Study on the Click Context of Web Search Users for Reliability Analysis*

Rongwei Cen, Yiqun Liu, Min Zhang, Liyun Ru, and Shaoping Ma

State Key Laboratory of Intelligent Technology and Systems,
Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology, Tsinghua University, Beijing, China
crw@mails.tsinghua.edu.cn

Abstract. User behavior information analysis has been shown important for optimization and evaluation of Web search and has become one of the major areas in both information retrieval and knowledge management researches. This paper focuses on users' searching behavior reliability study based on large scale query and click-through logs collected from commercial search engines. The concept of reliability is defined in a probabilistic notion. The context of user click behavior on search results is analyzed in terms of relevance. Five features, namely query number, click entropy, first click ratio, last click ratio, and rank position, are proposed and studied to separate reliable user clicks from the others. Experimental results show that the proposed method evaluates the reliability of user behavior effectively. The AUC value of the ROC curve is 0.792, and the algorithm maintains 92.8% relevant clicks when filtering out 40% low quality clicks.

Keywords: User behavior analysis; click reliability; search user; search engine.

1 Introduction

User feedback provides useful information for analyzing, estimating and optimizing the performance of Web retrieval systems. It is an important topic for both IR researchers and search engine system engineers. Previous studies indicate that users are unwilling to provide explicit feedback for search engines [3]. Therefore, more studies (e.g. [4][5][19]) looked into implicit feedback information extracted from click-through data. This kind of feedback information has developed into an important research topic in the area of information retrieval and knowledge management, and has also been emphasized by commercial search engine community.

Unfortunately, practical Web data sources as well as click through logs contain lots of noise. Individual users may behave irrationally or maliciously, or may not even be real users, and we cannot treat each user as an individual "expert" [4]. By performing eye-tracking study, Joachims et al. [5] showed that individual user clicks include bias

* Supported by the Chinese National Key Foundation Research & Development Plan (2004CB318108), Natural Science Foundation (60621062, 60503064, 60736044) and National 863 High Technology Project (2006AA01Z141).

G.G. Lee et al. (Eds.): AIRS 2009, LNCS 5839, pp. 397–408, 2009.

and was not able to be used as judgments of absolute relevance directly. Therefore, state-of-the-art approaches require a large volume of click-through data to extract credible user feedback information based on the wisdom of the crowd (eg. [1][2][4][5] [12][13][18][19]). The consequent problem is that these methods only deal with hot queries with large number of access users and are not applicable for long-tail queries with rare user accesses.

To solve the problem, this paper defines the click reliability as whether a click is treated as one of an individual "expert" for relevance labeling probabilistically. Based on large scale log analysis, we study the contexts of individual user clicks and look into user decision process. Five context features of user clicks are proposed and employed to estimate click reliability.

The remaining part of the paper is as follow. Section 2 introduces the related work. Section 3 gives definition on users' click reliability in a probabilistic notion. In section 4, context of search behavior is been studied empirically and several features are proposed and verified for reliability analysis. Conclusions and future work are drawn in the section 5.

2 Related Work

In recent years, using implicit feedback information has been receiving much attention in the information retrieval area. Several approaches are proposed to mine relevant information from click-through data and some applications are implemented based on user behavior information, such as re-ranking search results [18], learning ranking [19], evaluating performance automatically [12], et al.

Tan et al. [6][7] detected robot behaviors for increasing the robustness of data and mining techniques applied to Web logs. Baeza-Yates et al. [8] and Kammenhuber et al. [9] modeled user clicks, query formulations and pages visited and revealed several interesting aspects of user behavior: users tend to formulate short queries, click on few pages and majority of users refine their initial queries in order to retrieve relevant documents. Sadagopan et al. [17] identified typical and atypical user sessions in click streams based on detecting outliers using Mahalanobis distance. These studies interpreted and modeled general user behaviors, detected robot behaviors or atypical ones, without analyzing and comparing different behaviors of one user.

In 2005, Joachims et al. [5] studied a work called "Eye-tracking", and mined implicit feedback information through analyzing user decision process. Based on this work, Agichitein et al. [4] depicted the rank bias of clicks, proposed background model and several user behavior models to mine relevant query-doc pairs. Recently, Craswell et al. [10] and Guo et al. [11] drew a cascade model, in which users view results from top to bottom and leave as soon as they see a worthwhile document, for explaining position bias of user behavior. However, some of these studies were performed over controlled lab conditions, which is not clear whether these techniques and rules will work for general real-world search. Some studied user behavior using statistics methods without analyzing individual user behaviors, which needed a large volume of user clicks for each query and cannot be adapted to long tail queries.

This paper analyzes user behaviors based on the contexts of user click process, looking into user decision process, studying user click preference, and estimating click reliability.

3 Search Process and Click Reliability

Before analyzing user behaviors, we assume the interactive process between user and search engine and look into user decision process. The concept of click reliability is defined, which has applications in performance evaluation of context features.

3.1 User Search Process

There is an interactive process between user and search engine. Generally, a user submits a query to a search engine. Then search engine presents results. By comparing the information of search result list, such as title, snippets, URL, other results, the searcher clicks results. When satisfied, he might be left. Otherwise, he returns to the result list, clicks more pages or refines his query and keeps on searching.

Table 1. A case of an interaction between user and search engine for the same topic

No.	Time	Query	Rank	Page Clicked
1	20:58:58	丰田(*Toyata*)	6	www.autohome.com.cn/526/
2	21:02:34	丰田(*Toyata*)	5	www.autohome.com.cn/110/
3	21:03:23	丰田(*Toyata*)	6	www.autohome.com.cn/526/
4	21:04:11	上海大众 (*Shanghai Volkswagen*)	5	www.che168.com/che168/cardb/brand/brand_58.html
5	21:06:14	广州本田 (*Guangzhou Honda*)	3	car.autohome.com.cn/brand/32/
6	21:09:23	丰田(*Toyata*)	2	car.autohome.com.cn/brand/63/
7	21:10:20	丰田(*Toyata*)	4	price.pcauto.com.cn/brand.jsp?bid=31
8	21:11:20	丰田(*Toyata*)	10	www.che168.com/che168/cardb/brand/brand_24.html
9	21:12:43	丰田卡罗拉 (*Toyota Corolla*)	1	www.autohome.com.cn/526/
10	21:19:12	丰田卡罗拉 (*Toyota Corolla*)	11	www.autohome.com.cn/526/options.html

Table 1 shows a case of user click process, from which we conjecture that a user want to find information for purchasing a car, and it is likely that the goal is *Toyota Corolla*(丰田卡罗拉). In the search processing, the user refers to other cars, *Shanghai VolksWagen*(上海大众) and *Guangzhou Honda*(广州本田). He finally clicks a page of *Toyota Corolla*'s configuration, and we can guess the user's need is the information about car configuration. Based on user search process analysis, we have the idea

that each click happens in the context of interactive process between user and search engine system, which is able to derive user decision process, and user click reliability can be estimated and judged to mine the preference information.

3.2 User Click Reliability

By performing eye-tracking studies and analyzing users' decision process, Joachims et al. [5] show that clicks are informative but biased, and it is difficult to make the interpretation of clicks as absolute relevance and relative preferences derived from clicks are reasonably accurate on average. This paper estimates the click relevance using click reliability.

Definition: *User Click Reliability* $\Re$ is an estimated probability of the relevance between query q and document d, given the context F of click c, and is formalized as:

$$\Re(c(q,d)) = P(R(c(q,d)) = 1 \mid F) . \tag{1}$$

where $c(q,d)$ is a click c of query q and document d, $R(c(q,d))$ presents the relevance of q and d, when $R(c(q,d)=1$ means relevant and $R(c(q,d))=0$ means irrelevant, and F is the click context, such as other clicks and queries in the current user session.

The concept defined here is different from the traditional studies which select relevant clicks by counting large scale logs statistically. This definition evaluates the relevance of query-document pairs from individual user clicks.

Based on Bayesian theorem, we have:

$$P(R(c) = 1 \mid F) = \frac{P(F \mid R(c) = 1)}{P(F)} P(R(c) = 1) . \tag{2}$$

Here $P(R(c)=1)$ is the likelihood of relevant clicks in whole click set. If we just compare the values of click reliability in a given click corpus, $P(R(c)=1)$ can be regarded as a constant value and wouldn't affect the comparative results. The equation is rewritten as:

$$P(R(c) = 1 \mid F) \propto \frac{P(F \mid R(c) = 1)}{P(F)} . \tag{3}$$

Now consider the terms in equation (3), $P(F)$ is the probability of context feature F which can be estimated using the proportion of F in a given click corpus. $P(F|R(c)=1)$ is the probability of feature F in relevant click set and equals to the proportion of clicks with feature F in relevant click set. According to equation (3), the reliability of click c with context feature F is proportional to $P(F|R(c)=1)$ and inversely proportional to $P(F)$. Therefore, the expression $P(F|R(c)=1)/P(F)$ is able to estimate the performance of feature F and we define the concept of *Click Reliability Value (CRV)* as follows:

$$ClickReliabilityValue(CRV) = \frac{P(F \mid R(c) = 1)}{P(F)} . \tag{4}$$

According to equation (2), when *CRV* is larger than 1, namely $\frac{P(F \mid R(c)=1)}{P(F)} > 1$, then we have $P(R(c)=1|F)>P(R(c)=1)$, and it means that the clicks with feature *F* is more reliable than the clicks in whole corpus in a probabilistic notion.

4 Empirical Study on the Context of User Behavior

Traditionally, user clicks are considered as a proof of relevance between queries and documents, and state-of-the-art approaches requires extensive user interaction data to guarantee statistical reliability. However, for long-tail queries, the challenge is that there is insufficient click data for statistical analysis. To assure our approach working for long-tail queries, we extracted features based on individual user clicks instead of relying on global statistics oriented features. Hence, we look into user decision process, analyze user click behaviors, then observe and propose several features from individual user at click level.

4.1 Data

The former study [4] showed that the Web search is not controlled and the techniques from the controlled lab may not work for general real world. To study user behavior in the real world, we collected search engine access logs from Sep. 10, 2008 to Oct. 24, 2008, with the help of a commercial search company. These access logs contain more than 194 million user clicks, 91 million unique queries, and 58 million user search sessions. Information extracted from the access logs is shown in Table 2.

Table 2. Information sources in the click-through logs collected

Item	Record Content
Query	The user query submitted
URL	URL of the result clicked by the user
Rank	The rank of the result clicked by the user
Order	The order of the result in the click sequence
User ID	Automatically assigned user's identification code
Time	Data and time of the clicking or querying event

For evaluating different context of relevant clicks and irrelevant ones, we randomly sampled 3000 queries from query logs. For each query, the top 20 results returned by 5 search engines in China were manually annotated as relevant or not by three assessors from a search engine company. Each assessor annotated about one third of the whole pooling set. The correctness of their annotation was also examined by co-checking each other's judgment results over a small subset containing 1000 query-doc pairs, and the kappa coefficient [20] measures agreement among these three assessors. Table 3 shows that the kappa value between any two assessors is large than 0.8, which means that our annotation is good reliability.

Table 3. The kappa statistic value of the manual annotation

	Assessor 1	Assessor 2	Assessor 3
Assessor 1	1.00	0.84	0.86
Assessor 2	0.84	1.00	0.87
Assessor 3	0.86	0.87	1.00

After annotation process, we have 89 thousands relevant query-doc pairs. With these pairs, 1.295 million click logs are picked out which have the same query-doc pairs. These clicks are treated as relevant ones reliably and annotated as Rel-Set and the whole click logs in our click-through data are annotated as Whole-Set.

4.2 Click Context Features

In the interaction between Web user and search engine systems, there are uncertainties in the process of user search and click. Firstly, we study the uncertainties in user query and click process and it is a possibility that there are two types of uncertainties, query uncertainty and click uncertainty. When a user clicks an irrelevant result, he/she is not satisfied and still need more information. Then the query tends to be refined and resubmitted to search engine, or more results will be clicked. We summarize these two types of uncertainties of click context as query number feature and click entropy feature.

(I) *QueryNum*: the unique query number submitted in current search session.

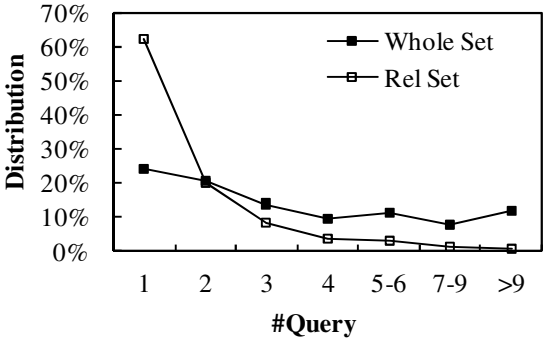

Fig. 1. QueryNum distributions of Rel-Set and Whole-Set. The category axis represents unique query number in current search session.

Fig. 1 shows that the context of 62% relevant clicks only contain one submitted query, which is larger than Whole-Set. The *QueryNum* of most clicks (82.1%) in Rel-Set is less or equal than 2 (only 44.9% for Whole-Set). Based on Fig. 1, we are able to derive that clicks with more submitted queries in their session context are less reliable. The *Click Reliability Value (CRV)* is 2.55, when *QueryNum* equals to 1, according to equation (4), which means that it works well for identifying click reliability. This context feature illustrates the existence of query uncertainty in process of query summiting.

(II) *ClickEntropy*: the information entropy (proposed by Shannon in 1948 [15]) of user click distribution in current search session, which is calculated as follows:

$$ClickEntropy = -\sum_{p_i} p_i \log(p_i) . \tag{5}$$

Here, p_i is the click distribution, estimated using the proportion of click on result i in all clicks in a user session and calculated as:

$$p_i = \frac{\#(click\ on\ result\ i)}{\#(total\ clicks\ in\ a\ user\ session)} . \tag{6}$$

p_i is different from the traditional Click-through Rate (CTR) metric [4][19]. The CTR is a statistic metric based on all user clicks, while p_i is based on current session.

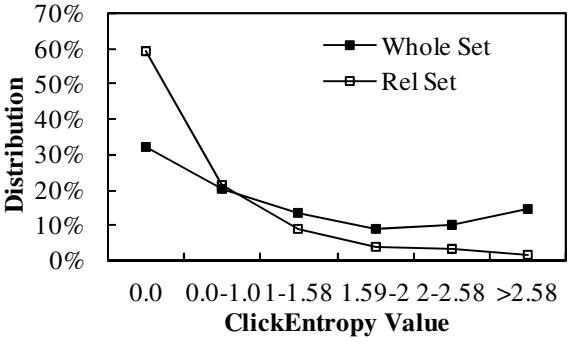

Fig. 2. ClickEntropy distributions of Rel-Set and Whole-Set. The category axis represents ClickEntropy value.

Fig. 2 shows that the *ClickEntropy* of Rel-Set is lower than Whole-Set's. When *ClickEntropy* equals to 0, namely user only click one page (one time or click the same page several times), it is 60% for Rel-Set, while it is only 32.2% for Whole-Set. Similar to *QueryNum* feature, the *ClickEntropy* of most clicks (81.1%) in Rel-Set is less than or equal to 1 (only 52.4%% for Whole-Set). According to *ClickEntropy* feature, it is able to derive that the click with more pages clicked in its session context is less reliable and *CRV*(*ClickEntropy*=1) is 1.85. This context feature illustrates the existence of click uncertainty in process of clicking result pages.

In a search process, each click is able to be sorted according to click time. Secondly, we conclude that the reliability is different for different order of click sequence. In [5], Jochims et al. proposed a rule, "Click > Earlier Click", which means that the later click is more relevant than earlier click, though the result of [13] shows that the rules of contradicting the existing search order perform worse compared to the rules that fully or partially reinforce the existing order of search result.

(III) *FirstClickRatio*:
which is deined as ***FirstClickInSession/FirstClickInQuery***: a click is the first one of a click sequence of session or query, or not. Here, we observe user click in two different scales of click context, session scale for the same user session and query scale for

the same query submitted by user. Since a user may submit several queries in current session, the scale of query is smaller than the one of session.

Fig. 3 shows that the 62% clicks in Rel-Set is the first ones of click sequence in user session, while it is 25.8% for clicks in Whole-Set. Similarly, 55.4% clicks in Rel-set are the first ones in query click sequence (25.8% for clicks in Whole-Set). The *CRV*(*FirstClickInSession*=*yes*) is 2.15 and the *CRV*(*FirstClickInQuery*=*yes*) is 1.42, which means that clicks with first position of click sequence in user session or query are more reliable than other ones. These two context features may be interpreted by the phenomenon that users may prefer to pay more attention to result lists and compare more information of each result before clicking any results and after first clicks, users' clicks tend to be less informative.

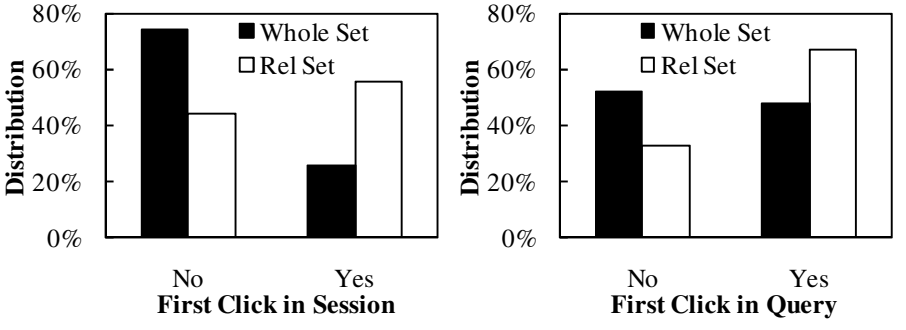

Fig. 3. FirstClickInSession/FirstClickInQuery distributions of Rel-Set and Whole-Set

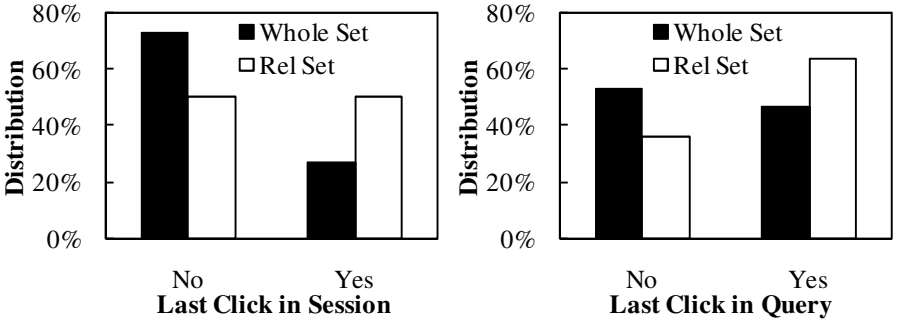

Fig. 4. LastClickInSession/LastClickInQuery distributions of Rel-Set and Whole-Set

(IV) ***LastClickRatio*****:**
which is defined as ***LastClickInSession/LastClickInQuery***: a click is the last one of a click sequence of user session or query, or not, analogous to the features *FirstClick-InSession* and *FirstClickInQuery*.

Fig. 4 displays the distributions of these two context features and shows that 50% clicks in Rel-Set are the last ones in user session (63.6% in user query), while it is 26.7% in user session for Whole-Set (36.4% in user query). These two features can be interpreted by that users tend to stop interaction with search engines when they finally

take the satisfying documents. These two features is special cases of the rule "Click > Earlier Click" in [5].

These four sequence context features look into different orders of click sequences based on session and query scale respectively, and the performances of session scale are better than the ones of query scale. By log analyzing, 71% users submit query only once in user session (Yu et al. have similar findings in [14]). For sessions with one query, the features, *FirstClickInSession* and *LastClickInSession*, is consistent with the other two, *FirstClickInQuery* and *LastClickInQuery*. Therefore, there are high correlations between these two groups of features with different scale. By studying the logs, the correlation between *FirstClickInSession* and *FirstClickInQuery* is 0.654 and it is 0.654 between other two features, which shows that these two groups of features are dependent.

Due to the high correlations, the first/last clicks of session sequences are filtered out from the sets of first/last clicks of query sequences. Fig. 5 shows the distributions of modified features of first/last clicks in query sequences, which are almost the same for both sets. According to the distributions in Fig. 5, we can derive that the first/last clicks of session sequences perform well, while the ones of query sequences fail to filter out reliable clicks. The user decisions under first/last clicks of query sequence, not session sequence, have the similar properties to middle clicks in sequence.

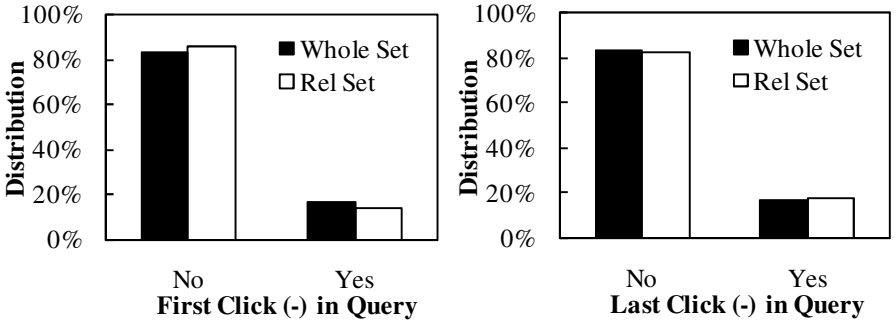

Fig. 5. The distributions of first clicks and last clicks in query sequences, not in sessions

Besides the features above, there are other context features, such as rank position of results, which is considered as rank bias [4][5][10][11] and causes the difficulties of log mining. To rank position factors, there are two different viewpoints. One is that the results at top positions have more possibilities to be viewed and clicked, the other is that search engines are experts in ranking the relevant documents at top positions. Here, we study user clicks with different result positions to look into the rank feature.

(IV) *RankPosition*: the rank position of click result.

Fig. 6 presents the rank distribution of Rel-Set and Whole-Set. For both of the sets, the results at top positions have more chance to be viewed and clicked than lower positions, and this phenomenon is defined as rank bias [4][5][10][11]. 30.3% users click the result at first position for Whole-Set, while it is 47.8% for Rel-Set, which means that the search engine may supply more intelligence for the first position than general rank bias. The reason may be that search engineers pay more attention to first rank position, and employ more rules or strategies.

According to the above analysis of five context features, we find out that processes of user clicks are influenced by result lists returned, and user decisions are applicable for interactions between users and search engines. By studying the context features in user click logs, we can look into user decision process and mine performing information to estimate click reliabilities.

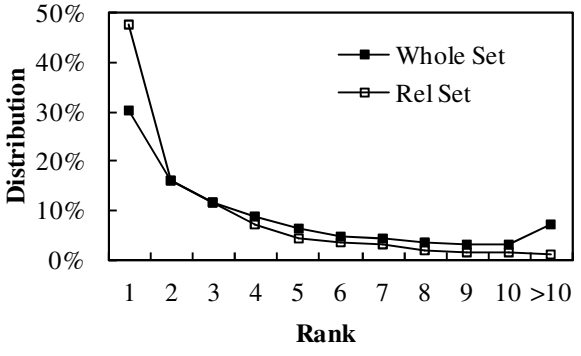

Fig. 6. The distribution of clicks at different rank positions

4.3 Experiments on Click Reliability Estimation

In this session, we estimated user clicks with context features proposed above, and selected high reliable clicks. According to the formula (3) in Session 3.2, clicks were estimated using one feature. Naïve Bayes theorem allowed us to use multiple features to estimate click reliability and Naïve Bayes method is applied to export the possibility value of a click with context features being a reliable one.

For evaluation, the Whole-Set and Rel-Set introduced in Session 4.1 were used to train and test, 2/3 data for training and 1/3 for test. After Bayesian learning, each click was assigned a score. We choose ROC (Receiver Operating Characteristic) curves and corresponding AUC (Area under the ROC Curve) values to evaluate the performance of our method. ROC is a useful technique for organizing classifiers and visualizing their performance and it is also adopted by quality estimation [16].

After learning and testing, ROC curves of our algorithm are shown in Fig. 7. From this figure, we see that our method is able to select reliable clicks probabilistically, which is better than random selecting method. The AUC value for the algorithm's ROC curve is 0.792, which means our estimation algorithm has 79.2% chances to rank a reliable click higher than a non-reliable click, while the AUC value for the random curve is 0.5.

The ROC curve shows that the high reliable clicks are able to be selected using our algorithm probabilistically. Table 4 lists that when we filter out 80% low reliable clicks, the algorithm can maintain 60% relevant click and we filter out 40% low reliable clicks, the algorithm can maintain 92.8% relevant click, which shows the effectiveness performance of our algorithm.

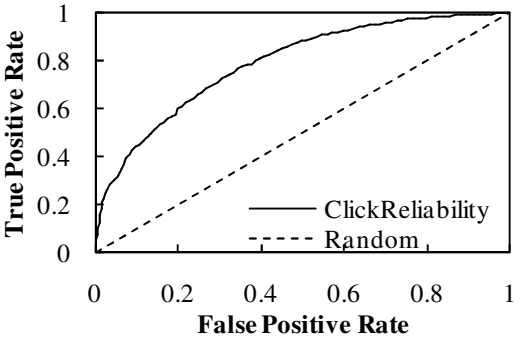

Fig. 7. ROC curves to evaluate the performance of click selecting method, compared with random method

Table 4. Different cleansed data size and corresponding relevant click recalls (the proportion of retained relevant clicks)

Cleansed data size	20.0%	40.0%	60.0%	80.0%
Relevant click recall	60.0%	81.4%	92.8%	98.4%

5 Conclusion and Future Work

In this paper, we analyze the contexts of user click behavior in interaction processes between users and search engine systems, and look into user decisions. The definition of click reliability is defined in a probabilistic notion. Five user context features are proposed and analyzed. The main conclusions are listed as follows:

[1] There are uncertainties in user query and click process, and clicks with more certainties are more reliable;
[2] The first and last clicks in click sequences have higher click reliability than the others;
[3] User decision process and search results influence user click behaviors and context features are effective for finding reliable clicks.

In the future, more work will be done on the application of reliability click estimation, such as improving Web search ranking, evaluating search engine performance, detecting click spam, finding bad search cases, etc.

References

1. Yates, R., Tiberi, A.: Extracting semantic relations from query logs. In: Proceedings of the 13th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, pp. 76–85. ACM, New York (2007)
2. Fuxman, A., Tsaparas, P., Achan, K., Agrawal, R.: Using the wisdom of the crowds for keyword generation. In: Proceeding of the 17th international Conference on World Wide Web, pp. 61–70. ACM, New York (2008)

3. Joachims, T., Freitag, D., Mitchell, T.: WebWatcher: a tour guide for the world wide Web. In: IJCAI 1997, vol. 1, pp. 770–777. Morgan Kaufmann, San Francisco (1997)
4. Agichtein, E., Brill, E., Dumais, S., Ragno, R.: Learning user interaction models for predicting web search result preferences. In: Proceedings of the 29th ACM SIGIR Conference on Research and Development in information Retrieval, pp. 3–10. ACM, New York (2006)
5. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: Proceedings of the 28th ACM SIGIR Conference on Research and Development in information Retrieval, pp. 154–161. ACM, New York (2005)
6. Tan, P., Kumar, V.: Modeling of web robot navigational patterns. In: Proceedings ACM WebKDD Workshop (2000)
7. Tan, P., Kumar, V.: Discovery of web robot sessions based on their navigational patterns. Data Mining and Knowledge Discovery 6, 9–35 (2002)
8. Yates, R., Hurtado, C., Mendoza, M., Dupret, G.: Modeling user search behavior. In: Proceedings of the 3th Latin American Web Congress. LA-WEB, p. 242. IEEE Computer Society, Los Alamitos (2005)
9. Kammenhuber, N., Luxenburger, J., Feldmann, A., Weikum, G.: Web search clickstreams. In: Proceedings of the 6th ACM SIGCOMM Conference on internet Measurement, pp. 245–250. ACM, New York (2006)
10. Craswell, N., Zoeter, O., Taylor, M., Ramsey, B.: An experimental comparison of click position-bias models. In: Proceedings of the international Conference on Web Search and Web Data Mining, pp. 87–94. ACM, New York (2008)
11. Guo, F., Liu, C., Wang, Y.M.: Efficient multiple-click models in web search. In: Proceedings of the 2nd ACM international Conference on Web Search and Data Mining, pp. 124–131. ACM, New York (2009)
12. Liu, Y., Cen, R., Zhang, M., Ru, L., Ma, S.: Automatic Search Engine Evaluation Based On User Behavior Analysis. Journal of Software 19(11), 3023–3032 (2008)
13. Agrawal, R., Halverson, A., Kenthapadi, K., Mishra, N., Tsaparas, P.: Generating labels from clicks. In: Baeza-Yates, R., Boldi, P., Ribeiro-Neto, B., Cambazoglu, B.B. (eds.) Proceedings of the 2nd ACM international Conference on Web Search and Data Mining (2009)
14. Yu, H., Liu, Y., Zhang, M., Ru, L., Ma, S.: Research in Search Engine User Behavior Based on Log Analysis. Journal of Chinese Information Processing 21(1), 109–114 (2007)
15. Shannon, C.E.: A Mathematical Theory of Communication. Bell System Technical Journal 27, 379–423, 623–656 (1948)
16. Svore, K., Wu, Q., Burges, C., Raman, A.: Improving Web Spam Classification using Rank-time Features. In: Proceedings of AIRWeb 2007, pp. 9–16. ACM, New York (2007)
17. Sadagopan, N., Li, J.: Characterizing typical and atypical user sessions in clickstreams. In: Proceedings of the 17th international Conference on World Wide Web, pp. 885–894. ACM, New York (2008)
18. Agichtein, E., Brill, E., Dumais, S.: Improving web search ranking by incorporating user behavior information. In: Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval, pp. 19–26. ACM, New York (2006)
19. Dou, Z., Song, R., Yuan, X., Wen, J.: Are click-through data adequate for learning web search rankings? In: Proceeding of the 17th ACM Conference on information and Knowledge Management, pp. 73–78. ACM, New York (2008)
20. Carletta, J.: Assessing Agreement on Classification Tasks: The Kappa Statistic. Computational Linguistics 22(2), 249–254 (1996)