Improving Personalized Web Search using Result Diversification

Filip Radlinski
Cornell University
Ithaca, NY, USA
filip@cs.cornell.edu

Susan Dumais
Microsoft Research
Redmond, WA, USA
sdumais@microsoft.com

ABSTRACT

We present and evaluate methods for diversifying search results to improve personalized web search. A common personalization approach involves reranking the top N search results such that documents likely to be preferred by the user are presented higher. The usefulness of reranking is limited in part by the number and diversity of results considered. We propose three methods to increase the diversity of the top results and evaluate the effectiveness of these methods.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms: Algorithms, Measurement

Keywords: Personalized web search, Result diversity

1. INTRODUCTION

Personalizing search results for individual users is increasingly being recognized as an important future direction for web search [3, 5, 6, 7]. Providing results specific to individual users is particularly important because different users expect different information even given the same query [4]. One proposed approach involves providing a user profile to the search engine, which can then use it to bias search results toward the user's interests. However, this requires the search engine to perform the personalization at additional computational expense, and requires that the user trusts the search engine with her profile. In this work, we focus on an alternative approach that runs entirely client-side, where the client requests a larger number of search results and reranks them such that documents more likely to interest the user are presented higher [3, 5].

The primary limitation of client side reranking is that the system can only rerank the top N results. While reranking may allow effective personalization when web pages of particular interest to the user are present, it cannot be effective if all top N results are similar. Anagnostopoulos et al. recently proposed a method to sample search results to avoid homogeneity [1]. We propose an alternative of using query-query reformulations to understand the variety of user intents and improve the effectiveness of client-side reranking. By observing how large numbers of users modify their search queries, we see which kinds of results tend to be missing from the top of search results, from the user's perspective. For example, by looking at logs from a large web search engine, we observed that the query "windows" is often followed by specializations such as "windows xp" or clarifications such as "house windows". This suggests that if we want to personalize results for a user who issued the query "windows", we may also want to consider results for both of these reformulations. We believe that analyzing query-query reformulations adds interesting diversity within the result set by focusing on user intents that are not well represented in the original results. We also believe that such a method could be used to diversify general web search results, although we do not address this question here.

We first present our general strategy for diversifying search results using query-query reformulations, then propose how to select the reformulations to consider. Next, we describe a method to measure result diversity. Finally, we present an evaluation of the effectiveness of our approach.

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.
ACM 1-59593-369-7/06/0008.

2. DIVERSIFICATION METHODS

Assume that we want to personalize search by reranking 100 results using query reformulations to introduce diversity into those results. Given a query q, we generate a set of k related queries $R(q)$. We then take $\frac{100}{k+1}$ results from each query in $R(q)$ and from q. This gives D, a set of 100 results to rerank. For our experiments we used $k \in \{0,\ 2,\ 4,\ 9,\ 19\}$. When k=0, the top 100 results from the original query are considered for reranking, and when k=19, the top 5 results from q and from 19 reformulations are considered. We now describe the data and algorithms used to select $R(q)$.

To obtain query-query reformulations, we analyzed a large sample of the query logs from a popular web search engine over about 6 weeks. For each query q_i we measured n_i, the number of times the query was observed, and p_i, the empirical probability that q_i was followed by any other query within a thirty minute time window. For a pair of queries (q_i, q_j), let n_{ij} be the number of times q_i was followed by q_j. $p_{ij} = \frac{n_{ij}}{n_i}$ is the empirical probability of q_i being followed by q_j. p^*_{ij} is the related symmetric measure $p^*_{ij} = \sqrt{p_{ij}p_{ji}}$.

We developed three methods for generating $R(q)$. The Most Frequent (MF) method sets $R(q_i)$ to the queries q_j with highest n_{ij}. These are the queries that most often follow q_i. The Maximum Result Variety (MRV) method greedily selects queries that are both frequent reformulations (using p_{ij}) and different from other queries that have already been selected (using p^*_{jk}). We used a weighted combination of these two factors, $\arg\max_{q_j} \lambda p_{ij} - (1-\lambda) \max_{q_k \in R(q_i)} p^*_{jk}$, with $\lambda = 0.5$. MRV is motivated by the MMR measure of Carbonell and Goldstein [2] and aims to select a set of queries that are related to q_i yet different from each other.

Finally, the Most Satisfied (MS) method sets $R(q_i)$ as the set of queries q_j with minimum p_j and $p_{ij} > 0.001$ and $n_{ij} \geq 2$. This method finds queries that tend not to be further reformulated yet occur with some minimum frequency.

3. DIVERSITY EVALUATION

Let $match(d_i, u)$ measure how well document d_i matches the interests of the user u. The maximum match in D, $diversity(D) = \max_{d_i \in D} match(d_i, u)$, reflects the extent to which at least one result is very similar to a user's interests. We used the average value of $diversity(D)$ across all users as a measure of diversity for each method. We also looked at measuring diversity using the top 2-5 match scores (rather than just the maximum).

We measure $match(d_i, u)$ using the relevance-feedback approach for reranking developed by Teevan et al. [5]. They proposed a modification of the standard BM25 weighting scheme, in which relevance information is obtained from a local representation of a user's interests:

$$w_t = \log \frac{(r_t + 0.5)(N - n_t + 0.5)}{(n_t + 0.5)(R - r_t + 0.5)},$$

where N is the number of documents in the corpus, R is the number for which we have relevance feedback, and n_t and r_t are the number of documents in N and R that contain the term t. We computed these weights for both individual words (unigrams) and for pairs of adjacent words (bigrams):

$$match_{unigram}(d_i, u) = \sum_{t \in d_i} w_t$$

$$match_{bigram}(d_i, u) = \sum_{t_i, t_j \in d_i} w_{t_i, t_j}$$

To compute these measures, we estimate the four parameters for both unigrams and bigrams. N and n_t were computed from a sample of 1.5 billion web pages. R and r_t were computed for each user using a full text index of the files, emails and web pages to represent their interests as in [5].

4. EXPERIMENT AND RESULTS

We evaluated these methods for 33 volunteers. Figure 1 shows the bigram match results for the diversification methods and five values of k. The MS method did not generate enough reformulations for some of the user-specific queries so we omit it for simplicity. The unigram match and alternative diversity measures follow the same general trends.

The evaluation was performed on two types of queries. The lower three curves show the results for a set of 30 fixed queries chosen from the search engine log. The queries varied in frequency, topic, typical reformulation patterns, etc. The upper two curves show the results for the most recent queries in each user's browser history, averaging 76 queries per user. The main effect of query type (fixed, user) is reliable ($F(1, 32) = 4.82, p = 0.022$), showing that the match score for queries of interest to the user are higher. The main effect of diversification method is marginally reliable ($F(1, 32) = 3.30, p = 0.079$), with MRV leading to somewhat higher diversity scores. The main effect of k is reliable ($F(4,128) = 3.82, p = 0.006$), showing that diversity scores increase as the number of reformulations considered increases. Interestingly for the MF method the first few reformulations reduce the result diversity. This suggests that the most frequent reformulations are not very different in topic from the original query. Even with this small initial dip, the linear correlation between k and diversity score is strong and significant ($r = 0.90, p = 0.037$).

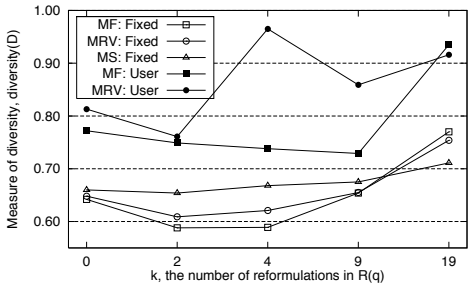

Figure 1: Evaluation of MF, MRV and MS diversity methods on a fixed query set (Fixed) as well as on queries taken from users' web browser cache (User).

Incorporating results from reformulations leads to higher computational cost for each query. However we expect the load to increase sub-linearly in k since results for common reformulations will be in cache, and a user will reformulate less often if they are satisfied with their first query.

5. CONCLUSIONS

In this poster, we presented a number of methods to collect diverse results for a given query using past query reformulations. Our evaluation suggests that this is a promising method to improve personalized reranking of search results. We will next evaluate these diversification techniques in an end-to-end web search personalization system.

We would like to thank Eric Horvitz for useful discussions, Ed Cutrell for assistance with statistical analysis, Robert Ragno for assistance with collecting data and the volunteers in the user study.

6. REFERENCES

[1] A. Anagnostopoulos, A. Z. Broder, and D. Carmel. Sampling search engine results. In *International World Wide Web Conference*, pages 245–256, 2005.

[2] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Poster at SIGIR*, 1998.

[3] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *International World Wide Web Conference*, pages 675–684, 2004.

[4] J. Teevan, S. T. Dumais, and E. Horvitz. Beyond the commons: Investigating the value of personalizing web search. In *Workshop on New Tech. for Personalized Information Access (PIA)*, pages 84–92, 2005.

[5] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *SIGIR 2005*, pages 449–456, 2005.

[6] Y. Zhang, J. Callan, and T. Minka. Novelty and redundancy detection in adaptive filtering. In *SIGIR 2002*, pages 81–88, 2002.

[7] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *International World Wide Web Conference*, pages 22–32, 2005.