Estimating Credibility of User Clicks with Mouse Movement and Eye-Tracking Information*

Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma

[1] State Key Laboratory of Intelligent Technology and Systems
[2] Tsinghua National Laboratory for Information Science and Technology
[3] Department of Computer Science and Technology, Tsinghua University, Beijing 1000084, China
maojiaxin@gmail.com, {yiqunliu,z-m,msp}@tsinghua.edu.cn

Abstract. Click-through information has been regarded as one of the most important signals for implicit relevance feedback in Web search engines. Because large variation exists in users' personal characteristics, such as search expertise, domain knowledge, and carefulness, different user clicks should not be treated as equally important. Different from most existing works that try to estimate the credibility of user clicks based on click-through or querying behavior, we propose to enrich the credibility estimation framework with mouse movement and eye-tracking information. In the proposed framework, the credibility of user clicks is evaluated with a number of metrics in which a user in the context of a certain search session is treated as a relevant document classifier. With an experimental search engine system that collects click-through, mouse movement, and eye movement data simultaneously, we find that credible user behaviors could be separated from non-credible ones with a number of interaction behavior features. Further experimental results indicate that relevance prediction performance could be improved with the proposed estimation framework.

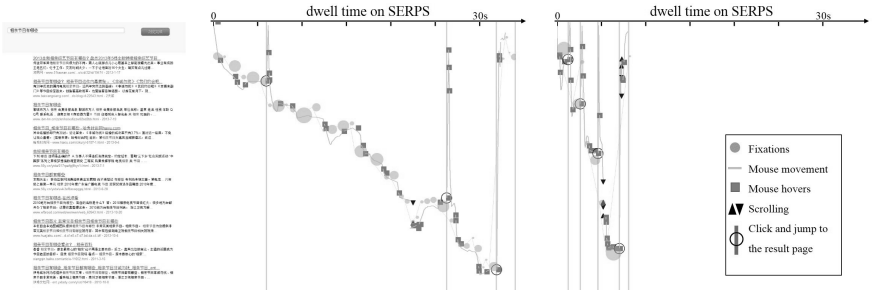
Introduction

With the growth of information available on the Web, search engines have become a major information acquisition platform. Search engines try to retrieve and display *relevant* results with respect to the query issued by the user. To a large extent, the success of a Web search engine depends on how well it can estimate the relevance between a query-result pair. Recently, the interaction logs of users, especially the click-through data, have been shown to have great value in improving Web search engines. This is because they can be used as implicit relevance feedbacks from users and they are easy to collect on a large-scale (e.g. [11,12]). The most common approaches to exploit click-through data involve the training of *generative models* (named click models) to model users' click behaviors and extract relevance information (e.g. [3,2]).

However, not every click is a good indicator of relevance. During a search session, some users tend to rely on the snippets provided by search engines to determine the

* This work was supported by National Key Basic Research Program (2015CB358700) and Natural Science Foundation (61472206, 61073071) of China.

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 263–274, 2014.



(a) SERP (b) Session A (c) Session B

Fig. 1. Two search sessions of a same query. (a) shows the SERP of the query, (b) is for Session A and (c) is for Session B. In (b) and (c), we plot the y coordinates of the eye and mouse against the dwell time on the SERP. Note that the y coordinates are aligned across (a),(b) and (c). The size of the red circle indicates the duration of the fixation.

relevance of the results. They may read the SERPs (Search Engine Result Pages) and click on search results after thorough consideration. In such situations, these clicks are more likely to be informative indicators of users' relevance judgments. Other users may not be skilled at making relevance judgments or may click on multiple results in a short time without attentively selecting the results. In those case, the extent to which these clicks can be used as implicit relevance feedbacks is questionable.

Previous studies showed that search engine users have different behavioral patterns while browsing and clicking results on SERPs [21]. Users who use advance syntax [23] or possess domain knowledge [22] are regarded as *expert users* [7], and are expected to be more likely to provide credible relevance feedback information. These findings suggest that the clicks made by different users should be treated differently. Xing et al. [24] characterized the expertise of search engine users and proposed an unsupervised method to find expert users for improving relevance estimation. We follow the idea that we can regard the user as a relevant document classifier and then use some *metrics* to measure its performance. To take this analysis a step further, we claim that the credibility of clicks is not only related to the users' expertise in finishing search tasks but also depends on other factors, including the concentration and willingness of users. Different from previous works, which are usually based on merely click-through logs, we try to exploit a more comprehensive behavioral log, which involves eye-tracking and mouse movement behavior of search users. Rather than extracting *user-level* credibility estimation, we characterize click credibility on a *session-level*[1] because we believe that the relevance judgment performance of a certain user varies from query to query and should not be regarded as the same among different queries. Figure 1 shows two search sessions on a same query. The user in session A read the results in an approximate linear order, with the mouse following the gaze position, leaving many hovers on the results, and clicked on result 2, 7, 10. The user in session B browsed the SERP in a hurry, left fewer gazes and mouse hovers, and clicked on result 1, 3, 5, 9. Since the relevant results

[1] By *session* we mean a unique user-query pair in a continuous time period.

of this query are result 1, 2 and 7, we can see that click behaviors in Session A are more reliable because 2 of 3 clicks are on relevance results. This case serves as an example that the click credibility varies across sessions and user behavioral logs can be used to estimate it.

In this research, we try to determine the extent to which we can *trust* the click-through data by exploring more interaction behavior information collected from users. To investigate this problem, we built an experimental search engine with which we can collect detailed user behavioral data. The collected data include the click logs, mouse movement logs and eye-tracking information. We also hired accessors to manually annotate all the query-result pairs used in the experiment in 4-level relevance score. With this comprehensive dataset, we compared a variety of metrics to characterize the credibility of user clicks.

The major contributions of this paper can be summarized as follows:

- We proposed a framework and a variety of metrics to characterize the credibility of clicks on a session level.
- Based on a number of interaction behavior features extracted from mouse movement, eye movement and click-through behavior, we constructed a learning-based framework to estimate the credibility of search users' behaviors.
- We demonstrated that the predicted credibility of clicks can be used to improve the relevance estimation of the query-result pairs, which could help to improve the ranking performance of search engines.

Related Work

Click Models

Although click-through data are informative, previous works showed that they can not be interpreted as implicit relevance feedback directly because they are biased [12]. One of the major biases is the *position bias*. That is, the results that are ranked higher in the SERPs are more likely to be examined by users and therefore to be clicked by users. To address this bias, *click models* were proposed to estimate the probability of search result being examined and being clicked (e.g.[2,3]). In most of the click models, the *examination hypothesis* [3,16], that the user will click on a result *if and only if* he or she examines it and regards it as relevant, is made. Thus, after acquiring the examination information and the click-through data by mining the log, the relevance can be estimated.

Identifying Expert Users

According to previous works, the Web search engine users behave quite differently. [21] used a search trail extraction method to investigate the behavioral variability among users and among queries. They reported two extreme classes of users. The 'Navigators' have consistent interaction patterns, whereas the 'Explorers' have interaction patterns with a much higher variability. [23] compared the users who use advance syntax with other users. In addition to a significant difference in behavioral patterns, they found that the relevance scores of the results clicked by the advance syntax users were higher.

Similar research on users who are domain experts [22] also showed that these domain expert users are more likely to successfully find relevance results when issuing a in domain query.

Exploiting Eye-Tracking and Mouse Movement Information

Eye-tracking technology has attracted extensive attention from search engine researchers. Eye-tracking devices can record users' real-time eye movements and help to elucidate how users browse and interact with the SERP. [4] found that users spend more time examining the top results and that adding information to snippets will improve the performance for informational tasks. [1] found that search engine users have two evaluation styles. The Exhaustive style user will read more snippets before making a click, whereas the Economic style user only scans a few results before the first action.

Mouse movement information is another behavioral data source that worth investigating. Previous works [8,17] suggested that mouse positions and movements can be used to predict gaze positions. [6] exploited mouse movement data and click logs to predict the success of search sessions. Huang et al. found that *hovering* on results can be interpreted as implicit relevance feedback [10], and can be considered as a signal of examination in click models [9]. [18] built an end-to-end pipeline that can collect relevance annotations, click data and mouse movement data simultaneously. They also used several mouse movement features and the collected relevance annotations to train a classifier to predict the relevance of other results.

Our work is related to but different from these works. The click models enable us to estimate relevance by merely mining the click-through data. However, the click-through data are not always reliable, which motivates us to estimate the credibility of clicks. Additionally, the differences at the user level (e.g. the user's search expertise) is one of, but not the most, decisive influencing factors for the credibility. Thus in this paper, we leverage more interaction information, especially mouse movement data, to estimate click credibility at a session level, not a user level.

User Behavior Dataset

To analyze the session-level click credibility, we built an experimental search engine and recruited 31 subjects (15 males and 16 females) to conduct an experiment. The subjects were first-year university students from two different majors and with a variety of self-reported expertise in using search engines.

In the experiment, each subject was asked to accomplish 25 search tasks selected from NTCIR IMine task [2]. Each task corresponded to a specific query and necessary explanation to avoid ambiguity. We selected 5 navigational queries, 10 informational queries, and 10 transactional queries to cover various types of information needs. The queries and explanations are listed in Table 1. We crawled the first result page for each task from a Chinese commercial search engine and because the vertical results and the advertisement may affect users' behaviors [20], we filtered out the vertical results and advertisement. For each task, we restricted all the subjects to use the same query and browse the same SERP with 10 ordinary results.

[2] (`http://www.thuir.cn/imine`)

Table 1. Queries and explanations of 25 search tasks. For intent, 'I' indicates Informational, 'T' indicates Transactional, and 'N' indicates Navigational.

ID	Query	Translation	Explanation	Intent
1	央金兰泽的歌曲	Yangjinlanze's song	Find the information of the music sung by the singer	I
2	卫子夫	Empress Wei Zifu	Find the introduction of the historical figure	I
3	天梭手表官网	Official website of Tissot	Find the official website of a watch brand	N
4	温柔的谎言	Gentle lies	Find the online watching resources of a TV drama	T
5	佛教音乐	Buddhist music	Find Buddhist music download resources	T
6	声卡是什么	What is a sound card	Find the definition and principal of sound cards	I
7	qq加速器下载	qq accelerator download	Find the download resources of a software	T
8	浏览器下载	Browser download	Find the download resources of Web browsers	T
9	相亲节目有哪些	What dating shows are there on TV	Find major TV dating shows	I
10	遮天	Zhe Tian	Find online reading resources of a novel	T
11	工行网上银行个人网上银行	Personal online bank of CCBC	Find the website of a bank	N
12	冬季恋歌国语全集	Chinese collections of Winter Sonata	Find the online watching resources of a TV drama	T
13	乘法口诀	Multiplication table	Find the multiplication table for pupils	I
14	学雷锋作文	Writings of 'learn from Lei Feng'	Find writing samples of a given topic for pupils	I
15	驾驶证考试网	Website of driving license exam	Find the website for the driving license exam	N
16	春雨的诗句	Poems of spring rain	Find famous poems for spring rain	I
17	初恋这件小事	First Love	Find online watching resources for a movie	T
18	魅族官网	Official website of Meizu	Find the official website of a digital brand	N
19	斯特拉马乔尼	Stramaccioni	Find the resume of a soccer coach	I
20	哈利波特	Harry Potter	Find online watching resources for a movie	T
21	电脑桌面壁纸	Wallpaper for computer	Find wallpaper pictures for a desktop computer	T
22	清明上河园	Millennium City Park	Find photos of a viewpoint	I
23	安卓2.3游戏下载	Android 2.3 game download	Find download resources of mobile games	T
24	陈楚生演唱会	Chen Chusheng's concert	Find the concert information of a singer	I
25	联通网上营业厅	Online service of China Unicom	Find the official website of a mobile company	N

To analyze each user's examination processes, we used a Tobii X2-30 eye-tracker to record each subject's eye movements during the experiment. The data recorded by the eye-tracker are sequences of fixations and saccades (fast eye movements between two fixations). According to [13], when reading, the subject will process only the content that he or she fixates on. Therefore, to reconstruct the examination sequence, we only took the fixations into account and regarded the fixation on a certain search result as a signal that the subject had examined it.

A variety of mouse events, including mouse movements, hovers on results, scrolling and clicks on result, were logged by injecting Javascript code into the crawled SERPs. We also hired 3 annotators to give objective relevance judgments of all the results. The relevance score had 4 levels, with 1 for least relevant and 4 for most relevant. We regarded results with a relevance score of 3 or 4 as relevant results, and the result with a score of 1 or 2 as not relevant. When disagreement occurred, we used the *median* of the scores from the three annotators as the final score.

From these 31 subjects, we collected 774 valid query sessions (one of the query sessions was abandoned because the eye-tracker malfunctioned during the experiment). Each query session has a comprehensive behavioral log and can be uniquely identified by the <*user, query*> pair.

Estimating Session-level Credibility

Credibility Metrics

When using a Web search engine, users always aim to find information or to accomplish a certain task, related to the issued query. During a search session, the user will obtain some information from the SERP and then click on the result that he or she believes to be helpful in fullfilling the information need. We can view this as a classification task [24]. The user is a classifier that takes the results as samples, the obtained information as input features, and outputs corresponding relevance judgments, by performing clicks or skips (no clicking after examination) over the results. Therefore, to use click-through data as implicit relevance feedbacks is to aggregate the outputs of multiple classifiers. The credibility of the click-through data generated in the session *is* the credibility of the output of the classifier, given a specific SERP as the input features. This analogy inspires us to use metrics that measures the performance of a binary classifier to characterize the session-level click credibility.

The confusion matrix [19] is a common evaluation metric in classification. We regard relevant results as positive samples and irrelevant results as negative samples. Each row of the matrix represents the instances in one of the predicted classes, which in our case is whether the user click or skip on the result, whereas each column represents the instances in one of the actual classes, that is, whether the result is actually relevant to the query. We assume that a user's click credibility is independent of the content (the results on SERPs) that he or she did not examine. This assumption is necessary because a user may not always examine all the results on the SERPs, and it does not make sense to judge the user's click credibility according to the results he or she did not examine. We used eye-tracking data as an explicit information source for examination to filter out the unexamined results. Based on the examination sequence, the click log of a query

session and the relevance score given by the annotators, we can compute the confusion matrix of the session. From the matrix, we derived 3 intuitive metrics for session-level click credibility:

- *Accuracy*: $\frac{\#\{TP\}+\#\{TN\}}{\#\{TP\}+\#\{TN\}+\#\{FP\}+\#\{FN\}}$[3]. The proportion of correctly classified samples.
- *True Positive Rate*: $\frac{\#\{TP\}}{\#\{TP\}+\#\{FN\}}$. The proportion that a relevant result is clicked by the user. True positive rate is also referred to as recall, and related to the user's ability to find relevant result on SERPs.
- *True Negative Rate*: $\frac{\#\{TN\}}{\#\{TN\}+\#\{FP\}}$. The proportion that an irrelevant result is skipped by the user. It relates to the user's ability to prevent unnecessary clicks, and it is called specificity in some situations.

The statistics on the metrics are showed in Table 2. For each of the metrics, we group the <*user, query*> pairs by the users and by the queries, separately compute the means and standard deviations of each group and show the macro average mean M and macro average standard deviation SD. We noticed obvious variability in all 3 metrics, which indicates the credibility of clicks varies across sessions. The standard deviations of the accuracy and true positive rate are lower when grouped by query, which reveals that the query is a more determining factor than the user. Therefore, it is necessary to estimate the credibility of clicks on the session level, not the user level.

Table 2. Statistics for metrics of click credibility

Metrics		Accuracy	True Positive Rate	True Negative Rate
Grouped by user	M	0.613	0.557	0.813
	SD	0.236	0.274	0.239
Grouped by query	M	0.613	0.557	0.811
	SD	0.200	0.257	0.247
Single session	M	0.613	0.557	0.812
	SD	0.249	0.306	0.269

Predicting Click Credibility

We have proposed 3 metrics to characterize the session-level click credibility. However, our primary concern is to find credible clicks to improve automated relevance feedback and the computation of the proposed metrics themselves requires manually annotation of the relevance. Therefore, we need to *predict* the metrics of click credibility when the relevance annotation is not available. In this section, we try to use the behavioral logs as features and predict the click credibility through a *regression* process. Because all the proposed metrics are probabilities $p \in [0, 1]$, it is more convenient to predict the corresponding *log-odds* $\alpha \in (-\infty, +\infty)$ of them:

$$\alpha = \text{logit}(p) = \log(\frac{p}{1-p})$$

[3] Here $\#\{TP\}$, $\#\{TN\}$, $\#\{FP\}$ and $\#\{FN\}$ are used to denote the number of true positives, true negatives, false positives and false negatives.

Feature Extraction. To make large scale deployment possible, we only used data that can be collected in practical application scenarios of search engines. That is, we did not use eye-tracking data and relevance annotations to generate the features. The features that we used and their correlations (measured in Pearson's r) with the metrics are list in Table 3. The click features are the features that can be generated from click-through logs. The session features are some time-related features of a certain session. The mouse movement features include hovers, mouse movement and scrolling information. Feature 10, the unclicked hovers, were proposed in [10]. The click entropy [5] of a query is defined as:

$$\text{ClickEntropy}(q) = -\sum_i P(C_i = 1|q) log(P(C_i = 1|q))$$

The N Results Satisfied Rate [14] is the proportion of the sessions in which only first N results were clicked. For the user features, we computed the click entropies of users, which is similar to the click entropies of queries, and also includes the total click numbers and time-related features.

Evaluations. The ground truths of proposed metrics are calculated using the eye-tracking data and the click-through data. As the whole dataset was limited in size (774 samples), we use the Leave-One-Out Cross Validation (LOO-CV) to access the generalized prediction performance. Because Web search engines posses a large number of users and response to all kinds of queries, it is reasonable to assume that when predicting the session-level click credibility, the user and the query are *unseen* in the training set. Therefore, in every iteration, we chose one query and one user, the corresponding <*user, query*> pair was then used as test set, and the <*user, query*> pairs of the other 24 queries and the other 30 users formed the training set.

A series state-of-the-art regression methods, including the SVR, random forest regression, gradient boosting tree regression and Lasso[4], were tested using the training data. We finally used SVR, which is relatively stable and gains promising results, to predict the click credibility.

If behavioral information is not available, the best estimate of each of the metrics is the arithmetic mean of the metrics of the training set sessions. We used it as the baseline estimation method because to our best knowledge, there are no existing works which try to estimate user behavior credibility with mouse movement information. The results, measured in Mean Squared Error (MSE) and Mean Absolute Error (MAE) with the ground truth, are listed in Table 4.

Estimating Relevance

Having obtained the predicted log-odds α, we can compute the corresponding predicted metrics by $p = e^{\alpha}/(1 + e^{\alpha})$, and use them to improve the relevance estimation. Recalling the *examination hypothesis* [3], we can estimate the relevance score r_i by:

$$r_i = P(C_i = 1|E_i = 1) \tag{1}$$

[4] Implementations are provided in scikit-learn [15].

Table 3. Features extracted from the behavioral log with Pearson's r between features and metrics (TP:True Positive Rate, TN:True Negative Rate). * indicates $r \neq 0$ with $p < 10^{-3}$.

No.	Description	Accuracy	TP	TN
Click Features				
1	number of clicked results	-0.02	0.24*	-0.55*
2	lowest rank of clicks	-0.11	0.05	-0.36*
3	average difference in ranks between two clicks	-0.06	0.03	-0.16*
Session Features				
4	average time spent on the SERP for each click	-0.04	-0.05	0.09
5	total time spent on the search task	-0.07	0.03	-0.19*
6	total time spent on the SERP	-0.12*	-0.01	-0.22*
7	maximal continuous time spent on the SERP	-0.14*	-0.16*	0.06
Mouse Movement Features				
8	number of hovered results	-0.15*	-0.06	-0.26*
9	lowest rank of hovers	-0.18*	-0.10	-0.24*
10	number of results that are hovered over but not clicked	-0.18*	-0.23*	0.07
11	moving time of the mouse	-0.12	-0.02	-0.17*
12	idle time of the mouse	-0.06	0.10	-0.34*
13	dwell time of the mouse in the result region	-0.11	-0.03	-0.16*
14	length of the mouse trails	-0.08	0.08	-0.34*
15	velocity of mouse movement	0.12*	0.17*	-0.16*
16	horizontal moving distance of the mouse	-0.05	0.07	-0.19*
17	vertical moving distance of the mouse	-0.08	0.08	-0.37*
18	maximal y coordinate that the mouse has reached	-0.18*	-0.11	-0.22*
19	total distance of scrolling	-0.11	-0.09	-0.13
20	maximal displacement in y axis of scrolling	-0.15*	-0.13*	-0.14*
Query Features				
21	click entropy of the query	-0.17*	-0.13*	-0.13
22	N Results Satisfied Rate of the query	-0.08	-0.03	0.01
User Features				
23	click entropy of the user	-0.16*	-0.06	-0.21*
24	user's total number of clicks	-0.02	0.14*	-0.32*
25	average time that the user spends on each search task	-0.08	0.02	-0.20*
26	average time that the user spends on the SERP	-0.10	-0.05	-0.12

Table 4. Results of predicting click credibility, * indicates the improvement over baseline is significant with $p < 10^{-3}$

	Baseline		SVR	
	MSE	MAE	MSE	MAE
Accuracy	0.071075	0.224086	0.061015(-14.1%*)	0.206813(-7.7%*)
True Positive Rate	0.109941	0.290940	0.082651(-24.8%*)	0.219068(-24.7%*)
True Negative Rate	0.086662	0.192934	0.069311(-20.0%*)	0.165896(-16.6%*)

Where $E_i = 1$ indicates that result i is examined and $C_i = 1$ indicates that result i is clicked. If we know the accuracy of the session, a_s, then we have:

$$\begin{aligned} P(C_i = 1|E_i = 1) &= r_i \times a_s + (1 - r_i) \times (1 - a_s) \\ P(C_i = 0|E_i = 1) &= (1 - r_i) \times a_s + r_i \times (1 - a_s) \end{aligned} \tag{2}$$

If we know the true positive rate TP_s and the true negative rate TN_s, we have:

$$\begin{aligned} P(C_i = 1|E_i = 1) &= r_i \times TP_s + (1 - r_i) \times (1 - TN_s) \\ P(C_i = 0|E_i = 1) &= (1 - r_i) \times TN_s + r_i \times (1 - TP_s) \end{aligned} \tag{3}$$

By the criterion of maximum likelihood estimation, we can estimate r_i. We use the fixation information recorded by eye-tracker as explicit signal to estimate E_i. That is, during a session, if the user fixates on result i, we assume that he or she has examined this result and set $E_i = 1$. We use the relevance score given by (1) as the baseline, and refer to (2) as the accuracy model, (3) as the confusion matrix model.

To evaluate the performance in estimating the relevance, we use the predicted relevance score r_i to re-rank the results on SERP in a descending order, and use the Mean Average Precision (MAP) to measure the ranking performance. Higher MAP is associated with better estimation performance. We use the metrics computed by LOO-CV as the predicted metrics. The results are in Table 5. Both models can improve original rankings and significantly outperform the baseline; therefore, the predicted metrics can be utilized to improve relevance estimation. The accuracy model is slightly better than the confusion matrix model.

Table 5. Results of relevance estimation, * indicates the improvement is significant with $p < 0.05$ and ** indicates $p < 0.01$

	Original Ranking	Baseline (Examination hypothesis without credibility estimation)	Accuracy Model	Confusion Matrix Model
MAP	0.805124	0.843075	0.884604	0.877654
Improvement over original ranking	-	+4.7%	+9.9%**	+9.0%**
Improvement over baseline	-	-	+4.9%**	+4.1%*

Conclusions and Future Works

To estimate the credibility of click-through data of Web search engines, we conducted a user behavioral experiment to analyzed a detailed behavioral log. Based on the analogy that a search engine user can be regarded as a classifier that classifies the search results into a relevant class or an irrelevant class, we used 3 metrics, i.e, the accuracy, the true positive rate and the true negative rate, which are designed for measuring classification

performance, to measure session-level click credibility. We reported the correlations between these click credibility metrics and user behavioral features, and demonstrated that we can used the predicted metrics to improve implicit relevance feedback of search engines.

These metrics not only provide us with extra dimensions in understanding the interactions between users and search engines, but can be also applied to improving search engines. Therefore, our work can be considered as a demonstration of a framework that exploits the abundant but complex behavioral data. That is, we can extract some properties (e.g. the proposed click credibility metrics), which are either intuitive or useful, from the query sessions and try to predict them using the behavioral data and a supervised learning method. This process is a user behavior analysis that may produce new insights. And at the same time, we can use these properties to improve the search engine itself.

In future works, we will further explore more behavior-related properties, which may either help the user behavioral analysis or the enhancement of other search engine functionalities. On the other direction, we will try to collect mouse movement logs on a larger scale and validate our approach in a real-world dataset.

References

1. Aula, A., Majaranta, P., Räihä, K.-J.: Eye-tracking reveals the personal styles for search result evaluation. In: Costabile, M.F., Paternó, F. (eds.) INTERACT 2005. LNCS, vol. 3585, pp. 1058–1061. Springer, Heidelberg (2005)
2. Chapelle, O., Zhang, Y.: A dynamic bayesian network click model for web search ranking. In: Proceedings of the 18th International Conference on World Wide Web, pp. 1–10. ACM (2009)
3. Craswell, N., Zoeter, O., Taylor, M., Ramsey, B.: An experimental comparison of click position-bias models. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 87–94. ACM (2008)
4. Cutrell, E., Guan, Z.: What are you looking for?: an eye-tracking study of information usage in web search. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 407–416. ACM (2007)
5. Dou, Z., Song, R., Wen, J.R.: A large-scale evaluation and analysis of personalized search strategies. In: Proceedings of the 16th International Conference on World Wide Web, pp. 581–590. ACM (2007)
6. Guo, Q., Lagun, D., Agichtein, E.: Predicting web search success with fine-grained interaction data. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 2050–2054. ACM (2012)
7. Hölscher, C., Strube, G.: Web search behavior of internet experts and newbies. Computer Networks 33(1), 337–346 (2000)
8. Huang, J., White, R., Buscher, G.: User see, user point: gaze and cursor alignment in web search. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1341–1350. ACM (2012)
9. Huang, J., White, R.W., Buscher, G., Wang, K.: Improving searcher models using mouse cursor activity. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 195–204. ACM (2012)
10. Huang, J., White, R.W., Dumais, S.: No clicks, no problem: using cursor movements to understand and improve search. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1225–1234. ACM (2011)

11. Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 133–142. ACM (2002)
12. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 154–161. ACM (2005)
13. Just, M.A., Carpenter, P.A.: A theory of reading: From eye fixations to comprehension. Psychological Review 87, 329–354 (1980)
14. Liu, Y., Zhang, M., Ru, L., Ma, S.: Automatic query type identification based on click through information. In: Ng, H.T., Leong, M.-K., Kan, M.-Y., Ji, D. (eds.) AIRS 2006. LNCS, vol. 4182, pp. 593–600. Springer, Heidelberg (2006)
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
16. Richardson, M., Dominowska, E., Ragno, R.: Predicting clicks: Estimating the click-through rate for new ads. In: Proceedings of the 16th International Conference on World Wide Web, pp. 521–530. ACM (2007)
17. Rodden, K., Fu, X., Aula, A., Spiro, I.: Eye-mouse coordination patterns on web search results pages. In: CHI 2008 Extended Abstracts on Human Factors in Computing Systems, pp. 2997–3002. ACM (2008)
18. Speicher, M., Both, A., Gaedke, M.: Tellmyrelevance!: Predicting the relevance of web search results from cursor interactions. In: Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management, pp. 1281–1290. ACM (2013)
19. Stehman, S.V.: Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment 62(1), 77–89 (1997)
20. Wang, C., Liu, Y., Zhang, M., Ma, S., Zheng, M., Qian, J., Zhang, K.: Incorporating vertical results into search click models. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2013, pp. 503–512. ACM, New York (2013), http://doi.acm.org/10.1145/2484028.2484036
21. White, R.W., Drucker, S.M.: Investigating behavioral variability in web search. In: Proceedings of the 16th International Conference on World Wide Web, pp. 21–30. ACM (2007)
22. White, R.W., Dumais, S.T., Teevan, J.: Characterizing the influence of domain expertise on web search behavior. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining, pp. 132–141. ACM (2009)
23. White, R.W., Morris, D.: Investigating the querying and browsing behavior of advanced search engine users. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 255–262. ACM (2007)
24. Xing, Q., Liu, Y., Zhang, M., Ma, S., Zhang, K.: Characterizing expertise of search engine users. In: Banchs, R.E., Silvestri, F., Liu, T.-Y., Zhang, M., Gao, S., Lang, J. (eds.) AIRS 2013. LNCS, vol. 8281, pp. 380–391. Springer, Heidelberg (2013)