. Their basic idea is that documents should be selected sequentially according to the probability of the document being relevant conditioned on the documents that come before.

Researchers also have gained some results on the query understanding and used it to improve the search result. Song et al. [10] studied ambiguous queries. Their experiments showed that 16% of the queries are ambiguous queries. Radlinski and Dumais [8] used the query log to extract some related queries to a given query.

Using these related queries, they simply diversify and rerank the results. It is shown that their method is promising to improve the performance of a search engine. Hu et al. [7] design a method to predict user's query intent using Wikipedia. They predefine 3 domains: namely, personal name, travel and job. For each domain, they use random walk on the graph created from Wikipedia to generate a probability vector. Then when a query comes in, it will map to a set of articles in Wikipedia and then add the probability together; if the summation exceeds a threshold, the query will be considered as the intent of that domain.

Recently, Gollapudi and Sharma [5] present an approach to characterizing diversification system using a set of natural axioms. The choice of axioms presents a method that can make objective functions independent with the distance and relevance functions. Agrawal et al. [1] formulate the problem of diversifying search formulation theoretically. The idea is to assume that users only consider the top k returned results. The objective is to maximize the probability that average users find at least one useful result within the top k results. They prove this problem is a NP-hard problem and also propose a greedy algorithm.

For the evaluation on diversity task, Zhai et al. [13] proposed a framework for evaluating subtopic retrieval which generalizes the traditional precision and recall metrics by accounting for intrinsic topic difficulty as well as redundancy. They also present two ways to measure the novelty and then combine them with relevance under the strategy of MMR. Clarke et al. [4] also propose a evaluation framework: alpha-nDCG. They improve the famous evaluation method nDCG by adding diversity and novelety under probability ranking principle (PRP).

3 Method

Unlike most previous methods, our method directly incorporates user intent data into the ranking model. In specific, we discover subtopics of a given query from frequently issued similar queries and human knowledge collections; and then estimate user's interest in each subtopic using previous query frequencies.

Figure 1 shows an example to illustrate how our algorithm works. In this example, there are three subtopics under the given query c_1, c_2, and c_3. Under these subtopics, there are web pages p_1 through p_8 that are relevant to at least one subtopics. Although p_3 is more relevant to one of the subtopics c_2 than p_5 to c_3, given that c_2 attracts less user interest than c_3, p_3 should still be ranked lower than p_5.

We formalize our method below. Given a query q, the probability that a re-

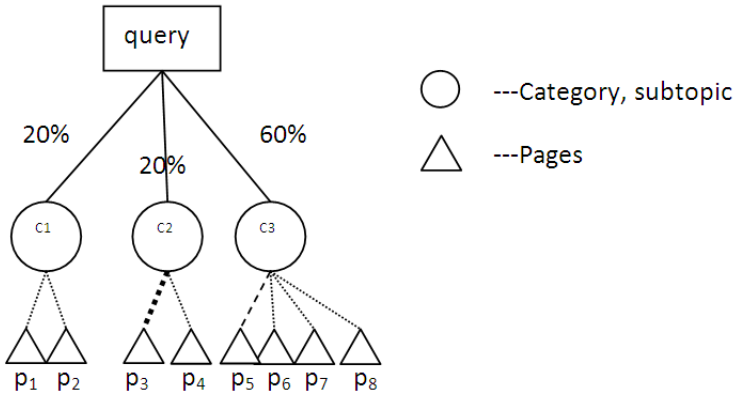

Figure 1: An example of user intent sensitive ranking.

trieved document d meets user's intent can be written as

$$P(I|q,d) = \frac{P(I=1|q)P(d|I=1,q)}{P(d|q)} \tag{1}$$

Since every document d for a given query q is retrieved by a basic retrieval method, we assume that, without considering user's intent, the probability of each retrieved document given the query $P(d|q)$ is the same accross all retrieved documents. Thus, we ignore $P(I=1|q)$ and $P(d|q)$ in the following, which leads us to

$$P(d|I=1,q) \propto P(I|q,d) \tag{2}$$

Now we take subtopic information into consideration, where c_i $(i = 1..k)$ represents a subtopic associated with query q.

$$\begin{aligned} P(d|I=1,q) &= P(d|c_1,I=1,q) \times P(c_1|I=1,q) \\ &\quad +P(d|c_2,I=1,q) \times P(c_2|I=1,q) \\ &\quad +\ldots \\ &\quad +P(d|c_k,I=1,q) \times P(c_k|I=1,q) \\ &= \sum_c P(d|c,I=1,q) \times P(c|I=1,q) \end{aligned}$$

Thus, we have

$$P(d|I=1,q) \propto \sum_c P(d|c,I=1,q) \times P(c|I=1,q) \tag{3}$$

From 1 and 3, we have

$$P(I|q,d) \propto \sum_{c} P(d|c, I=1, q) \times P(c|I=1, q) \tag{4}$$

We generate the final ranking using the above computed $P(I|q,d)$ value. Therefore, in order to rank the retrieved documents, we need to discover the set of subtopics c_i, and estimate $P(d|c, I=1, q)$ and $P(c|I=1, q)$ for each c_i. We will describe them in the following section.

4 Experiments

In the previous section, we transformed the problem of ranking retrieved documents into estimating two probabilities $P(d|c, I=1, q)$ and $P(c|I=1, q)$ for each subtopic c_i of a given query. Here, we describe our method to estimate them.

4.1 Estimate $P(c|I=1, q)$

We used Google Insights for Search [6] and Wikipedia to extract the possible subtopics. From Google Insights for Search, we can obtain a set of related queries as well as their relative weights. Such queries can be considered candidate subtopics. Figure 2 shows the information we can get for the query "map" from Google Insights for Search. The relative weight for each related queries gives us a convenient way to estimate $P(c|I=1, q)$. However, as we can see from the example, some of the related queries cannot be seen as reasonable subtopics.

We also obtain the list of articles from Wikipedia disambiguation pages for a given query, and use these articles as candidate subtopics. Due to its human maintained nature, the list of subtopics obtained from Wikipedia is more reasonable than the list from Google. However, the lack of relative weight as present in Google introduces a challenge to estimate $P(c|I=1, q)$.

We also use Wikipedia to try to refine these related queries. Wikipedia is a knowledge pool which is maintained manually. The list of article titles in a disambiguation page can be treated as subtopics of the title for this disambiguation page. We use disambiguation pages to find the related queries which are more reasonable than those from Google Insight for Search. However, there are no relative weights on the related queries from Wikipedia. Then we merge the sets of related queries.

For a given query q, we denote the set of candidate subtopics obtained from Google as $Q_g(q)$; the subtopics from Wikipedia, $Q_w(q)$. The two sets of subtopics are merged using the following rules.

1. $r \in Q_g(q) \cap Q_w(q)$

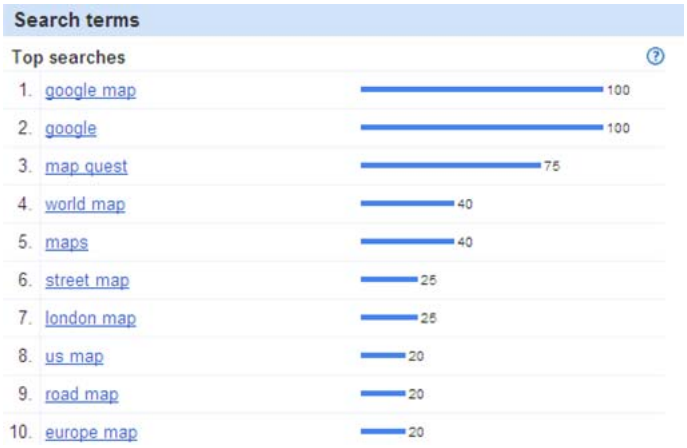

Figure 2: Information from Google's Insight for Search for query "map".

We just use it as a subtopic and its relativity as $P(c|I = 1, q)$

2. $r \in Q_g(q) - Q_w(q)$

   We consider that it may not be reasonable and drop it.

3. $r \in Q_w(q) - Q_g(q)$

   Since most of the related queries from Wikipedia are reasonable, they should be kept as subtopics. However, the problem is that we cannot estimate $P(c|I = 1, q)$. Here, it does not appear in $Q_g(q)$, so we assume that there may be very few people who want this subtopic. We just put the minimal relativity (the set of relativity numbers which come from the top 10 related queries $Q_g(q)$)on this related query.

Note that besides the subtopics obtained using the above method, we also consider the original query q as one of the subtopics of itself, and assign it the maximal weight. Then, we normalize the weights of subtopics to sum to 1. Such weights are used as the estimated value for $P(c|I = 1, q)$.

4.2 Estimate $P(d|c, I = 1, q)$

For each subtopic, we treat it as a query to retrieval 1000 documents by use the standard BM25 [9]. Then, the scores which are generated by BM25 are normalized to sum to 1. We use these normalized the BM25 scores as the probability $P(d|c, I = 1, q)$.

	>median	=median	<median
alpha-nDCG	28	13	9
IA	27	16	7

Table 1: For each query, we compare our evaluation scores with the median scores. >median means the number of queries where our score is better than the median score. =median means the number of queries where our score is equal to the median scores. <median means the number of queries where our score is worse than the median score. There are total 50 queries.

For each document, after we get $P(d|c, I = 1, q)$,$P(c|I = 1, q)$, we combine and rerank all the documents which are retrieved by each subtopic by using equation 4.

4.3 TREC Diversity Results

The TREC organizers provide two performance metrics. The first is alpha-nDCG which are defined by Clarke et al.[4] In evaluation, alpha-nDCG@10 is used and the parameter alpha = 0.5. The second is an "intent aware"(IA) version of precision as defined by Agrawal et al.[1], where all intents are given equal weight. For each query,we compare our evaluation scores to the median scores of all participants. There are total 50 queries in the test data. Table 1 show the results of comparing our scores with the median scores. The results show that the evaluation results are similar in two measurement. For more than half queries, the evaluation socres of our methods are better than the median, for more than ten queries, our scores are equal to the median scores and for only less than queries, our scores are worse than the median scores. Overall, the result show that our method is effective and but still need be improved.

5 Summary and Future Work

We studied the problem of diversifying search results for ambiguous web queries. Understanding the sub topic of the ambiguous queries and user intent on these subtopics are very important for this problem. Our method maximizes the satisfaction of average users by ranking documents relevant to a variety of usubtopics considering the probability of user intent given that query. The TREC evaluations show our method is effective on the diversity task.

Further work is still needed. The first problem is how to better estimate $P(c|I = 1, q)$ and $P(d|c, I = 1, q)$. In our experiment, for $P(c|I = 1, q)$, we

only borrow the information from Google Insights for Search and Wikipedia. In the future, we plan to mine query log and document content to obtain potential subtopics. Second, we did not detect duplicates. Duplicate detection is expected to affect our diversity method. In the future, we plan to detect near duplicate documents before the stage of diversification. Finally, we will consider finer-grained granularity of diversity based on enough user intent data. Multi-level diversity model can also be built.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Number IIS-0545875.

References

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14, New York, NY, USA, 2009. ACM.

[2] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, New York, NY, USA, 1998. ACM.

[3] H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 429–436, New York, NY, USA, 2006. ACM.

[4] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666, New York, NY, USA, 2008. ACM.

[5] S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 381–390, New York, NY, USA, 2009. ACM.

[6] Google insight for search, 2009. http://www.google.com/insights/search/.

[7] J. Hu, G. Wang, F. Lochovsky, J.-t. Sun, and Z. Chen. Understanding user's query intent with wikipedia. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 471–480, New York, NY, USA, 2009. ACM.

[8] F. Radlinski and S. Dumais. Improving personalized web search using result diversification. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 691–692, New York, NY, USA, 2006. ACM.

[9] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proc. of the 17th Annual Int'l ACM SIGIR Conf. on Research and Development in Info. Retrieval*, pages 232–241, 1994.

[10] R. Song, Z. Luo, J.-R. Wen, Y. Yu, and H.-W. Hon. Identifying ambiguous queries in web search. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 1169–1170, New York, NY, USA, 2007. ACM.

[11] E. Vee, U. Srivastava, J. Shanmugasundaram, P. Bhat, and S. A. Yahia. Efficient computation of diverse query results. In *ICDE '08: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pages 228–236, Washington, DC, USA, 2008. IEEE Computer Society.

[12] C. Yu, L. Lakshmanan, and S. Amer-Yahia. It takes variety to make a world: diversification in recommender systems. In *EDBT '09: Proceedings of the 12th International Conference on Extending Database Technology*, pages 368–378, New York, NY, USA, 2009. ACM.

[13] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 10–17, New York, NY, USA, 2003. ACM.

[14] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 22–32, New York, NY, USA, 2005. ACM.