International Journal of Innovative Computing, Information and Control
Volume **12**, Number **5**, October **2016**

ICIC International ©2016 ISSN 1349-4198
pp. 1745–1760

EXAMINING PERSONALIZATION HEURISTICS BY TOPICAL ANALYSIS OF QUERY LOG

Wei Song, Ying Liu, Lizhen Liu and Hanshi Wang

Information and Engineering College
Capital Normal University
No. 105, West 3rd Ring North Road, Beijing 100048, P. R. China
{ wsong; liuying; hswang }@cnu.edu.cn; xxgccnu@126.com

Received April 2016; revised August 2016

Abstract. *Personalization has been considered to be a promising way for future web search. Many algorithms have been proposed based on different heuristics. There is little work on systematically examining such heuristics. In this article, we examine typical personalization heuristics based on a commercial search engine query log. We aim to provide evidence to show the necessity, potential and limitations for personalization. Specially, we utilize automated query topic classification to simulate query intents and user preferences, which ensures that our study can be conducted on large scale query log data. We examine the heuristics on query ambiguity, query intent, user preferences and their interactions. We find that only a few queries with multiple user intents are inherently ambiguous. Queries with a specific meaning also have various user intents on different query aspects. About 48% of such queries involve named entities. We introduce new metrics to measure user preferences and preference distribution. These metrics show that searchers do have their preferences and focus on only a few interested topics. We assess the potential and difficulty for personalization based on search history. The results show that users with long search history benefit more from personalization, while the difficulty is nearly the same for users with either long or short search history. These observations could provide references for potential directions of web personalization.*

Keywords: Personalized search, Query log, Query classification, User modeling

1. **Introduction.** In the last decades, the World Wide Web has been changing dramatically. More and more data is available on the web. Search engines have become one of the most important tools for finding information from the web. The user enters a query, which contains typical 2 or 3 keywords, and the search engine returns a list of documents that are relevant to the query. However, in the response to a single query, the search engine may return hundreds of thousands of search results, covering various topics. It is impossible for users to explore all these results. Therefore, it is challenging for search engines to provide the most relevant documents to satisfy every user.

Personalized search has attracted much attention in Information Retrieval (IR) community which aims to provide information to users based on the contextual information of individuals [1]. There has been much work on personalized search. Many of proposed approaches rely on some heuristics such as "queries are inherently ambiguous" and "users have consistent preferences". These heuristics are used as basis for designing personalization algorithms. However, little work has been done to study whether these heuristics are true and to what extent, these heuristics help personalization.

In this article, we summarize personalization heuristics and systematically examine to what extent these heuristics are reliable. Our work relies on the query log of a commercial search engine. As a result of user interactions with search engines, query logs contain rich user search behaviors. Therefore, analyzing query logs is a promising way to understand user interactions with search engines and know about the world. Although much work has been proposed on inspecting the query length, query frequency and topic trends, most of existing work focuses on studying the behaviors of mass population, while little work deals with the behaviors and interests of individual users.

To fully take the advantage of large scale of the query log data, we adopt automated query topic classification to represent query intents and user preferences. A query topic classifier is trained based on the query log data as well with minimum human labors by considering the categories of the corresponding URLs that some users had clicked on. Based on the classifier, the preference of a user could be represented according to the categories of all his past clicks; the query intent can be simulated according to the clicks of all users who have submitted the query. Therefore, it is possible for us to observe the variations of query intents of different users and how user interests evolve along with time. We can also assess the potential and difficulty for personalized search based on the above observations.

We examine personalization heuristics related to query ambiguity, query intent and user preferences on large scale query log data and find some interesting observations. (1) Queries tend to have multiple user intents. However, users focus on narrow topics. (2) Among the queries with many topics, there are only a few of them having more than one meaning. However, for queries with a clear meaning, such as named entity queries, users have various intents as well. The reason is that these queries have multiple aspects or subtopics. (3) Users do have preferences and their preference distribution is highly unbalanced. Users usually dwell on a few interested topics. (4) Users with longer search history available have larger personalization potential while the difficulty is nearly the same for all users. These observations complement previous research.

The remaining parts of this article are organized as follows. We begin the article with a brief introduction of related work. In this section, we state and compare some extant works. In Section 3, we present a summarization of typical personalization heuristics. In Section 4, we present our dataset and the method we used to classify user clicked URLs and user queries. In Section 5, we analyze 2 main types of personalization heuristics. Next, we conclude with a discussion about our observations and future work.

2. Related Work.

Our work is highly related to several research areas: personalized search, query log analysis and query ambiguity.

2.1. Personalized search.

Personalized search assumes that contextual user information could be used for distinguishing different information needs among users. Many methods have been proposed and most of them rely on some personalization heuristics. Matthijs and Radlinski [2] adopted user browsing history for personalization. Their heuristic is that user long term preferences are consistent. Bennett et al. [3] gave the first study to assess how short-term (session) behavior and long-term (historic) behavior interact, and how each may be used in isolation or in combination to optimally contribute to gains in relevance through search personalization. White et al. [4] mined historic search-engine logs to find other users performing similar tasks to the current user and leveraged their on-task behavior to identify Web pages. Yan et al. [5] proposed a characterization and evaluation of the use of cohort modeling to enhance search personalization to access to sufficient information about each user's interests and intentions. In [6] a user profile was generated using a domain ontology. They expanded personalized query to reflect user's interests and intents. Aiming to improve the accuracy of personalized search, Wang et al. [7] proposed a preference recommendation scheme to complement users' conditional preference networks. Their recommender system can be applied in personalized search on the premise that user preferences are consistent. To provide accurate preferences of users for effectively personalized search, Xu et al. [8] captured and merged the independent user context in a user session after short-term query context is captured. They think short-term context is more suitable for personalized search. However, these heuristics are considered to be true and little work has been done to examine such heuristics. We analyzed the user preference related heuristics in this paper and measured the potential and difficulty for personalization using search history.

2.2. **Query log analysis.** In the early years, query log analysis focused on the basic characteristics such as query frequency and session related analysis [9,10]. Later, more work had been done for exploring the long term effects of query logs [11,12] or topical analysis of query logs. Wedig and Madani [13] analyzed a large scale query logs by accessing the opportunity of personalized search. Our work is inspired by their work. However, they mainly focused on user preference related information but ignored the query related information. Besides, they classified a query according to its content. It means a query has only one category no matter who submits it or when it is submitted. Instead, we link the query topic to user's real actions and assign topics to queries according to user intents. Teevan et al. [14] measured the potential for personalization based on both explicit and implicit judgements. They found people's judgements on the same queries differ greatly and quantified the potential by measuring the gap between the optimal rating for an individual and the optimal rating for a group of users. Our work extends their work from a different view. We analyze the potential and difficulty for personalization making use of user search history. Eickhoff et al. [15] presented an in-depth analysis of sessions and investigated within-session and cross-session developments of expertise, focusing on how the language and search behavior of a user on a topic evolves over time. Pu and Jiang [16] made an investigation of the academic information finding and re-finding behavior through various methods. Their observations provide opportunities for personalization by ranking pages that users have viewed before higher. Part of our work is similar to their work. For example, we examine the topic distributions of queries that have been issued by the same users multiple times. We found that the topic of re-finding queries is narrow which is consistent with their observations but from a different angle. Whiting et al. [17] presented Chinese language Sogou 2012 query log and released it. However, their work is not on personalization. Potey et al. [18] reviewed and compared some of the available methods to give an insight into the area of query log processing for information retrieval. They classified web query intent based on knowledge extraction from query log analysis. Jiang and Yang [19] captured comprehensive information in search query log, and proposed three frameworks to infer query intents. We do not aim to make infer in this work, but look forward to finding evidences supporting inferring.

2.3. **Query ambiguity.** A strong motivation is that queries are inherently ambiguous. As a result, some work has been done to distinguish ambiguous queries. Song et al. [20] proposed a taxonomy for query ambiguity and designed learning based algorithm to distinguish ambiguous queries from broad queries and clear queries. Zhu and Wei [21] introduced a method to measure the ambiguity of user queries to help dynamically adjusting the number of expanded terms. Luo et al. [22] proposed a query representation approach named "query2vec", and they built a query ambiguity identification framework taking user behavior features collected from click-through logs into consideration to tell the differences between clear and ambiguous queries. Kamal et al. proposed a hybrid query disambiguation adaptive approach in 2015 [23] and then they proposed a Post-search Ambiguous query Classification Method in 2016 [24], to make ambiguous queries clear and address the inherent ambiguous queries to improve the accuracy of search results. Our work considered query ambiguity from two aspects: user intents and query content. We first analyzed query intents according to user click information, and then sampled queries with multiple intents for analysis.

3. Personalization Heuristics.

Heuristic refers to experience-based techniques for problem solving, learning, and discovery (defined in Wikipedia). Extensive existing personalization approaches rely on some kinds of heuristics. These heuristics lay the foundation for web personalization.

We first distinguish three terminologies that may result in confusion. ***Query ambiguity*** reflects whether a query has multiple meanings, considering the query properties only. In contrast, ***query intent*** indicates the potential information need of users on a specific query. Moreover, ***user preference***s represent the preferred interests or topics of individual users. These concepts actually talk about different perspectives in search personalization.

The motivation of personalized search is highly related to these perspectives. Therefore, we roughly categorize personalization heuristics according to these perspectives. However, in addition to studying them separately, we are also interested in the interactions between them.

In this section, we would summarize some heuristics that are commonly used in previous work. In next section, we will show how to represent query ambiguity, query intent and user preference based on automated query topic classification. Then we will examine these heuristics in Section 5 in order to gain more insights.

3.1. Query ambiguity heuristic.

Heuristic 1. *Queries are inherently ambiguous.*

Query ambiguity is a strong motivation for personalized web search. Classical examples like "apple" (referring to a kind of fruit or a company) and "eclipse" (literally referring to an astronomical natural phenomenon, or referring to a song, a film, even a software development) are extensively used in many works. However, ambiguity is a fuzzy concept. Song et al. [20] constructed a taxonomy for query ambiguity consisting of 3 types:

- Ambiguous query: a query that has more than one meaning;
- Broad query: a query that covers a variety of subtopics;
- Clear query: a query that has a specific meaning and covers a narrow topic.

It is a big step towards identifying ambiguous queries. However, there are two limitations of the work. (1) The experiments were conducted on small sampled queries data set and relied on manual examination of query ambiguity. (2) Although this taxonomy reflects the characteristics of the query ambiguity, they do not analyze the relation between query ambiguity and user intent.

We investigate the above questions based on the query log which contains large scale of users' behaviors. Moreover, we measure the query ambiguity and user intent based on automated query topic classification. This allows us to examine personalization heuristics on the large scale query log data.

We look forward to answering the following research questions:

(1) Are queries ambiguous inherently?
(2) How many queries are inherently ambiguous?

3.2. Query intent heuristic.

Heuristic 2. *For the same query, there are diverse query intent (different users have different goals).*

Heuristic 2 is the fundamental assumption of personalized search. Submitting query "iPhone", different users may want to search for diverse information. Their intents may be the price of this phone, the phone's capabilities or the pictures of the phone. Many work on query intents or subtopic mining task [6,25,26] are all based on this heuristic. We will examine the heuristic and look forward to answering the following questions:

(3) How many topics does a query have from the users' perspective?

(4) What types of queries tend to have multiple user intents?

3.3. User preference heuristics.

Heuristic 3. *Users have distinct preferences.*

Another fundamental heuristic for developing personalized algorithm is that users have distinct preferences. Therefore, user future interests could be predicted based on past preferences. User preferences could be divided into two types: short-term preference and long-term preference.

Heuristic 4. *Sequential queries tend to have similar topics.*

This heuristic is related to short-term preferences. Short-term preference refers to user interests within a short period, which throws light on a user's current information need in a single session. A session can be considered as a period consisting of all interactions for the same information need. In response to a single query, the search engine may return hundreds of thousands of search results. Many of these results may be irrelevant to the user's real intent because of the inaccuracy and ambiguity of the query. To access to accurate information, the user usually adjusts his query content by changing or adding some specific keywords. So these sequential queries issued by the user may have similar topics. A network hot topic can be a short-term information need. In [8,27], the searchers all think short-term context is more suitable for personalized search.

Heuristic 5. *User preferences are consistent.*

This heuristic is about user long-term preferences. Long-term preference refers to user interests over a long period of time. Generally, long-term preference refers to information that is stable for a long time and is often accumulated over time, such as a user's education level and general interest. Long-term preference can be applicable to all sessions. The application in [2] and [7] are both on the premise that user preferences are consistent.

Short-term preference is more random compared to long-term preference which is usually considered as consistent. And we summarize the heuristics and by examining the heuristics, we hope to answer the following research questions:

(5) How probably the topic of current query is the same as previous ones within a short period of time?

(6) How does user topical interest distribute?

(7) How much will user search history help personalization?

(8) How much is the personalization difficulty for users with different search history?

4. Dataset and Topical Categorization.

4.1. Dataset.

Our investigation is based on one month's Chinese query log which is released by Sogou search engine[1]. The dataset includes 28 days' query logs that were sampled in June, 2008. The elements we used in this study include: a user ID, the day and time of the query submission, the query string and corresponding user clicked URL. In total, there are 51537393 query instances, 5655036 distinct queries and 10679396 users.

[1] http://www.sogou.com/

4.2. **Topical classification of URLs and queries.** We want to examine user intents about queries and user interests. Since it is difficult to describe query intents precisely, we use a taxonomy of categories to model query intents and user interests. There are various ways to classify a query (Beitzel et al. [28]; Shen et al. [29]). Fixing a query's topic for any user at any time is not a good choice, because both query topics and user interests vary in different contexts. So the topic of a query should depend on the user actions at a particular time. We assign the topic of the corresponding user clicked URL to a query. We adopt the method used in [30] by directly looking up in a categorized directory of web. In their study, they used the top level categories from Open Directory Projects (ODP) consisting of 15 topics.

The 15 categories may be too broad to differentiate different user intents. Therefore, we use a taxonomy with moderate categories based on KDD cup 2005 taxonomy [31]. The taxonomy, which consists of two level hierarchies, has 67 categories in total. The top level has 7 categories. We used the second level of categories directly. We collected categorized websites from ODP for Chinese[2], YahooCN[3] and Baidu[4]. We then manually assigned the categories at second level of each of these Chinese directories into the taxonomy of KDD cup 2005. In this process, we merged some categories in KDD cup 2005 taxonomy, since these categories are considered to be the same in Chinese directories. Finally, we got a taxonomy with 59 categories and about 56000 distinct categorized websites.

For each URL, we determine its category by looking up in the used taxonomy. If there is no matched entry, prune the postfix of the URL and look up again. This process is repeated until a match is found or a miss is detected. We only retain the URLs that could be classified into only one category. Other URLs and corresponding query samples are discarded.

In total, we got about 23.1 million categorized queries and 7.4 million users. From Table 1, we can see that the simple automatic classification method can cover about 63.5% of distinct queries and 70.1% of users in the original query log.

Table 1. Statistics of the original and the categorized query log. The percentage of categorized query log to the original one is also shown.

	original (millions)	categorized (millions)	percentage
#query	51.5	23.1	44.8%
#distinct query	5.6	3.5	63.5%
#user	10.6	7.4	70.1%

4.3. **The performance of the query classifier.** Our query topical classifier depends on the URLs only and is fully automated. To confirm its effectiveness, we randomly sampled 1000 query and category pairs. We asked two labelers to check the query in each pair, and manually assigned the best category among the 59 categories. The agreement between two labelers is 0.84. We used the agreed 840 query category pairs as the test data and compared the automated predicted categories with the manually labeled categories. We used the p@N as a metric. The experimental results show than p@1 can reach 91% and p@2 can reach 96%. This indicates that the query classifier is reliable enough for our research.

[2] http://www.dmoz.org/World/Chinese_Simplified/

[3] http://site.yahoo.com.cn/

[4] http://site.baidu.com/

5. **Experiments.**

In this section, we provide the experimental results. From the experimental results, we find some observations and answer our interested research questions. First, we examine the query intents and ambiguity. As introduced before, we use topics to simulate query intents, since it is difficult to describe query intents accurately. Generally, we assume different topics lead to different intents. Then we model and analyze the user preferences by considering their topical interests.

5.1. **Models.**

The taxonomy used is represented as $T = \{t_1, \cdots, t_n\}$, where n is the number of categories. In our case, $n = 59$. Suppose a query q had been issued for l times. Each time, a query category is assigned based on the category of the clicked URL. The query could be represented with a sequence of categories $C = \{c_1, c_2, \cdots, c_l\}$.

Definition 5.1. *The query intent is represented as* $Intent(q) = (w_1, \cdots, w_n)$*, where* $w_i = \frac{count(t_i,C)}{l}$*,* $count(t_i, C)$ *represents the number of the clicked URLs for query* q *belonging to* t_i*.*

Similarly, for user u with search history $H = (q_1, \cdots, q_m)$, the corresponding category sequence is $C_u = (c_1, \cdots, c_m)$.

Definition 5.2. *The user interests are modeled as* $Preference(u) = (v_1, \cdots, v_n)$*, where* $v_i = \frac{count(t_i,C_u)}{|H|}$*,* $count(t_i, C_u)$ *represents the number of the URLs clicked by* u *belonging to* t_i*.*

5.2. **Query intent and ambiguity analysis.**

We start query intent and ambiguity analysis by presenting some basic results about query topics. For a distinct query, we aggregate its topics from all its occurrences in query log. That is to get the topic distribution among all users who had submitted this query. There are about 2.4% of queries having at least 10 topics. As summarized in Table 2, for queries with more than 10 instances, the average number of topics is 3.99, and 31.8% of such queries have at least 5 topics. 52.7% of queries with at least 50 instances have at least 5 topics. This figure increases to 59.7% for queries with at least 100 instances.

The results show that high frequency queries probably have more diversified intents. Since we only use topics to simulate query intents, in reality, it may be more diversified. The answer of the research question (3) can be extracted obviously from Table 2.

Table 2. Percentage of the number of topics for queries with a certain frequency

%	Query Frequency		
#Topic	≥ 10	≥ 50	≥ 100
1-4	68.2	47.3	40.3
5+	31.8	**52.7**	**59.7**
10+	2.5	10.2	15.1

It is interesting to see what types of queries tend to have multiple topics. We sampled 500 queries with at least 10 topics and explored the types of such queries. Figure 1 illustrates the basic characteristics of such queries. Surprisingly, in reality, we find only a few queries are inherently ambiguous that have more than one meaning or refer to more than one entity. Most of these queries have clear meaning but they still have multiple topics. We checked the topics that a clear query “辣妹 维多利亚 (spice girl Victoria)” covers as shown in Table 3. We can see that categories it covers reflect different aspects of the entertainment star. People may look for the information about some aspects of the same object by inputting simple queries to search engines.

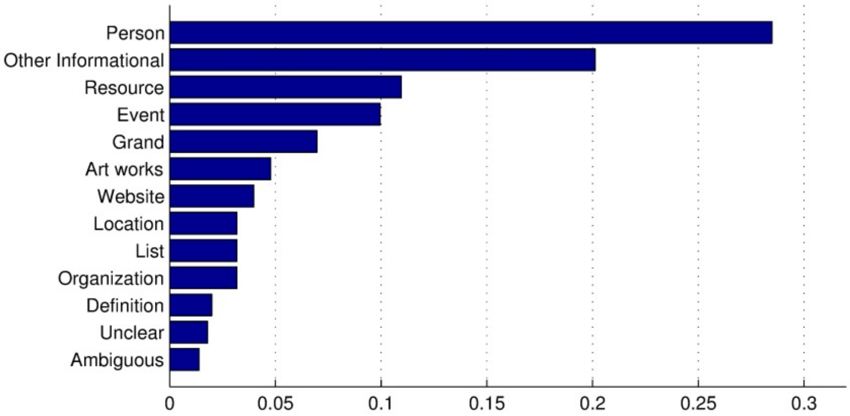

Figure 1. Types of sampled queries, each of which has at least 10 topics

Table 3. The category distribution of query "spice girl Victoria"

Category	Percentage(%)
Living.Health Fitness	34.5
Sports.News Scores	12.7
OnlineCommunity.Forums Groups	10.9
Entertainment.TV	7.3
Living.Book Magazine	7.3
Computers.Internet Intranet	7.3
Information.Arts Humanities	3.6
Living.Fashion Apparel	3.6
Living.Car Garage	3.6
Information.Companies Industries	1.8
Entertainment.Pictures Photos	1.8
Living.Finance Investment	1.8
Information.Education	1.8

Among these queries with multiple topics, name entity queries comprise a large proportion. About 42% of such queries are name entity queries themselves and 48% of queries contain at least one name entity as a substring. That is, name entity queries are prone to having multiple user intents in a very great degree.

Another question is: are there any dominating intents for queries? People may care about several aspects of a query, but some aspects are common. For a given query q, we construct $Intent(q)$ according to Section 5.1. Based on $Intent(q)$, we define $SortedIntent(q)$ as:

Definition 5.3. *$SortedIntent(q) = (w_{s_1}, \cdots, w_{s_n})$, where $w_{s_i} \geq w_{s_j}$ if $i < j$.*

$SortedIntent(q)$ ranks user interested topics for query q from the most interested one to the least interested one.

We represent w_{s_i} as $SortedIntent(q, i)$. Given a set of queries Q, we can define the accumulated version of $SortedIntent$:

Definition 5.4. *$AccSortedIntent(Q) = (a_1, \cdots, a_n)$, where $a_i = \frac{\sum_{q \in Q} SortedIntent(q,i)}{|Q|}$.*

$AccSortedIntent$ can reflect how much attention people paid on topics with different interest levels over all queries. a_i indicates how much attention people pay on the i-th most interested topic. For Figure 2, we can see for a query, most of people focus on 1 to

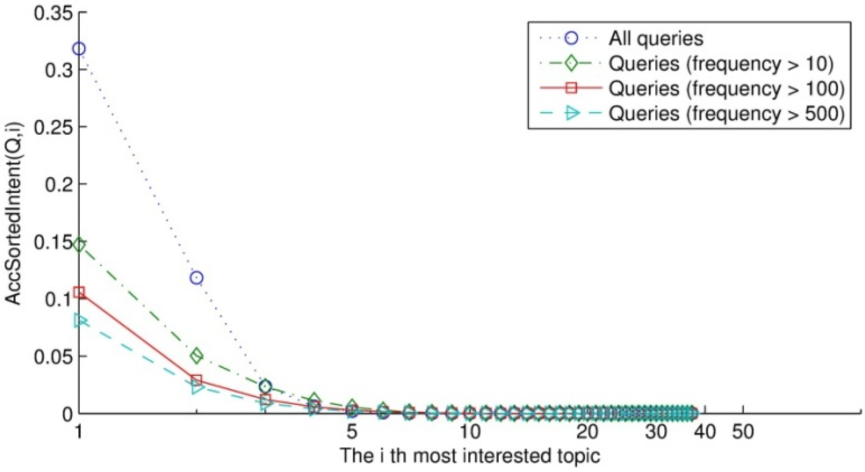

Figure 2. *AccSortedIntent* for queries with different frequency. The X-axis represents the sorted topics from most interested one to the least interested one. The Y-axis represents the *AccSortedIntent*.

3 aspects of the query. If the most interested topics for each query could be identified, search engines could return more documents on these topics. The result could satisfy more people's interests, although it is not tailored for individuals.

Based on the above analysis, we find that many queries have multiple user intents. Next, we examine the topic trends for the same query submitted by a single user and multiple users. We also show the difference of viewing the same query between a single user and more users. For this evaluation, we extracted sample categorized queries which satisfy the condition that a query must be submitted by at least 2 users and each of these users submitted this query at least 2 times.

From Figure 3 and Figure 4, we can see if people look for information about the same queries over time, the average number of topics is small. The overall average number of topics considering re-finding behavior is 1.68. This indicates if a user looks for information about a query repeatedly, he or she may focus on narrow aspects of the query. It is consistent with previous work [32], that user clicks tend to converge for the same query.

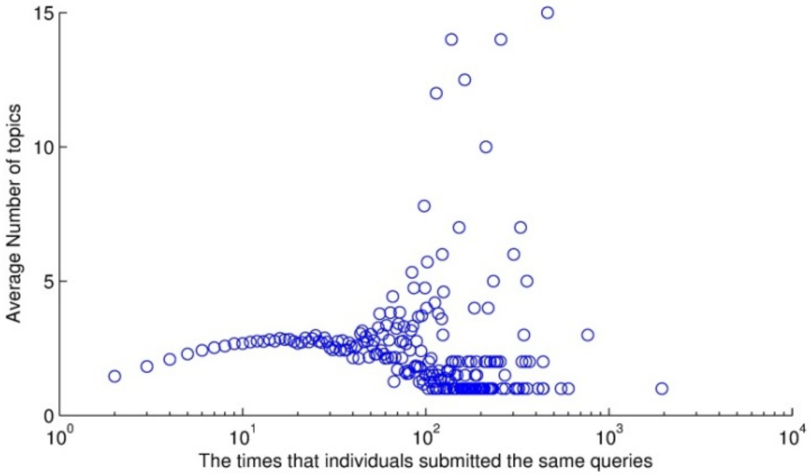

Figure 3. The average number of topics for queries that have been submitted by the same users for a certain times

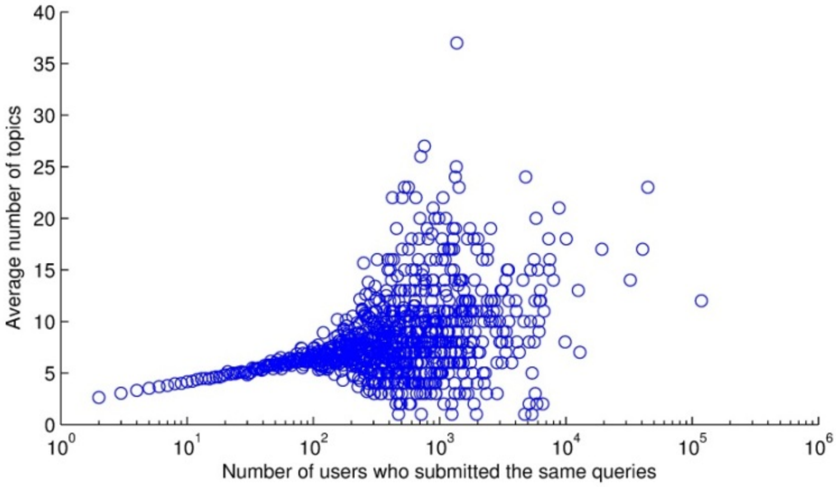

FIGURE 4. The average number of topics for queries that have been submitted by multiple users

In contrast, the overall average number of topics is 3.21 when we consider more users. The main trend increases with the number of persons submitting the same queries. It means for the same queries, user intents are diverse.

5.3. **User preference analysis.** In this section, we analyze user preferences and preference distribution. We also introduce new metrics for the analysis. First, we sort user interested topics and extract the accumulated version of the sort. By this way, we can conclude the distribution of preference. In addition, we check the user search history and examine the consistency of user interests based on user search history. To examine the consistency, we introduce new metrics to measure the potential and the difficulty for personalization.

5.3.1. *Basic analysis.* First, we show some experimental results about how many topical interests users usually have. Figure 5 illustrates the number of users that have different number of topical interests. The figure obtained by analyzing all users might not reflect the reality, since many users in query log issued only few queries. In contrast, for users

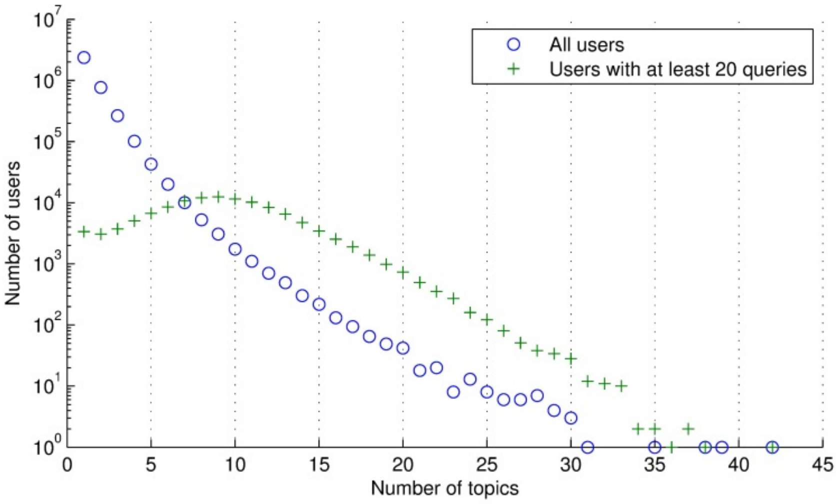

FIGURE 5. Number of users that have a certain number of topics

with longer search history (issued more than 20 queries in query log), 79.6% of users have 5 to 15 interested topics, and only 2% of these users have interested topics more than 20. This indicates user searches usually do not cover broad topics. Users have their preferences on relative narrow topics.

Next, we examine the query distribution over different user interested topics.

Definition 5.5. *$SortedPreference(u) = (v_{s_1}, \cdots, v_{s_n})$, where $v_{s_i} \geq v_{s_j}$ if $i < j$.*

SortedPreference sorts user interested topics from most interested one to least interested one. Given a set of users U, we define the accumulated version of *SortedPreference* as:

Definition 5.6. *$AccSortedPreference(U) = (a_1, \cdots, a_n)$, where $a_i = \frac{SortedPreference(u,i)}{|U|}$.*

Figure 6 presents the *AccSortedPreference* curves for users with different search history length. *AccSortedPreference* represents the percentage of queries that users submitted from the i-th most interested topic. We can see for Figure 6 that for users with search history longer than 20 queries, the curves become stable. On average, users pay much more attention on their most interested topics than less interested topics. It is to be recalled from Figure 5, most users have more than 5 topics. However, 44% of their queries are from the most interested topic. More than 73% of their queries belong to the top 3 most interested topics. It shows that users' searches focus on narrow topics. That means that user topical interest distribution is unbalanced and users usually dwell on a few interested topics. Since users focus on narrow topics, the possibility of that the topic of current query is the same as previous ones within a short period of time in great.

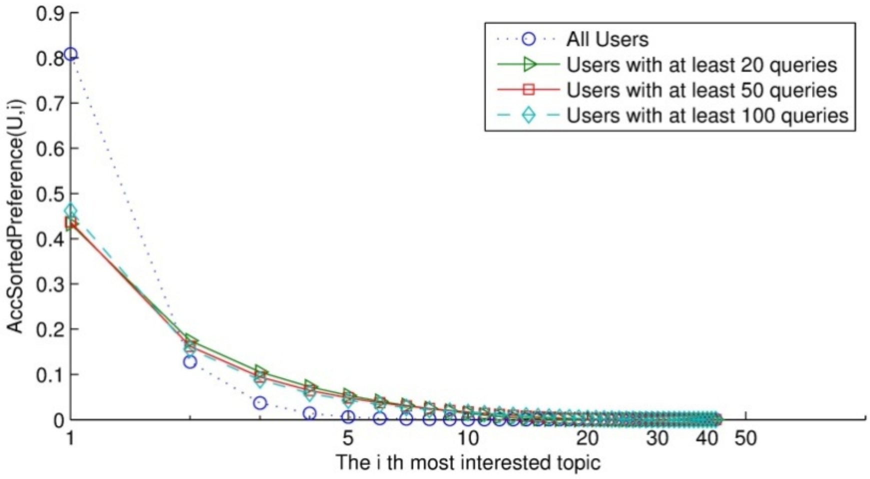

FIGURE 6. The curve of *AccPreference*(U) for users with different length of search history

We also check the user search history length distributions. Figure 7 shows the correlation between user search history and the number of users. In addition, Table 4 presents the concrete numbers of users that have certain search history length. The result shows most users issued queries less than 5%. Only 2% of users have history length more than 20.

5.3.2. *Consistency of user preferences.* Now we want to examine the consistency of user interests. Our examination will be based on the following two assumptions: topics within a session are more consistent than across sessions; user interests in long search history are consistent as well.

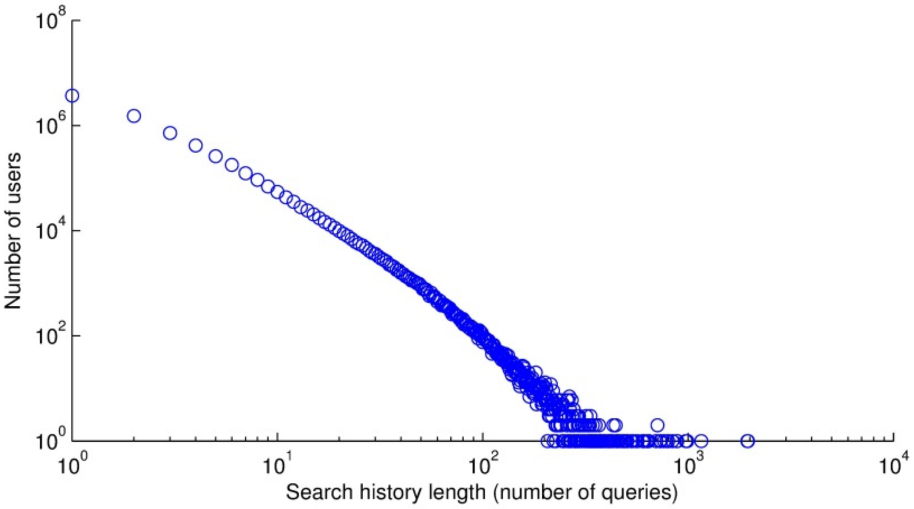

FIGURE 7. The correlation between user search length and the number of users

TABLE 4. Percentage of number of users with a different search history length

history length	[1-5]	(5-10]	(10-20]	(20-100]	[100,+)
#users	6643289	518811	217538	106418	3264
percentage(%)	88.7	6.9	2.9	1.4	0.44

We examine the transitions between consecutive queries within the same session. Our observations are similar with [30]: transitions from a topic to itself is most common. The average transition probability within sessions is much bigger than the average transition probability between two sessions.

Next, we examine the consistency of user history from the view point of measuring the potential and the difficulty for personalization based on user search history.

Suppose user u has search history $H_u = \{q_1, \cdots, q_m\}$, and the corresponding category sequence $C_u = \{c_1, \cdots, c_m\}$. We can divide the user history into two parts based on chronological order: H_u^A and H_u^B. The corresponding categories sequences are C_u^A and C_u^B. We consider the H_u^A as user search history, and H_u^B as the future user interests.

Definition 5.7. *For query $q \in H_u^B$, if the true topic of q is included in C_u^A, say q is predictable for u, noted as $Predictable(q) = \begin{cases} 1, & c(q) \in C_u^A \\ 0, & otherwise \end{cases}$, where $c(q)$ represents the category of query q.*

Predictable is used to indicate whether the topic of a future query could be predicted based on search history instead of using the dominating topics in search history. We assume that if the true topic is included in history topics, then it is predictable; otherwise, it is unpredictable.

Definition 5.8. *Define $AvgPredictable(Q)$ as the average Predictable score over a set of queries Q submitted by user U, noted as*

$$AvgPredictable(Q) = \frac{1}{\sum u \in U} \sum_{u \in U} \sum_{q \in H_u^B} Predictable(q).$$

Similarly, we define metrics to evaluate how much personalization could help single users and global users.

Definition 5.9. *For user u, define $UserPredictable$ as the percentage of queries in H_u^B that are predictable, noted a* $UserPredictable(u) = \frac{1}{|H_u^B|} \sum_{q \in H_u^B} Predictable(q)$.

$UserPredictable(u)$ represents the potential for personalization to a single user. It is easy to understand. Personalization may not always work. $UserPredictable(u)$ tells the upper bound proportion of cases that personalization works for an individual.

Definition 5.10. *For a group of users U, define*

$$AvgUserPredictable(U) = \frac{1}{|U|} \sum_{u \in U} UserPredictable(u).$$

$AvgUserPredictable(U)$ indicates the potential of personalization for large scale of users. Note that in reality, two queries having the same topic does not mean they present the same information need. Thus, the metrics introduced above provide an upper bound of the potential for personalization based on search history.

In contrast to the potential for personalization, we also introduce metric to measure the difficulty for personalization based on search history.

Definition 5.11. *For $q \in H_u^B$, all queries with different topics with q in H_u^A are considered as noise for predicting the topic of q. We define $HistoryNoise(q)$ as the percentage of noisy queries in H_u^A, noted as* $HistoryNoise(q) = \frac{|H_u^A| - |C_u^A(c(q))|}{|H_u^A|}$.

We measure the difficulty for personalization by measuring the noise in search history. The basic assumption is that when we predict the future user interests, only part of information from history plays a positive role, while the remaining search information, which is not related to the true topics, makes the personalization based on search history difficult.

Definition 5.12. *For a set of queries Q submitted by user U, define*

$$AvgHistoryNoise(Q) = \frac{1}{\sum u \in U} \sum_{u \in U} \sum_{q \in H_u^B} HistoryNoise(q).$$

We still base on the assumption that if the true topic is included in history topics, then it is predictable. So the queries with true topic in history are all considered to play a positive role for predicting the true interest. In reality, this may be not always true. As a result, $AvgHistoryNoise(q)$ stands for the lower bound of difficulty for personalization based on search history.

We simply divide users' whole search history equally into two parts. We can draw conclusions from Figure 8. The search history length plays an important role for personalization. Users with longer search history probably benefit more from personalized search. One reason is that users focus on narrow topics, and when search history is long enough, user interests tend to converge. As a result, future queries will be about the old topics with a high probability. While the difficulty for personalization for users with different length of search history is similar, ranging from 32% to 36%. The reason is similar. When search history is long enough, it covers broad topics. Therefore, for a particular interest, the remaining topics become noise. The result means that for any user, when we apply personalized search, at least 32% of information is noisy for current query. The key point, for personalization based on search history, is to filter the relevant part in history for current information need.

In this section, we summarize and experimentally examine some personalization heuristics. Taking full advantage of query logs, we adopt automated query topic classification to represent query intents and user preferences. From the experimental results, we solve

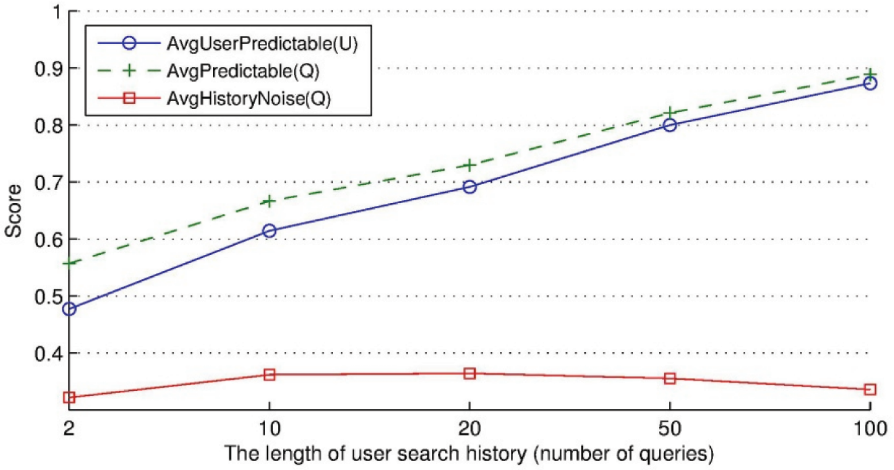

Figure 8. The curves for users $AvgUserPredictable$, $AvgPredictable$ and $AvgHistoryNoise$ with different search history length

some research questions and find some interesting observations. Queries tend to have multiple user intents. Among these queries, only a few queries are inherently ambiguous, and users focus on narrow topics. 48% of these queries involve named entities. Users do have preferences and their preference distribution is unbalanced. Along with the analysis of users search history, we find the topic of current query is the same as previous ones within a short period of time with a high probability. Users with longer search history available have larger personalization potential while the difficulty is nearly the same for all users.

6. **Conclusions.** We have explored typical heuristics assumptions based on which different personalization algorithms have been proposed. We mainly evaluated 2 types of personalization heuristics: query ambiguity and user preference. The dataset we used is a large scale query log with millions of queries and users. We modeled query topics and user interests by topical categorization of the web pages the users had clicked corresponding to the queries. In this way, the query topic is connected to the user behavior instead of using query content.

This article has summarized typical personalization heuristics and attempts to provide evidence to support these heuristics. Looking forward, we plan to develop methods for automatically mining query intents or subtopics, and employ subtopics for identifying user information need. Also, subtopics may help us to diversify search results. It is also necessary to develop algorithms for filtering relevant information from search history for current information need, especially for users with long search history.

Acknowledgment. This work was supported in part by National Science Foundation of China under Grant Nos. 61303105 and 61402304; the Beijing Natural Science Foundation under Grant No. 4154065; the Beijing Educational Committee Science and Technology Development Plan under Grant No. KM201610028015; the Humanity & Social Science General Project of Ministry of Education under Grant No. 14YJAZH046 and Beijing Key Disciplines of Computer Application Technology.

REFERENCES

[1] J. Pitkow, Personalized search: A content computer approach may prove a breakthrough in personalized search efficiency, *Communications of the ACM*, vol.45, 2002.

[2] N. Matthijs and F. Radlinski, Personalizing web search using long term browsing history, *Proc. of the 4th ACM International Conference on Web Search and Data Mining*, pp.25-34, 2011.

[3] P. N. Bennett, R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk et al., Modeling the impact of short- and long-term behavior on search personalization, *Proc. of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.185-194, 2012.

[4] R. W. White, W. Chu, A. Hassan, X. He, Y. Song and H. Wang, Enhancing personalized search by mining and modeling task behavior, *Proc. of the 22nd International Conference on World Wide Web*, pp.1411-1420, 2013.

[5] J. Yan, W. Chu and R. W. White, Cohort modeling for enhanced personalized search, *Proc. of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp.505-514, 2014.

[6] G. J. Hahm, M. Y. Yi, J. H. Lee and H. W. Suh, A personalized query expansion approach for engineering document retrieval, *Advanced Engineering Informatics*, vol.28, pp.344-359, 2014.

[7] H. Wang, S. Shao, X. Zhou, C. Wan and A. Bouguettaya, Preference recommendation for personalized search, *Knowledge-Based Systems*, vol.100, pp.124-136, 2016.

[8] Z. Xu, H.-Y. Chen and J. Yu, Generating personalized web search using semantic context, *The Scientific World Journal*, vol.2015, 2015.

[9] C. Silverstein, H. Marais, M. Henzinger and M. Moricz, Analysis of a very large web search engine query log, *ACM SIGIR Forum*, pp.6-12, 1999.

[10] B. J. Jansen, A. Spink and T. Saracevic, Real life, real users, and real needs: A study and analysis of user queries on the web, *Information Processing & Management*, vol.36, pp.207-227, 2000.

[11] M. Richardson, Learning about the world through long-term query logs, *ACM Trans. the Web*, vol.2, p.21, 2008.

[12] N. Buzikashvili, Query topic classification and sociology of web query logs, *Computación y Sistemas*, vol.19, pp.633-646, 2015.

[13] S. Wedig and O. Madani, A large-scale analysis of query logs for assessing personalization opportunities, *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.742-747, 2006.

[14] J. Teevan, S. T. Dumais and E. Horvitz, Potential for personalization, *ACM Trans. Computer-Human Interaction*, vol.17, p.4, 2010.

[15] C. Eickhoff, J. Teevan, R. White and S. Dumais, Lessons from the journey: A query log analysis of within-session learning, *Proc. of the 7th ACM International Conference on Web Search and Data Mining*, pp.223-232, 2014.

[16] H.-T. Pu and X.-Y. Jiang, An investigation of the academic information finding and re-finding behavior on the web, *Journal of Library & Information Studies*, vol.12, 2014.

[17] S. Whiting, J. M. Jose and O. Alonso, SOGOU-2012-CRAWL: A crawl of search results in the Sogou 2012 Chinese query log, *Proc. of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.709-712, 2016.

[18] M. A. Potey, D. A. Patel and P. Sinha, A survey of query log processing techniques and evaluation of web query intent identification, *2013 IEEE the 3rd International Advance Computing Conference*, pp.1330-1335, 2013.

[19] D. Jiang and L. Yang, Query intent inference via search engine log, *Knowledge and Information Systems*, pp.1-25, 2016.

[20] R. Song, Z. Luo, J.-Y. Nie, Y. Yu and H.-W. Hon, Identification of ambiguous queries in web search, *Information Processing & Management*, vol.45, pp.216-229, 2009.

[21] K. Zhu and F. Wei, A new query expansion method based on user logs mining, *Computer Applications and Software*, vol.6, p.032, 2012.

[22] C. Luo, Y. Liu, M. Zhang and S. Ma, Query ambiguity identification based on user behavior information, *Information Retrieval Technology*, Springer, pp.36-47, 2014.

[23] R. Ibrahim, S. Kamal, I. Ghani and S. R. Jeong, A hybrid query disambiguation adaptive approach for web information retrieval, *KSII Trans. Internet and Information Systems*, vol.9, pp.2468-2487, 2015.

[24] S. Kamal, R. Ibrahim and I. Ghani, Post-search ambiguous query classification method based on contextual and temporal information, *Asian Conference on Intelligent Information and Database Systems*, pp.575-583, 2016.

[25] S.-J. Kim and J.-H. Lee, Subtopic mining using simple patterns and hierarchical structure of subtopic candidates from web documents, *Information Processing & Management*, vol.51, pp.773-785, 2015.

[26] S.-J. Kim, J. Shin and J.-H. Lee, Subtopic mining based on three-level hierarchical search intentions, *European Conference on Information Retrieval*, pp.741-747, 2016.

[27] X. Shen, B. Tan and C. Zhai, Context-sensitive information retrieval using implicit feedback, *Proc. of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.43-50, 2005.

[28] S. M. Beitzel, E. C. Jensen, D. D. Lewis, A. Chowdhury and O. Frieder, Automatic classification of web queries using very large unlabeled query logs, *ACM Trans. Information Systems*, vol.25, p.9, 2007.

[29] D. Shen, R. Pan, J.-T. Sun, J. J. Pan, K. Wu, J. Yin et al., Query enrichment for web-query classification, *ACM Trans. Information Systems*, vol.24, pp.320-352, 2006.

[30] X. Shen, S. Dumais and E. Horvitz, Analysis of topic dynamics in web search, *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, pp.1102-1103, 2005.

[31] Y. Li, Z. Zheng and H. K. Dai, KDD CUP-2005 report: Facing a great challenge, *ACM SIGKDD Explorations Newsletter*, vol.7, pp.91-99, 2005.

[32] S. K. Tyler and J. Teevan, Large scale query log analysis of re-finding, *Proc. of the 3rd ACM International Conference on Web Search and Data Mining*, pp.191-200, 2010.