The Impact of History Length on Personalized Search

Yangbo Zhu
yangboz@cs.cmu.edu

Jamie Callan
callan@cmu.edu

Jaime Carbonell
jgc@cs.cmu.edu

Language Technologies Institute, School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213, USA

ABSTRACT

Personalized search is a promising way to better serve different users' information needs. Search history is one of the major information sources for search personalization. We investigated the impact of history length on the effectiveness of personalized ranking. We carried out task-based user study for Web search, and obtained ranked relevance judgments for all queries. Query contexts derived from previous queries in the same task are used to re-rank results for the current query. Experimental results show that the performance of personalization generally improves as more queries are accumulated, but most of the benefits come from a few immediately preceding queries.

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

General Terms: Measurement, Experimentation.

Keywords: Personalized Search, Web Search, Evaluation

1. INTRODUCTION

Web search queries are usually short and ambiguous. Personalized search aims to better understand information needs by modeling user interests. User profiles are built from various sources including searching, clicking and browsing history. This study focuses on the utilization of short-term context [2] for personalization. We try to answer one research question: How many related queries do we need to understand the current query?

The effectiveness of traditional information retrieval systems is evaluated using binary relevance judgments. Web search systems use multivalued relevance assessments because there are many relevant documents, with varying degrees of relevance. Teevan et al. [3] obtained relevance ratings on a 3-point scale (highly relevant, relevant, non-relevant), and used Normalized Discounted Cumulative Gain (NDCG) [1] to characterize the potential of personalizing search. In this paper, we exploit the usage of ranked judgments, which is much finer-grained than multivalued ones. Compared with binary or multivalued judgments, ranked judgments are more expensive to obtain. However, they can be approximately induced from large amounts of click-through data.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

2. DATA COLLECTION

We conducted a pilot user study with two participants, both of whom are familiar with Web search. We designed 20 information seeking tasks, with 10 general-purpose and 10 technical ones. General-purpose tasks are those we encounter in our daily lives, such as vacation planning, home improvement and gift searching. Technical tasks include literature review, learning a new programming language, etc. Each participant was asked to pick 6 general-purpose tasks and 6 technical tasks that she felt best match her real life scenarios.

In order to complete a task, the participants self-generated a sequence of queries, and sent them to a search portal. The portal sent queries to a large commercial Web search engine, and presented top 10 results in a random order. For each query, the participant identified relevant and non-relevant results, and ranked relevant ones according to her preferences. For example, a typical set of ranked relevance judgments are (2,1,3,0,0), where the first three results are relevant, and the second result is the best from the user's perspective. The two participants evaluated 87 and 77 queries respectively, with 6.8 queries for each task on average. Table 1 shows four sample tasks and the corresponding queries.

Table 1: Sample tasks and user-generated queries

Tasks	Queries
Hobbies	photography, digital photography, photo composition, depth of field
Artists	Emma Lazarus, Emma Lazarus books, statue of liberty, Emma Lazarus poems
Skills	poster design, scientific poster design, Adobe Illustrator poster
Programming	Python tutorials, Python pros and cons, Python vs Java, Python performance

3. PERSONALIZATION

We follow one common paradigm for personalized search, which consists of three steps: (1) identify related queries for the current query, (2) construct a query context using these related queries and their search results, and (3) use the context to re-rank the original search results. Since we are interested in the impact of history length, we need to eliminate the influence of other factors. Therefore, we make two idealistic assumptions in our experiments: (a) related queries are accurately identified (queries for the same task are of course related), (b) the best result for each query can

be inferred using click-through data.

Given a query, we form its context by aggregating the user-ranked best results from previous queries for the same task. The context is a term vector with TFIDF term weighting. For each result of the current query, we calculate its cosine similarity to the context. We map the original ranking r to exponentially decreasing score $s = a^{-r}, a > 1$. Linear combinations of the original ranking scores and similarity scores are used to re-rank the top 10 results.

4. EVALUATION

The ranked relevance judgments enable us to calculate not only traditional binary style metrics, but also rank-based metrics. We use two existing and one novel metrics to measure the quality of personalized ranking.

Since the original NDCG [1] is sensitive to the base of the logarithm, we use a widely used variation of NDCG:

$$NDCG = \frac{1}{f(n)} \sum_{i=1}^{n} \frac{2^{v_i} - 1}{log(i+1)} \quad (1)$$

where n is the number of results, v_i is the relevance value of the ith result, $f(n)$ is a normalization factor. We map ranked judgments linearly to relevance values. For example, the best result has value 10, while the non-relevant ones share value 0.

Kendall-Tau distance is widely used for comparing two ranked lists. It is defined as the number of pair-wise disagreements normalized by the total number of pairs. However, it does not consider the relative importance of different results.

In Web search, putting highly relevant results at the top is extremely important. We propose a general weighted rank distance (WRD) to measure the quality of document ranking.

$$WRD = \frac{1}{z(n)} \sum_{i=1}^{n} w(R_i) d(R_i, i) \quad (2)$$

where n is the length of either list, R_i is the user ranking of the ith result, w is a non-increasing weighting function, d is a pair-wise distance function, and $z(n)$ is the normalizer that ensures $WRD \in [0, 1]$. $z(n)$ is calculated using the worst case (reversed user ranking). Note that when $d(R_i, i) = (R_i - i)^2$ and $w(R_i) = 1$, WRD resembles Spearman's rank correlation. In this paper, we let $d(R_i, i) = |R_i - i|$, $w(R_i) = R_i^{-1}$ for WRD(1), and $w(R_i) = R_i^{-2}$ for WRD(2).

We are interested in the impact of accumulating history on personalization. When history length is h, the context for current query is derived from the immediately preceding h query. Since we collected 164 queries from 24 tasks, there are 140 queries with at least one previous query. Table 2 presents the performance of personalization with varying history length. Considering the immediate preceding 1 or 2 queries improves the performance, but adding longer history only provides small further improvement, and occasionally hurts performance.

Why longer history does not help in our experiments? The value of personalization primarily lies in its disambiguation power. For perfectly clear queries, there is little room for personalization. For ambiguous queries, we need to find the right disambiguating direction. Longer history provides richer contextual information, but also brings in more noise. Our study suggests two common patterns [4] in the user behavior of forming queries. One pattern is from general queries to specific ones (e.g., "photography" in Table 1). During the process of completing a task, users get to know more specific terminologies. The other pattern is exploring different aspects of a given topic (e.g., "Python" in Table 1). Although these aspects are related, the useful vocabulary for one aspect might not be useful for another. We must be very careful when we use results from related aspects to bias the ranking of the current query.

Table 2: Improvement with varying length of history (For each metric, the three rows are the original ranking, the personalized ranking and the relative improvement. For NDCG, the higher the better. For the other three metrics, the lower the better.)

History Length	1	2	3	4	5
# of Queries	140	116	92	68	44
NDCG	0.643	0.655	0.659	0.687	0.652
	0.677	0.708	0.710	0.736	0.689
	5.3%	8.1%	7.7%	7.2%	5.7%
Kendall-Tau	0.333	0.337	0.338	0.320	0.335
	0.316	0.314	0.311	0.290	0.298
	5.1%	6.8%	8.2%	9.3%	11.0%
WRD(1)	0.599	0.601	0.593	0.582	0.616
	0.573	0.557	0.545	0.529	0.565
	4.4%	7.4%	8.0%	9.1%	8.3%
WRD(2)	0.545	0.550	0.541	0.536	0.573
	0.504	0.478	0.460	0.447	0.486
	7.5%	13.2%	14.8%	16.6%	15.1%

5. CONCLUSION

Empirical results suggest that incorporating short-term contexts works well for personalized search. Just a few preceding queries can provide valuable information for search engines to better understand the current query. In our experiments, considering longer history does not provide much further improvement. To take advantage of ranked relevance judgments, we propose weighted rank distance to measure ranking quality.

6. ACKNOWLEDGMENTS

This research was supported in part by National Science Foundation grant IIS-0707801. Any opinions, findings, conclusions, or recommendations are the authors' and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] Kalervo Järvelin and Jaana Kekäläinen. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proceedings of SIGIR 2000*, 41–48.

[2] Xuehua Shen and Bin Tan and Chengxiang Zhai. Context Sensitive Information Retrieval Using Implicit Feedback. In *Proceedings of SIGIR 2005*, 43–50.

[3] Jaime Teevan and Susan Dumais and Eric Horvitz. Characterizing the Value of Personalizing Search. In *Proceedings of SIGIR 2007*, 757–758.

[4] Tessa Lau and Eric Horvitz. Patterns of Search: Analyzing and Modeling Web Query Refinement. In *Proceedings of UM 1999*, 119–128.