Reexamination on Potential for Personalization in Web Search

Daren Li[1] Muyun Yang[1] Haoliang Qi[2] Sheng Li[1] Tiejun Zhao[1]

[1]School of Computer Science
Harbin Institute of Technology
{drli|ymy|tjzhao}@mtlab.hit.edu.cn, lisheng@hit.edu.cn

[2]School of Computer Science
Heilongjiang Institute of Technology
haoliang.qi@gmail.com

Abstract

Various strategies have been proposed to enhance web search through utilizing individual user information. However, considering the well acknowledged recurring queries and repetitive clicks among users, it is still an open issue whether using individual user information is a proper direction of efforts in improving the web search. In this paper, we first quantitatively demonstrate that individual user information is more beneficial than common user information. Then we statistically compare the benefit of individual and common user information through Kappa statistic. Finally, we calculate potential for personalization to present an overview of what queries can benefit more from individual user information. All these analyses are conducted on both English AOL log and Chinese Sogou log, and a bilingual perspective statistics consistently confirms our findings.

1 Introduction

Most of traditional search engines are designed to return identical result to the same query even for different users. However, it has been found that majority of queries are quite ambiguous (Cronen-Townsend et al., 2002) as well as too short (Silverstein et al., 1999) to describe the exact informational needs of users. Different users may have completely different information needs under the same query (Jansen et al., 2000). For example, when users issue a query "Java" to a search engine, their needs can be something ranging from a programming language to a kind of coffee.

In order to solve this problem, personalized search is proposed, which is a typical strategy of utilizing individual user information. Pitkow et al. (2002) describe personalized search as the contextual computing approach which focuses on understanding the information consumption patterns of each user, the various information foraging strategies and applications they employ, and the nature of the information itself. After that, personalized search has gradually developed into one of the hot topics in information retrieval. As for various personalization models proposed recently, Dou et al. (2007), however, reveal that they actually harms the results for certain queries while improving others. This result based on a large-scale experiment challenges not only the current personalization methods but also the motivation to improve web search by the personalized strategies.

In addition, the studies on query logs recorded by search engines consistently report the prevailing repeated query submissions by large number of users (Silverstein et al., 1999; Spink et al., 2001). It is reported that the 25 most frequent queries from the AltaVista cover 1.5% of the total query submissions, despite being only 0.00000016% of unique queries (Silverstein et al., 1999). As a result, the previous users' activities may serve as valuable information, and technologies focusing on common

Coling 2008: Poster Volume, pages 701–709,
Beijing, August 2010

user information, such as collaborative filtering (or recommendation) may be a better resolution to web search. Therefore, the justification of utilizing individual user information deserves further discussion.

To address this issue, this paper conducts a bilingual perspective of survey on two large-scale query logs publically available: the AOL in English and the Sogou[1] in Chinese. First we quantitatively investigate the evidences for exploiting common user information and individual user information in these two logs. After that we introduce Kappa statistic to measure the consistency of users' implicit relevance judgment inferred from clicks. It is tentatively revealed that using individual user information is what requires web search to face with after common user information is well exploited. Finally, we study the distribution of potential for personalization over the whole logs to generally disclose what kind of query deserves for individual user information.

The remainder of this paper is structured as follows. Section 2 introduces previous methods employing individual and common user information. In Section 3, we quantitatively compare the evidences for exploiting common user information and individual user information. In Section 4, we introduce Kappa statistic to measure the consistency of users' clicks on the same query and try to statistically present the development direction of current web search. Section 5 figures out utilizing individual user information as a research issue after well exploiting common user information. Section 6 presents the potential for personalization curve, trying to outline which kind of queries benefit the most from individual user information. Conclusions and future work are detailed in Section 7.

2 Related Work

With the rapid expansion of World Wide Web, it becomes more and more difficult to find relevant information through one-size-fits-all information retrieval service provided by classical search engines. Two kinds of user information are mainly used to enhance search engines: common user information and individual user information. We separately review the previous works focusing on using these two kinds of information.

Among various attempts to improve the performance of search engine, collaborative web search is the one to take advantage of the repetition of users' behaviors, which we call common user information. Since there is no unified definition on collaborative web search, in this paper, we believe that the collaborative web search assumes that community search activities can provide valuable search knowledge, and sharing this knowledge facilitates improving traditional search engine results (Smyth, 2007). An important technique of collaborative web search is Collaborative Filtering (CF, also known as collaborative recommendation), in which, items are recommended to an active user based on historical co-occurrence data between users and items (Herlocker et al., 1999). A number of researchers have explored algorithms for collaborative filtering and the algorithms can be categorized into two classes: memory-based CF and model-based CF. Memory-based CF methods apply a nearest-neighbor-like scheme to predict a user's ratings based on the ratings given by like-minded users (Yu et al., 2004). The model-based approaches expand memory-based CF to build a descriptive model of group-based user preferences and use the model to predict the ratings. Examples of model-based approaches include clustering models (Kohrs et al., 1999) and aspect models (J. Canny, 2002).

The other way to improve web search is personalized web search, focusing on learning the individual preferences instead of others' behaviors, which is called individual user information. Early works learn user profiles from the explicit description of users to filter search results (Chirita et al., 2005). However, most of users are not willing to provide explicit feedback on search results and describe their interests (Carroll et al., 1987). Therefore, recent researches on the personalized search focus on modeling user preference from different types of implicit data, such as query history (Speretta et al., 2005), browsing history (Sugiyama et al., 2004), clickthrough data (Sun et al., 2005), immediate search context (Shen et al., 2005) and other personal information (Teevan et al.,

[1] A famous Chinese search engine with a large number of Chinese web search users.

2005). So far, there is still no proper comparison between the two solutions. It is still an open question which kind of information is more effective to build the web search model.

Considering the difficulty in collecting private information, using individual user information seems less promising as the cost-effective solution to web search. To address this issue, some researches about the value of personalization have been conducted. Teevan et al. (2007) have done a ground breaking job to quantify the benefit for the search engines if search results were tailored to satisfy each user. The possible improvement by the personalized search, named potential for personalization, is measured by a gap between the relevance of individualized rankings and group ranking based on NDCG. However, it is less touched for the position of individual user information in contrast with common user information in large scale query log and how to balance the usage of common and individual information in information retrieval model.

This paper tentatively examines individual user information against common user information on two large-scale search engine logs in following aspects: the evidence from clicks on the same query, Kappa statistic for the whole queries, and overall distribution of queries in terms of number of submissions and Kappa value. The bilingual statistics consistently reveals the tendency of using individual user information as an equally important issue as (if not more than) using common user information) issue for researches on web search.

3 Quantitative Evidences for Using Common or Individual User Information

To quantitatively investigate the value of common user information and individual user information in query log, we discriminate the evidence for using the two different types of user information as follows:

(1) Evidence for using common user information: if there were multiple users who have exactly the same click sets on one query, we suppose those clicks sets, together with the query, as the evidence for exploiting common user information. It is clear that such queries are able to be better responded with other's search results. Note that common user information is hard to be clearly defined, in order to simplify the quantitative statistics we give a strict definition. Further analysis will be shown in following sections.

(2) Evidence for using individual user information: if a user's click set on a query was not the same as any other's, for that query, the search intent of the user who issue that query can be better inferred from his/her individual information than common user information. We suppose this kind of clicks, together with the related queries, as the evidence for exploiting individual user information.

Since users may have different search intents when they issue the same query, a query can be an evidence for using both common and individual user information. In our statistics, if a query has both duplicate click sets and unique click set, the query is not only counted by the first category but also the second category.

The statistics of the two categories are conducted in the query log of both English and Chinese search engines. We use a subset of AOL Query Log from March 1, 2006 to May 31, 2006 and Sogou Query Log from March 1, 2007 to March 31, 2007. The basic statistics of AOL and Sogou log are shown in Table 1. Notice that the queries in raw AOL and Sogou log without clicks are removed in this study.

Item	AOL	Sogou
#days	92	31
#users	6,614,960	7,488,754
#queries	7,840,348	8,019,229
#unique queries	4,811,649	4,580,836
#clicks	12,984,610	17,607,808

Table 1: Basic statistics of AOL & Sogou log

Table 2 summarizes the statistics of different evidence categories over AOL and Sogou log. Note that click set refers to the set of clicks related to a query submission instead of a unique query. As for evidence for using common and individual user information, there is no clear distinction in terms of number of records, number of users in two logs. However, in terms of unique query and distinct click set, one can't fail to find that evidence for using individual user information clearly exceeds

Log	The Condition		Number			
	Repeated queries	Click	Records	User	Unique Query	Distinct Click Set
AOL	3,745,088 (47.77% of total query submissions)	Same	2,438,284	277,416	382,267	461,460
		Different	2,563,245	343,846	542,593	1,349,892
Sogou	4,252,167 (53.02% of total query submissions)	Same	2,469,363	1,380,951	228,315	358,346
		Different	5,481,832	1,545,817	752,047	2,171,872

Table 2: Different click behaviors on repeated queries

that for using common user information, especially in Sogou log. Therefore, though making use of common and individual user information can address equally well for half users and half visits to the search engine, the fact that much more unique queries and click sets actually claims the significance of needing individual user information to personalize web results. And methods exploiting individual user information provide a much more challenging task in terms of problem space, though one may argue utilizing common user information is much easier to attack.

4 Kappa Statistics for Individual and Common user information

Section 3 has shown the evidence for using individual user information is prevailing than common user information in quantity for the unique queries in search engines. However, these counts deserve a further statistical characterization. In this section, we introduce Kappa statistic to depict the overall consistency of users' clicks in query logs.

4.1 Kappa

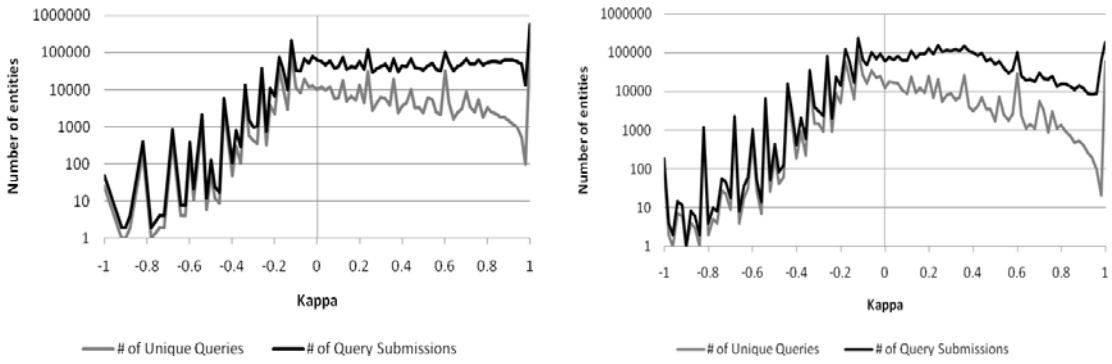
Kappa is a statistical measure introduced to access the agreement among different raters. There are two types of Kappa. One is Cohen's Kappa (Cohen, 1960), which measures only the degree of agreement between two raters. The other is Fleiss's Kappa (Fleiss, 1971), which generalizes Cohen's Kappa to measure agreement among more than two raters, denoted as:

$$\kappa = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}}$$

where, $\overline{P}$ is the probability that a randomly selected rater agree with another on a randomly selected subject. $\overline{P_e}$ is the expected probability of agreement if all raters made ratings by chance. If we use Kappa to measure the consistency of relevance judgment by different raters, $\overline{P}$ can be interpreted as the probability that two random selected raters consistently rate a random selected search result as relevant or non-relevant one. Similarly, $\overline{P_e}$ can also be construed as the expected probability of identical relevance judgment rated by different raters all by chance.

Teevan et al. (2008) used Fleiss's Kappa to measure the inter-rater reliability of different raters' explicit relevance judgments. We expand their work and employ Fleiss's Kappa to measure the consistency of implicit relevance judgments by users on the same query[2]. Here clicks are treated as a proxy for relevance: documents clicked by a user are judged as relevant and those not clicked as non-relevant (Teevan et al., 2008). As we all know that the result set of one query may change over time, so we select the longest time span to calculate Kappa value of a query, during which the result set of it preserves unchanged. From Kappa value of each query, we can statistically interpret to which extent users share consistent intent on the same query according to Table 3 (Landis and Koch, 1977). Though the interpretation in Table 3 is not accepted with no doubt, it can give us an intuition about what extent of agreement consistency is. In other words, Kappa is a measure with statistical sense. Meanwhile, Kappa values of queries with

[2] There may be more than two users who submitted the same query.



(a). AOL **(b). Sogou**

Figure 1: Number of unique queries and query submissions as a function of Kappa value.

κ	Interpretation
< 0	No agreement
0.0 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Table 3: Kappa Interpretation

various sizes of click sets are also comparable. That is also the reason we choose Kappa to measure consistency.

4.2 Distribution of Kappa

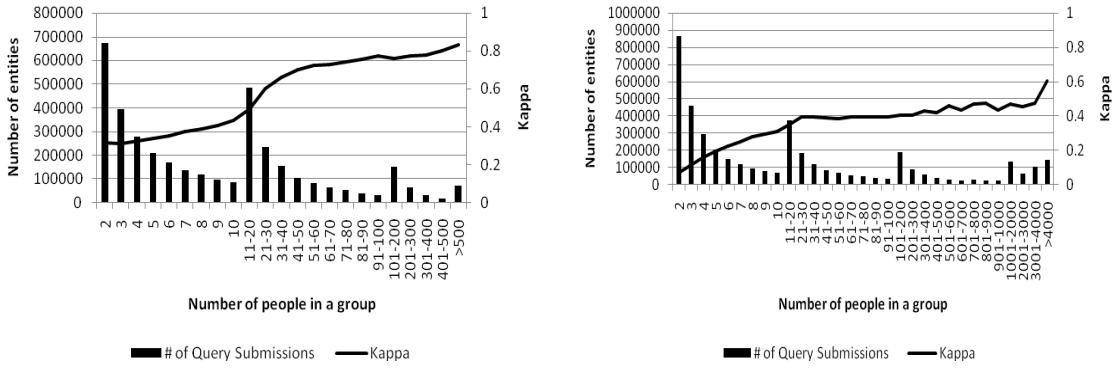
As introduced in Section 2, common user information is supposed to be the repetition of users' behaviors. We consider that the amount of repetition of users' clicks on one query is quantified by the consistency of its clicks. To statistically present the scale of repetition in current query log, we try to give an overview of consistency level of two commercial query logs.

Figure 1 plots distribution of Kappa value of the two logs in the coordinate with logarithmic Y-axis. About 34.5% unique queries (44.0% query submissions) in AOL log and only 13.9% unique queries (15.2% query submissions) in Sogou log have high Kappa values above 0.6. According to Table 3, click sets of these queries can be regarded as somewhat consistent. These queries can be roughly resolved by using common user information. On the other hand, for the rest of queries which constitute majority of the logs, users' click sets are rather diversified, which are hard to be satisfied by returning the same result list to them. As a whole, the queries in both AOL and Sogou can be characterized as less consistently in the clicks according to Kappa value, which is a statistical support for exploiting individual user information.

5 Individual or Common user information: A Tendency View

The above analyses quantitative analyses have shown that the repetition of search is not the statistically dominant factor, with the impression that employing individual user information is equally, if not more, important than common user information. This section tries to further reveal this issue so as to balance the position of individual user information and common user information from a research point.

Intuitively, a query can be characterized by the number of people issuing it, i.e. query frequency if we remove the resubmissions of one query by the same people. We try to depict the above mentioned query submissions and Kappa values as a function of number of people who issue the queries in Figure 2. In Figure 2, different numbers of users who issue the same query are shown on the x-axis, and the y-axis represents the number of different entities (left scale) and the average Kappa value (right scale) of the queries. We find that the number of queries becomes very small when the number of users in a group grows over 10, so we set a variant step length for them: with the length step of the group size falling between 2 and 10 set as 1, between 11 and 100 as 10, between 101 and 1000 as 100 and above 1000 as 1000.



(a). AOL **(b). Sogou**

Figure 2: Average Kappa value of queries as a function of number of people in a group who issue the same query (line) and the number of submissions of the queries issued by the same size of group (dark columns).

According to Figure 2(a), Kappa values of the queries in AOL log with more than 20 users are above 0.6, which indicates rather consistent clicks for them, accounting for about 29.4% of all query submissions. While for those queries visited by less than 20 users, the Kappa value declines gradually from 0.6 with the drop of users. For these queries occupying majority of query submissions, exploiting individual user information is supposed to be a better solution since the clicks on them are rather individualized.

According to Figure 2(b), though Kappa values of queries increase similarly with people submitting them in AOL, the overall consistency of the queries in Sogou log is much lower: with a Kappa value below 0.6 even for the queries visited by a large number of users. This fact indicates that Chinese users may be less consistent in their search intents, or partially reflects that the Chinese as a non-inflection language has more ambiguity, which can also be implied from Table 2. Therefore, individual user information may be more effective than common user information in Sogou log.

Summarized from Figure 2, it is sensible that common user information is appropriate for the queries in the right-most of X-axis. With most number of visiting people, such queries bear rather consistent clicks though covering only a small proportion of the distinct query set. Moving from the right to the left, we can find the majority of queries yield a less Kappa value, for which the individualized clicks require individual user information to meet the needs of each user. In this sense, how to exploit individual user information is predestined as the next issue of information retrieval if common user information was to be well utilized.

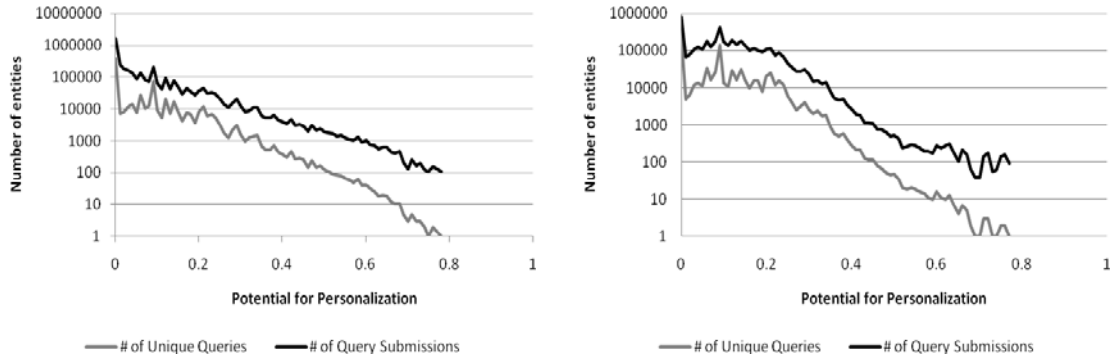
6 Queries for Personalization

Since using individual user information is a non-negligible issue in IR research, a subsequent issue is what queries can benefit in what extent from individual user information. In this section, we try to give an overview for this issue via a measure named potential for personalization.

6.1 Potential for Personalization

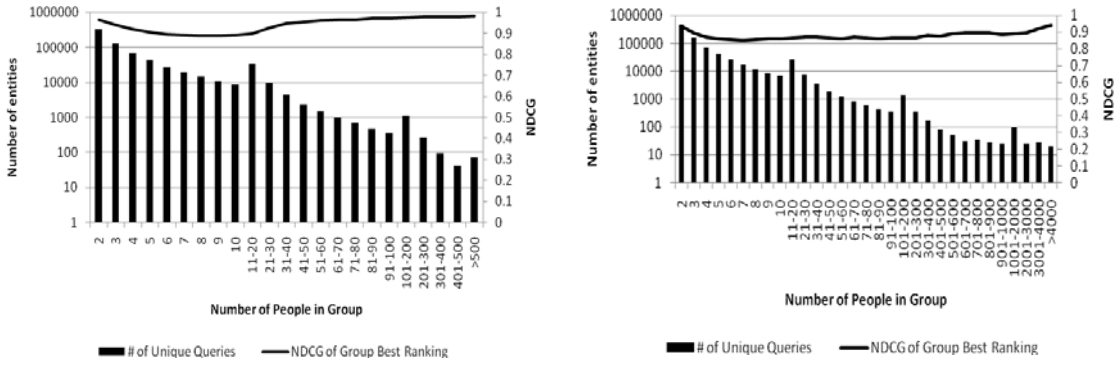
Potential for personalization proposed by Teevan et al. (2007) is used to measure the normalized Discounted Cumulative Gain (NDCG) improvement between the best ranking of the results to a group and individuals. NDCG is a well-known measure of the quality of a search result (Järvelin and Kekäläinen, 2000).

The best ranking of the results to a group is the ranking with highest NDCG based on relevance judgments of the users in the group. For the queries with explicit judgments, the best ranking can be generated as follows: results that all raters thought were relevant are ranked first, followed by those that most people thought were relevant but a few people thought were irrelevant, until the results most people though were irrelevant. In other word,



(a) AOL **(b) Sogou**

Figure 3: Number of unique queries and query submissions as a function of potential for personalization



(a) AOL **(b) Sogou**

Figure 4: The average NDCG of group best ranking as a function of number of people in group (solid line), combining with the distribution of the number of unique queries issued by the same size of group (dark columns)

the best ranking always tries to put the results that have the highest collective gain first to get the highest NDCG.

The previous work has shown that the implicit click-based potential for personalization is strongly related to variation in explicit judgments (J. Teevan et al., 2008). In this paper, we continue using click-based potential for personalization to measure the variation. Assuming the clicked results as relevant, we can calculate the potential for personalization of each query over the web search query log to present what kind of query can benefit more from personalization.

6.2 Potential for Personalization Distribution over Query Logs

Teevan et al. (2007) have depicted a potential for personalization curve based on explicit judgment to characterize the benefit that could be obtained by personalizing search results for each user. We continue using potential for personalization based on click-through to roughly reveal what kind of query can benefit more from personalization.

First we investigate the number of unique queries with different potential for personalization, which is shown in Figure 3. We find that there are about 53.9% unique queries in AOL log and 32.4% unique queries in Sogou log, whose potential for personalization is 0. For these queries, current web search is able to return perfect results to all users. However, for the rest of queries, even the best group ranking of results can't satisfy everyone who issues the query. So these queries should be better served by individual user information, covering

46.1% unique queries in AOL and 67.6% in Sogou.

Then, in order to further interpret what kind of query individual user information is needed most, we further relate potential for personalization to the number of users who submit the queries over AOL and Sogou query log as shown in Figure 4. For clarity's sake, we also set the same step length as in Figure 2.

According to Figure 4, the curve of potential for personalization is approximately U-shaped in both AOL log and Sogou Log. As the number of users in one group increases, performance of the best non-personalized rankings first declines, then flattens out and finally promotes[3]. Note that the left part of the curve is very similar to what Teevan et al. (2007) showed in their work.

Again in Figure 4, the queries which have the most potential for personalization are the ones which are issued by more than 6 and less than 20 users in AOL log. While in Sogou log, the queries issued by more than 6 and less than 4000 users have the most potential for personalization. Such different findings are probably caused by the content of query. There are many recommended queries in the homepage of Sogou search engine, most of which are informational query and clicked by a large number of users. Even when the size of group who issue the same query becomes very big, the query still has a wide variation of users' behaviors. So the consistency level of queries in Sogou log is much lower than the queries in AOL log at the same size of group.

7 Conclusion and Future Work

In this paper, we try to justify the position of individual user information comparing with common user information. It is shown that exploiting individual user information is a non-trivial issue challenging the IR community through the analysis of both English and Chinese large scale search logs.

We first classify the repetitive queries into 2 categories according to whether the corresponding clicks are unique among different users. We find that quantitatively the queries and clicks deserving for individual user information is much bigger than those deserving for common user information.

After that we use Kappa statistic to present that the overall consistency of query clicks recoded in search logs is pretty low, which statistically reveals that the repetition is not the dominant factor and individual user information is more desired to enhance most queries in current query log.

We also explore the distribution of Kappa values over different numbers of users in the group who issue the same query, concluding that how to utilize individual user information to improve the performance of web search engine is the next research issue confronted by the IR community when the repeated search of users are properly exploited.

Finally, potential for personalization is calculated over the two query logs to present an overview of what kind of queries that the optimal group-based retrieval model fails, which is supposed to benefit most from individual user information.

One possible enrichment to this work may come from the employment of content analysis based on text processing techniques. The different clicks, which are the basis of our examination, may have similar or even exact content in their web pages. Though the manual check for a small scale sampling from the Sogou log yields less than 1% probability for such case, the content based examination will be definitely more convincing than simple click counts. In addition, the queries for the two types of user information are not examined for their contents or the related information needs. Content analysis or linguistic view to these queries would be more informative. Both of these issues are to be addressed in our future work.

Acknowledgement

This work is supported by the Key Project of Natural Science Foundation of China (Grant No.60736044), and National 863 Project (Grant No.2006AA010108). The authors are grateful for the anonymous reviewers for their valuable comments.

[3] Note that the different step length dims the actual U-shape in the figure.

References

Canny John. 2002. Collaborative filtering with privacy via factor analysis. In *Proceedings of SIGIR' 02*, pages 45-57.

Carroll M. John and Mary B. Rosson. 1987. Paradox of the active user. *Interfacing thought: cognitive aspect of human-computer interaction*, pages 80-111.

Chirita A. Paul, Wofgang Nejdl, Raluca Paiu, and Christian Kohlschutter. 2005. Using odp metadata to personalize search. In *Proceedings of SIGIR '05*, pages 178-185.

Cohen Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20: 37-46

Dou Zhicheng, Ruihua Song, and Ju-Rong Wen. 2007. A Large-scale Evaluation and Analysis of Personalized Search Strategies. In *Proceedings of WWW '07*, pages 581-590.

Fleiss L. Joseph. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378-382.

Herlocker L. Jonathan, Joseph A. Konstan, Al Borchers, and John Riedl. 1999. An algorithmic framework for performing collaborative filtering. In *Proceedings of SIGIR '99*, pages 230-237.

Jansen J. Bernard, Amanda Spink, and Tefko Saracevic. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, pages 207-227.

Järvelin Kalervo and Jaana Kekäläinen. 2000. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of SIGIR '00*, pages 41-48.

Kohrs Arnd and Bernard Merialdo. 1999. Clustering for collaborative filtering applications. In *Proceedings of CIMCA '99*, pages 199-204.

Landis J. Richard and Gary. G. Koch. 1977. The mea-surement of observer agreement for categorical data. *Biometrics* 33: 159-174.

Pitkow James, Hinrich Schutze, Todd Cass, Rob Cooley, Don Turnbull, Andy Edmonds, Eytan Adar and Thomas Breuel. 2002. Personalized search. *ACM*, 45(9):50-55.

Shen Xuehua, Bin Tan and ChengXiang Zhai. 2005 Implicit user modeling for personalized search. In *Proceedings of CIKM '05*, pages 824-831.

Silverstein Craig, Monika Henzinger, Hannes Marais and Michael Moricz. 1999. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6-12.

Smyth Barry. 2007. A Community-Based Approach to Personalizing Web Search. *IEEE Computer*, 40(8): 42-50.

Speretta Mirco and Susan Gauch. Personalized Search based on user search histories. 2005. In *Proceedings of WI '05*, pages 622-628.

Spink Amanda, Dietmar Wolfram, Major Jansen, Tefko Saracevic. 2001. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226-234

Sugiyama Kazunari, Kenji Hatano, and Masatoshi Yoshikawa. 2004. Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of WWW '04*, pages 675-684.

Sun Jian-Tao, Hua-Jun Zeng, Huan Liu, Yuchang Lu and Zheng Chen. 2005. CubeSVD: a novel approach to personalized web search. In *Proceedings of WWW'05*, pages 382-390.

Teevan Jaime, Susan T. Dumais, and Eric Horvitz. 2005. Personalizing search via automated analysis of interests and activities. In *Proceedings of SIGIR '05*, pages 449-456.

Teevan Jaime, Susan T. Dumais and Eric Horvitz. 2007. Characterizing the value of personalizing search. In *Proceedings of SIGIR '07*, pages 757-758.

Teevan Jaime, Susan T. Dumais and Daniel J. Liebling. 2008. To personalize or Not to Personalize: Modeling Queries with Variation in User Intent. In *Proceedings of SIGIR '08*, pages 163-170.

Townsend Steve Cronen and W. Bruce Croft. 2002. Quantifying query ambiguity. In *Proceedings of HLT '02*, pages 613-622.

Yu Kai, Anton Schwaighofer, Volker Tresp, Xiaowei Xu, Hans-Peter Kriegel. 2004. Probabilistic Memory-based Collaborative Filtering. In *IEEE Transactions on Knowledge and Data Engineering*, pages 56-59.