Incorporating User Preferences into Click Models

Qianli Xing*, Yiqun Liu*, Jian-Yun Nie†, Min Zhang*, Shaoping Ma*, Kuo Zhang*
*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 100084, China
†University of Montreal, CP. 6128, succursale Centre-ville, Montreal, Quebec H3C 3J7, Canada
xingqianli@gmail.com, yiqunliu@tsinghua.edu.cn, nie@iro.umontreal.ca
{z-m,msp}tsinghua.edu.cn, zhangkuo@sogou-inc.com

ABSTRACT

Click models are developed to interpret clicks by making assumptions on how users browse the search result page. Most existing click models implicitly assume that all users are homogeneous and act in the same way when browsing the search results. However, a number of researches have shown that users have diverse behavioral patterns, which is also observed in this paper by eye-tracking experiments and click-through log analysis. As a uniform click model for all users can hardly capture the diverse click behavior, in this paper we incorporate user preferences into both a variety of existing click models and a novelly proposed click model. The experimental results on a large-scale click-through data set show consistent and significant performance improvement of the click models with user preferences integrated.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Human Factors

Keywords

Web search, click model, user preference

1. INTRODUCTION

Query logs of a search engine, in particular, the click-through data, contains rich information of user satisfaction with the search results and thus is a valuable source in relevance inference. A lot of studies [1, 8, 13, 14, 16] have been carried out to extract useful information from click-through data and to better understand user interactions to help improve and evaluate the search quality.

A fundamental problem in modeling click-through data is the position bias. That is, the probability of a document being clicked depends not only on its relevance, but also on its position in the search result page. Previous study [14] has shown that the examination probability decays as the ranking position increases. Such bias has been taken into account in the previous click models, such as dependent click model [11], click chain model [10], user browsing model [9], dynamic Bayesian network click model [4] and et al. Some studies even introduced more complex variables in the examination assumption, such as users' revisit behavior [24] and the influence of vertical search results [23]. These previous models showed success in fitting the real-world data and predicting future clicks. However, what we observe is that in all these models, the users are assumed to act in the same way, i.e. to have the same examination and click behavior. This is obviously not true in reality, and has been documented in many studies. For example, White et al. showed dramatic differences between Web search users in some key aspects of the interaction, such as issued query, clicked result, and post-query browsing [19, 20]. It is also reported that domain experts search differently from users with little or no domain knowledge with respect to session length, site selection and search effectiveness [2, 22, 7, 21]. All these studies indicate that users differ with respect to their search behaviors. Therefore, click models based on a uniform user behavior assumption can hardly account for personal preferences of users.

The personal preferences of a Web search user can be reflected in many aspects, such as domain interest, search intent and behavioral habit. Recent researches of click model have incorporated user personality with respect to domain interest [18] and search intent [12]. The results are promising. From another point of view, we observe that one's behavioral habits in browsing and clicking have impact on all her/his queries, and should be taken into account in click models. To the best of our knowledge, few studies have considered user personality from this aspect in a click model.

In this paper, we aim to develop click models that incorporate user search habits and preferences. Such models are motivated by the strong differences in search behavior of users observed in our analysis of click logs of a search engine as well as the eye-tracking experiments with human subjects. We observe that some users tend to examine more documents than the others; and some users tend to click more documents. Based on the observations, we introduce examination preference and click preference to describe the general behavioral preferences of a user in Web search environment. We then build a variety of user-specific click models by incorporating these preferences, and the experi-

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2505704.

mental results show consistent and significant performance improvement for click models with user preferences integrated. The main contributions of this work are:

- We investigate users' personal preferences in examining and clicking by eye-tracking study and click log analysis. The results show considerable difference in users' search behavior. According to that, we propose examination preference and click preference to describe user personality in a general behavioral aspect.
- This work is the first attempt to take into account user preferences from the general behavioral aspect for click models. We build a variety of user-specific click models by incorporating the proposed preferences into both existing click models and our proposed model. The experiment results show consistent and significant improvement, which verifies the effectiveness of the proposed user preferences for click model.

The rest of this paper is organized as follows: In section 2, we review the related work on click model. In section 3, we investigate the diversity of user behavior by log analysis and eye-tracking study. In section 4, we build user-specific click models by incorporating user preferences. The experiment results on a real-world click-through data are reported in section 5. We then discuss the results in Section 6 and give the conclusion in Section 7.

2. RELATED WORK

The cascade model [6] is a classic click model which makes the following assumption: users examine the result documents in the ranking order from top to bottom, and the session will be terminated once a document is clicked. It actually makes a very strong assumption that all users are equally patient in finding relevant documents, and it can not model the sessions with more than one click. The dependent click model [11] extends the one-click assumption of the cascade model by allowing users to continue examining the next document with a fixed probability λ after a document is clicked. The click chain model [10] further improves the dependent click model. Instead of letting λ being a fixed value, the probability is set to be related to the relevance of the clicked document in the click chain model. The user browsing model [8] is another popular click model for its clear representation, easy computation and good performance. It assumes that the probability of a document being examined depends both on its ranking position and the distance to the last clicked position, which is different from the previous models. Therefore, the user browsing model introduces 55 examination parameters in total.

Each of the abovementioned models imposes strong hypothesis on users' examination and click behavior. Although the results showed that such models can capture user clicks to certain extent, we observe that all these models have the same examination/click parameters for different users. In practice, it has been observed that users behave very differently in previous studies [19, 20]. For example, it is suggested by White et al. [20] that search users have dramatic differences in the interaction with the search engine. Two extreme user classes are reported in their paper. The 'Navigators' have consistent interaction patterns. They appear to tackle problems sequentially and are more likely to revisit the domains. The 'Explorers' have variable interaction patterns. They tend to branch frequently in search tasks and visit many new domains. According to their findings, the behavioral level user differences will certainly affect the probability of a user examining or clicking a document.

Some recent studies tried to incorporate some user factors into click models. Hu et al. [12] considered the possible different search intents, which would influence the clicks. It assumes a bias between the users' search intent and the returned results in each search session. Search intent is taken into account as a hidden variable in their model, and the click probability of a document then depends on both the document relevance and the search intent of the user. The experiments showed that the model can do better in interpreting clicks. However, search intent is a factor that is still difficult to recognize and capture in advance. As a consequence, the proposed model has a limited ability to make click prediction for an incoming search session. Shen et al. [18] noted the influence of the interested topics of individual search users and proposed a framework for personalized click models from the view of collaborative filtering. They argue that the global query-document relevance is not sufficient to reflect the interest of an individual user to a document. That is, one will be interested in a document when the underlying topics of the document match her/his own interested topics. In their paper, matrix factorization was used to characterize the latent factors of queries, documents and users. The latent factors of a user indicate the potential aspects of her/his interest. Thus for a query-document pair, their method will generate personalized probabilities of user being interested. Their experiment results showed that using the latent factors of queries and documents alone can significantly improve the accuracy of the baseline click model (user browsing model), and incorporating user latent factors can further boost the improvement. Besides Web search, click models are also applied in sponsored search to optimize the CTR of ads [17]. Cheng et al. [5] investigated the role of user personality in sponsored search and proposed two groups of user specific features. One group contains demographic features, such as gender, age, marriage status, interest and job title; the other group contains user click features, such as the CTRs at different levels. After adding these features to a baseline non-personalized click model, an improved accuracy of click prediction is obtained. This work shows the effectiveness of the personalized click model in sponsored search.

All the above approaches incorporated some user-specific features to improve click models. However, these models do not capture the user personality in a general behavioral level, i.e. a user has consistent behavioral habits across all her/his searches, which is content independent. In this paper, we will focus on studying the difference of users in the general behavioral level and build user-specific click models with this type of user preferences incorporated.

3. USER PREFERENCES

From a general behavioral aspect, search users may have different behavioral habits due to their personalities. For example, when searching a query, some users are willing to click many documents, whereas some others are reluctant to click the documents beyond top-ranking positions. Some users browse the search result page from top to bottom and make careful click selections, while some others simply trust the search engine to provide the most relevant documents in the top-ranking positions and will easily lose patience af-

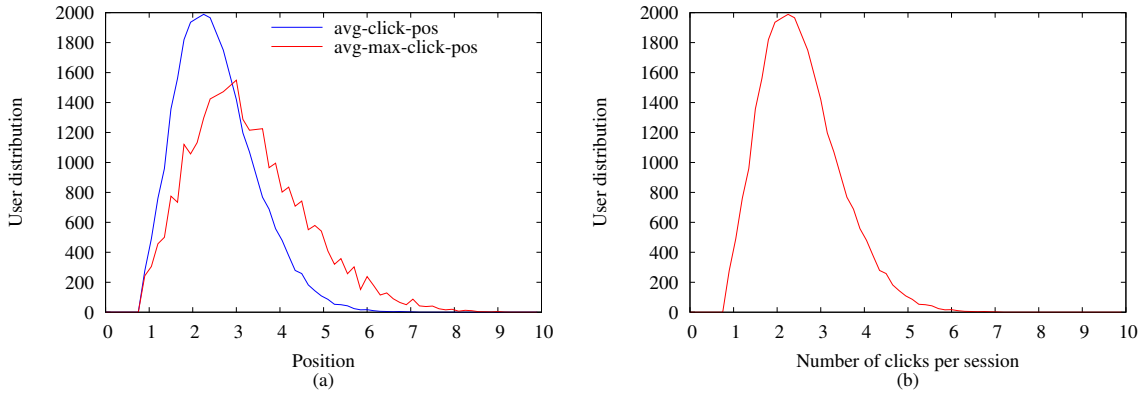

Figure 1: Results of the click log analysis. (a) shows user distributions on average click position and average maximum click position; (b) shows the user distribution on number of clicks per session. From both results we observe the diversity in users' click behavior.

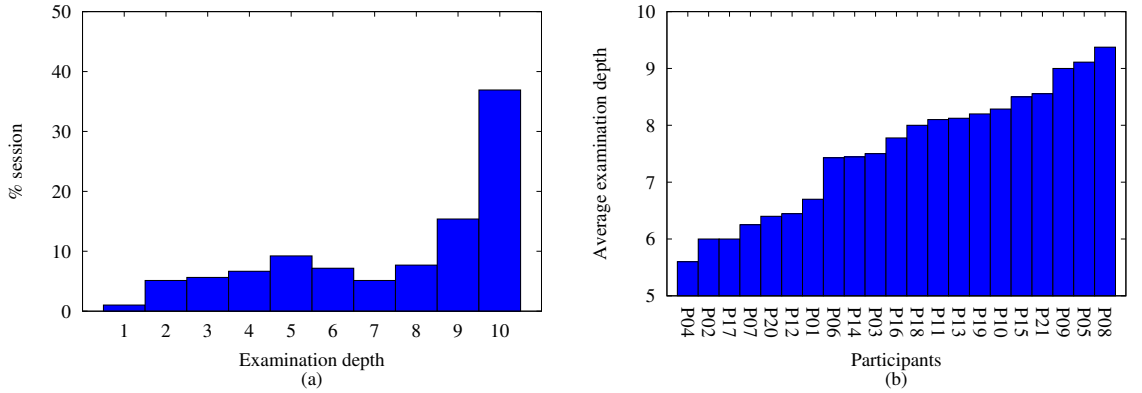

Figure 2: Results of the eye-tracking experiment. (a) shows the session distribution on examination depth, where sessions are dispersively distributed . (b) shows the average examination depth for each of the 21 participants over all her/his sessions, which differs a lot across participants

ter having examined several irrelevant documents. These personal preferences will affect the result that a click model captures user clicks. To verify the diversity of user behavior, in this section we study users' examining behavior by carrying out an eye-tracking experiment on 21 human subjects, and study users' clicking behavior by analyzing a click log of a real-world search engine.

3.1 Click log analysis

We first investigate how users click the documents in search result pages by analyzing the click log of a search engine. For each user, we compute: (i) the number of clicks per session, (ii) the average click position and (iii) the average maximum click position. In the click log, user is identified by cookie ID, which lasts a long enough time unless the user clears the cookie manually in the browser. A *session* is defined as the sum of the activities of a user searching one query in a continuous period of time, with a timeout of 30 minutes. To avoid noise, we only count the users who have no fewer than ten sessions and the result is presented in Figure 1.

The blue curve in Figure 1(a) shows the user distribution in terms of average click position. The majority of users (43.0%) have this value between 2 and 3. However, 27% of users are below 2 and the remaining 30% are above 3. The red curve in Figure 1(a) shows the user distribution in terms of average maximum click position, which is the last clicked position averaged across session. And we observe that over a half (54%) of the users have this value over 3. These dispersed distributions show user diversity in the clicking behavior. It may also imply the diversity of users in examination depth, which we will further verify in the following eye-tracking study. Figure 1(b) shows the user distribution over average number of clicks per session. Although the majority of the users have fewer than two clicks per session, there are still quite a lot of users (29%) who conducted more than two clicks per session. Besides the reason that the users may have examined different number of documents, another explanation for this diversity is that some users have stronger intent to click the returned documents than the others, which can be interpreted as click preference.

3.2 Eye-tracking Study

Click logs can not record the full information of users' examination. Although some simple rules can be used to infer whether a document is examined or not (e.g. the clicked document must be examined and the documents before a clicked document are supposed to be examined in most cases), we

Figure 3: The deepest examination positions of the 21 participants on one of the queries. Each triangle represents a participant. The size of a triangle indicates the fixation time. The fixation time threshold is 300 milliseconds for a position to be judged as 'examined'

still can not decide whether the documents after a clicked document are examined. Thus it is difficult to determine the exact depth that a user has examined in a search session using click logs alone. Instead, we carry out an eye-tracking experiment to obtain the explicit examination data of search users in a laboratory environment. We recruited 21 participants, who are college students from different disciplines. For the experiment, ten queries are randomly selected from the search engine query log as search tasks. Each participant is required to search all the ten queries using a real Web search engine on a computer with eye-tracking device installed.

The eye-tracking device detects the pupil of the subject and the software computes the corresponding fixation point on the computer screen. When a subject searches a query, the trail of her/his eye movements is then recorded. With the scan trail of each session (i.e. a subject searching a query), the examination depth can be obtained. Here examination depth is defined as the rank of the most bottom document whose snippet is examined by the subject. To determine whether a document is examined or not, we test if there is any point with a fixation time larger than 300 milliseconds in the corresponding area of the document. Figure 2(a) shows how the sessions are distributed on examination depth. The distribution is quite dispersed, showing the diversity of examination depth across all sessions. We notice that position 10 attracted the most sessions, which may not be the case in real search scenarios. One of the reasons is that the queries we use here are randomly sampled from the search engine query log, so most of them are not very frequent queries for which relevance documents can be easily found at early positions. For most our queries, the subjects will have to explore more documents to meet the information need. Although the participants are given the instruction that they should search the queries as normal, the laboratory environment may still be an factor that makes the participants behave differently from the real-world scenarios. However, as all the participants are in the same settings, it is still feasible to investigate the differences among them, which is the main purpose of this experiment.

To investigate the difference among the participants, we calculate the average examination depth of each participant over the ten queries. As demonstrated in Figure 2(b), the participants are diversely distributed over average examination depth. Participant P04 has the lowest average examination depth 5.6 while the number for P08 is 9.4, which indicates that the participant almost examined all the ten documents for each query. This fact is a strong indicator of the user diversity in examining behavior. More concretely, Figure 3 shows the examination depth of the participants on one of the queries, in which each triangle represents a participant and the position corresponds to the deepest examined position. This image gives us a direct impression on how different the search users are when examining a search result page.

3.3 Preferences

According to the results of click log analysis and eye-tracking study, the examining behavior and the clicking behavior exhibit considerable diversity across users. As suggested by the analysis above, the behavioral level user personality is one of the causes for this diversity. Therefore, we propose two user preference factors to describe the behavioral level personality of a user in Web search scenario:

- **Examination preference**: The probability of a document being examined differs across users. Some users are likely to examine more documents than the other. There can be multiple affecting factors for this preference such as user's patience, familiarity using the search engine and etc. Here we use a single factor to represent the overall influence of all possible factors, which we call as the examination preference. This preference is assumed to be consistent within all search activities of a user. Users with high examination preference are likely to examine more documents.

- **Click preference**: Given a document being examined, the click probability depends not only on the document relevance, but also on user's judgment whether to click the document. The perceived relevance of a document differs across users, meanwhile users may also have personal behavioral habits in clicking. For example, some users have strict standard to judge a document as relevant and some users are used to click as many as documents without seriously judging the relevance in advance and then check the landing pages one by one. Therefore, the click probability differs. Here we use a factor named click preference to describe user's personality in clicking. Users with high examination are assumed to be more likely to click a document after examining it.

In the following of this paper, we will focus on studying the influence of these two preference factors to click models.

In the next Section, we build user-specific click models by incorporating them into several click models.

4. CLICK MODELS

As we have found in the previous section that the personal preferences indeed exist in users' search behaviors, better performance can be expected for click models by taking into account the user preference factors. In this section, we build a variety of user-specific click models by incorporating the examination preference and the click preference into some existing click models and a novel click model we propose in this paper.

4.1 User Browsing Model

The user browsing model [9] is an effective and efficient model proposed by Dupret et al. Unlike the click models which assume that the examination probability of a document depends on its ranking position only, the user browsing model takes into account additional information when modeling the examination probability: the distance of the current document to the last clicked document in the search result page. This assumption increases the number of examination parameters from the typical 10 to 55 (we only consider the top ten returned documents in the first search result page), which brings in more flexibility to the model. Following the classic examination hypothesis, a document gets clicked only when it is both examined and relevant. Two binary variables E and A are associated with the examination and the perceived relevance respectively in this model, and they both follow bernoulli distribution:

$$\begin{aligned} P(A|u,q) &= \alpha_{u,q}^{A}(1-\alpha_{u,q})^{1-A} \\ P(E|r,d) &= \beta_{r,d}^{E}(1-\beta_{r,d})^{1-E} \end{aligned} \tag{1}$$

in which q is the query; u is the result document; r is the ranking position of u and d is the distance to the last clicked document. Let C be the variable indicating whether the document is clicked. The joint probability of the variables is written as:

$$P(C,A,E|u,q,r,d) = P(C|A,E)P(A|u,q)P(E|r,d) \tag{2}$$

where probability $P(C|A,E)$ is deterministic according to the examination hypothesis. A and E are independent from each other so their joint probability can be written as the product of two separate parts. In this calculation, user related factors are not considered. Therefore, it will derive the same click probability for different users. As indicated by our assumption above, both E and A should be user specific in our setting of user preferences. A straight forward way to introduce the user preference factors is to make E and A bernoulli variables with new parameters. That makes $A_{u,q,p} \sim$bernoulli$(\alpha_{u,q}\epsilon_p)$, and $E_{r,d,p} \sim$bernoulli$(\beta_{r,d}\gamma_p)$. p denotes the specific user. ϵ_p is the parameter for examination preference of user p and γ_p is the parameter for click preference. This transformation seems convenient before we later find out that it has no closed form solution in optimization with the expectation-maximization algorithm (also known as the EM algorithm), which is used for parameter estimation in the original user browsing model. Although gradient descent algorithms can be used instead, we still want to maintain the availability of the EM algorithm for its efficiency and elegant form after introducing the user preference factors. Sticking to EM for the new model also makes the following comparison between the two models fair. Therefore, instead of making new parameters for E and A, we directly add two new variables into the model, that are H and I. H indicates whether the user's examination preference is met and I indicates whether the click preference is met. We assume they follow bernoulli distribution as well:

$$\begin{aligned} P(H|p) &= \epsilon_p^H(1-\epsilon_p)^{1-H} \\ P(I|p) &= \gamma_p^I(1-\gamma_p)^{1-I} \end{aligned} \tag{3}$$

and the joint probability becomes:

$$\begin{aligned} & P(C,A,E,H,I|u,q,r,d,p) \\ = \; & P(C|A,E,H,I)P(A|u,q)P(E|r,d)P(H|p)P(I|p) \end{aligned} \tag{4}$$

given that a click only happens when H,I,A,E all take the value 1, the click probability becomes:

$$P(C=1|u,q,r,d,p) = \beta_{r,d}\epsilon_p\alpha_{u,q}\gamma_p \tag{5}$$

The form of this joint probability makes EM algorithm available for optimization. We note that it might be more flexible to let H and I be dependent on specific ranking position and etc. However, as we have stated in the previous section, we simply let H and I be consistent for a user within all her/his search sessions to avoid introducing too many parameters that will probably cause serious overfitting problems.

4.2 Position model

In the position model [6], the examination probability of a document is assumed to be related only to its ranking position, regardless of the previous clicks. Compared to the user browsing model, it has only ten examination parameters which is actually a stronger examination assumption. In this model, the click probability of a document is calculated as:

$$P(C=1|u,q,r) = \alpha_{u,q}\beta_r \tag{6}$$

and we incorporate the user preferences in the same way we did in the user browsing model. Variables H and I are introduced and the click probability becomes:

$$P(C=1|u,q,r,p) = \alpha_{u,q}\beta_r\epsilon_p\gamma_p \tag{7}$$

With only $P(E)$ being different from the user browsing model, the EM algorithm is also available for efficient parameter estimation in the position model.

4.3 Logistic model

The logistic model is proposed along with the user browsing model in [9]. It also follows the examination hypothesis that the click probability is the product of two separate probabilities, except that each of the two probabilities is calculated by a logistic function. In the logistic model, the click probability is rewritten as:

$$P(C=1|u,q,r,d) = \sigma(\beta_{r,d})\sigma(\alpha_{u,q}) \tag{8}$$

where σ is the logistic function:

$$\sigma(x) = \frac{1}{1+\exp(-x)} \tag{9}$$

This model is attractive because the logistic function always outputs a value ranged between 0 and 1, which has a perfect probability meaning. In such case the linear constrains for $0 \le \beta_{r,d} \le 1$ and $0 \le \alpha_{u,q} \le 1$ in optimization are removed. Without the constrains, algorithms such as

gradient descent can be easily adopted to estimate the parameters for this model. On the other hand, the logistic function also has a great advantage to plug in new features in a much easier way compared to the models in probabilistic frameworks, such as the user browsing model and the position model. To incorporate user preferences into the logistic model, we can simply rewrite the logistic functions with the new parameters:

$$P(C=1|u,q,r,d,p) = \sigma(\beta_{r,d}+\epsilon_p)\sigma(\alpha_{u,q}+\gamma_p) \quad (10)$$

Now ϵ_p and γ_p will have an influence to the probability along with $\beta_{r,d}$ and $\alpha_{u,q}$. The parameters can be estimated by maximizing the likelihood of the data set using the gradient descent algorithm.

Note that the examination parameter we use here is $\beta_{r,d}$, which follows the examination assumption of the user browsing model. Similarly, we can also use the examination assumption of the position model in this logistic model, for which the click probability becomes:

$$P(C=1|u,q,r,p) = \sigma(\beta_r+\epsilon_p)\sigma(\alpha_{u,q}+\gamma_p) \quad (11)$$

Equation 10 and 11 lead to two logistic models with different examination assumptions. We implement both of them in the following experiments.

4.4 Cascade model

The cascade model [6] assumes that users examine the documents one by one from top to bottom in the search result page until a document is clicked. It has shown success in explaining clicks at early ranking positions. In this model, the probability of the i_{th} document to be clicked is:

$$\begin{aligned} P(C_i=1|E_i=1) &= r_i \\ P(C_i=1) &= r_i\prod_{j=1}^{i-1}(1-r_j) \end{aligned} \quad (12)$$

in which r_i is the probability of the i_{th} document being relevant. By maximizing the likelihood of all session observations of a query, r_i can be efficiently estimated in a closed form calculation

$$r_i = \frac{n_i}{n_i+m_i} \quad (13)$$

where n_i is the number of times the document is clicked and m_i represents the number of times the document is skipped before a click. The solution is quite straightforward and very convenient to compute. Note that as the assumption that the users will examine all documents before a click is solidly embedded in this cascade model, the examination preference is not valid here. Therefore, we only incorporate the click preference into the cascade model and the click probability becomes:

$$P(C_i=1|q,p) = r_i\gamma_p\prod_{j=1}^{i-1}(1-r_j\gamma_p) \quad (14)$$

Now the probability of the i_{th} document to be clicked after examined by user p is dependent on both its relevance and the click preference of p. After this transformation, the EM algorithm is still available for optimization.

4.5 Dilution model

So far we have been using the 'global + personal' way to incorporate user preference factors into a variety of existing click models. On the other hand, instead of using this pattern, we can make all parameters user-specific as well, i.e. no global parameters. This approach is more straightforward but often suffers from over-fitting problems if too many personalized parameters are introduced. For example, in the user browsing model there are 55 examination parameters in total. If we make all these parameters user-specific to reflect personal examination preferences, we will have 54 times more examination parameters to estimate in the new model. In such case, there is no doubt that heavy over-fitting problems will rise up. However, if the number of parameters in the original model is small enough, it may be possible to assign each user a set of personal parameters. To compare with the previous method of incorporating user preferences, we propose a new click model which has only three parameters for the examination assumption.

Generally, more parameters bring more flexibility to the model, leading to a stronger fitting capacity. However, if there are some patterns embedded in the parameters (i.e. the parameters are correlated and thus redundant somehow), we may use fewer parameters with approximately the same representating ability. It has been found in the estimated result of the user browsing model that the examination probability decays when: 1) the ranking position increases; 2) the distance to the last click increases. Thus in this new model, we set two damping factors for the two cases respectively. Furthermore, to capture the additional information on how many documents have been clicked before, which is not considered in the user browsing model, we add the third damping factor: the examination probability decays as the number of documents clicked before increases (i.e with more documents clicked before, the user's information need is more likely to have been met). We call this model the dilution model.

Given that k documents are clicked before the r_{th} ranking position and the distance from r to the last clicked position is d, the examination probability of the r_{th} document in this dilution model is calcuated as:

$$P(E=1|r,d,k,p) = \beta_p^{r-1}\lambda_p^k\mu_p^d \quad (15)$$

in which β_p is the damping factor of p for the ranking position; λ_p is the damping factor for the clicks before; and μ_p is the damping factor for the distance to the last click. Each damping factor is a real valued parameter ranging from 0 to 1, which needs to be estimated. In the model, we assume that the first document in a search result page always has an examination probability of 1. And d is set to 0 if there are no clicks before.

Compared to the 'global + personal' way of incorporating user preferences in the previous click models, this dilution model allows each user to have personal examination parameters that are not shared by the others. It makes the model more descriptive and reasonable for a specific user. By knowing the damping factors of a user, we can have a direct sense of how likely she/he is going to examine a certain document. As to the click preference, we use the same method as in the previous models because there are too many query-document pairs and we can not make all the relevance parameters user-specific. After these modifications, the click probability turns out to be:

$$P(C=1|u,q,r,d,k,p) = \beta_p^{r-1}\lambda_p^k\mu_p^d\alpha_{uq}\gamma_p \quad (16)$$

in which γ_p is the click preference of user p and the parameter set $\{\beta_p,\lambda_p,\mu_p,\gamma_p\}$ describes the preferences of p. All the parameters are valued between 0 and 1.

5. EXPERIMENTS

5.1 Experimental setting

We use the click logs of a real-world Chinese search engine during a period of one month (November 2011) for experiment. The data is sampled by users to control the total size, i.e we select a random subset of users and use all of their click data in that month for the following experiments. To limit the noise in the data set, all the queries with fewer than 10 sessions are removed because there are not enough click data for the estimation of document relevance for these queries. Here a *session* is defined as a unique user-query pair in a continuous time period (with 30 minutes timeout). Thus query reformulations are treated as different sessions. In our data, user is identified by cookie ID. For the protection of users' privacy, all sensitive attributes, such as query string and document URL, are processed into numbers. As the users may clean the cookie from time to time, there are a big portion of users who have few search sessions in the month. To guarantee that there is enough data for each user when building the user-specific click models, we also remove the users that issued fewer than 10 distinct queries from the data. After the filtering, the final data set has 10,012 unique users, 53,048 unique queries and 668,105 sessions. We then split the data set into training part and test part in the following way: for each user, the first 80% of her/his sessions, ordered by timestamp, are used for training and the remaining 20% are used for testing. For each session, only the top ten returned documents (in the first search result page) are used for modeling.

In our experiments, we train the following models: user browsing model (UBM), position model (POS), logistic model (LOG-r and LOG-rd, using different examination assumptions), cascade model (CAS), dilution model (DIL) and their user-specific versions with user preferences incorporated (with postfix '-user' in the name). And a comparative study is carried out using multiple evaluation measures.

5.2 Perplexity

After learning the parameters of each model, we first evaluate their performances using *perplexity* as a metric. Perplexity is equivalent to cross-entropy but with an easier interpretation. It is originally an evaluation metric for language models, and is widely used in the evaluation of click models. It measures how well a trained model fits the real data, which is calculated as follows:

$$perplexity = 2^{-\frac{1}{N}\sum_i^N \log_2 p_i} \tag{17}$$

where N is the number of observations and p_i is the probability of the observation i being correctly predicted. Perplexity has a perfect value of 1 when the model is able to predict each single observation correctly. The lower the perplexity value, the better the model. When evaluating click models with perplexity, an event of click or skip is considered as an observation. For a click observation, p_i is the predicted probability of the document to be clicked. For a skip observation, p_i is the predicted probability of the document not to be clicked. We calculate the perplexity for click observations, skip observations and all observations separately and the results are shown in Table 1. To avoid zero values in equation 17, we let the predicted probability of an observation have a minimum value of 0.001 and a maximum value of 0.999. Also, the query-document pairs that appeared in fewer than five observations are excluded from the evaluation to limit the noise in the result.

It is observed that after incorporating user preferences, perplexity on the test set is consistently improved for all click models. Some of them gain notable improvements. LOG-rd-user gains as much as 13.7%[1] improvement over LOG-rd, followed by DIL-user(11.3%) and CAS-user(7.1%). Among the probabilistic models, only the cascade model gains significant improvement. The improvements for the user browsing model and the position model are however limited. The reason is that they both fail in improving perplexity for the skip observations. On the other hand, the logistic function based models benefit more from the integration of user preferences. This fact indicates that the current method we use to integrate user preferences is generally helpful for click models but is still not optimized.

It is expected that the models with distance information (distance to the last click) considered will perform better than the models without distance information. The former have richer information and more examination parameters, so they are supposed to be more powerful and flexible. In our experiment results, UBM is 12.5% better than POS; LOG-rd is 9.5% better than LOG-r. After incorporating user preferences, UBM-user and LOG-rd-user still beat POS-user and LOG-r-user respectively as expected. We also observe that the models with distance information considered benefit more from the integration of user preferences than the models without distance information, it is reasonable because a model with larger flexibility has larger room for improvement as well.

Among all the models presented in Table 1, the best performing model is LOG-rd-user. Given the fact that LOG-rd is not the best performing one among the models without user preferences. The significance of taking into account user preferences in a click model is revealed.

5.3 Click Prediction

Perplexity is an appropriate measure for the fitting ability of a click model, but using one metric alone is always not convincing enough in evaluation. In this section we evaluate the click models by predicting real clicks. The aim is to investigate how the clicks predicted by different models match the real data. The click models are used to predict the first clicked position and the last clicked position [11]. The gap between the predicted position and the real position is able to reflect the effectiveness of a model to certain extent.

Given a click model and the trained parameters, we simulate the user clicks for each session presented in the test set. Then the first click and the last click are identified from the simulated clicked. The simulated click positions are compared to the ground truth to compute the mean absolute error (MAE). For the click models without distance information in examination assumption, such as POS and LOG-r, clicks are simulated for each ranking position independently because the click probability of a document is independent from the previous clicks. But for the click models with distance information in the assumption, such as UBM and LOG-rd, the click probability of a document is dependent on the previous clicks so we have to simulate the clicks in order from top to bottom. In such case, if a click is wrong-

[1]as the ideal value for perplexity is 1, the improvement of perplexity p_1 over p_0 is calculated as: improvement=$(p_0 - p_1)/(p_0 - 1) \times 100\%$

Table 1: The perplexity on training set and test set. *click* indicates the click observations; *skip* indicates the skip observations; *total* is the sum of *click* and *skip*. The numbers under them are the amount of observations in each category. The numbers in brackets are the perplexity improvement of the models that have user preferences incorporated(with postfix '-user') compared to the original models

	Training set			Test set		
	total	click	skip	total	click	skip
complete data	4,704,953	517,772	418,7181	1,019,055	112,549	906,506
UBM	1.084	1.508	1.041	1.119	1.824	1.053
UBM-user	1.083	1.504	1.040	1.117(+1.7%)	1.793(+3.8%)	1.053(0%)
POS	1.100	1.648	1.046	1.136	2.014	1.058
POS-user	1.100	1.661	1.045	1.135(+0.7%)	1.989(+2.5%)	1.059(-1.7%)
LOG-rd	1.105	1.629	1.053	1.124	1.752	1.064
LOG-rd-user	1.095	1.612	1.044	1.107(+13.7%)	1.679(+9.7%)	1.052(+8.2%)
LOG-r	1.118	1.844	1.051	1.137	1.993	1.061
LOG-r-user	1.102	1.698	1.045	1.130(+5.1%)	1.949(+4.4%)	1.056(+8.2%)
DIL	1.113	1.165	1.060	1.124	1.709	1.067
DIL-user	1.090	1.466	1.051	1.110(+11.3%)	1.585(+17.5%)	1.062(+7.5%)
	total	click	skip	total	click	skip
cascade data	575,814	456,490	119,324	126,569	97,875	28,694
CAS	1.38	1.178	2.527	1.535	1.26	3.016
CAS-user	1.376	1.178	2.494	1.497(+7.1%)	1.239(+8.1%)	2.853(+8.1%)

ly predicted at a certain position, the following predictions will be affected, which will cause more uncertainty in the prediction result. During the simulation, there are chances that a session may have no simulated clicks. We only keep the sessions with at least one simulated click for evaluation.

Figure 4 shows the prediction errors of the first clicked position and the last clicked position. The cascade model only predicts one click each session so it is evaluated by the first click prediction only. According to the results, predicting the last clicked position turns out to be a more difficult task than predicting the first clicked position, indicated by the larger error bars in (b) than that in (a). In both (a) and (b), we observe that the errors are smaller for the models with user preferences, and this result is consistent for all the models. The trend that all click models get improved is consistent with the perplexity result as well. Besides, the performance gain ratio of the models is also similar to that in the perplexity result, i.e. LOG-RD has the largest performance gain, followed by DIL and LOG-r. And the variance of error among different models is also reduced. For the first clicked position, the variance of error is reduced from 0.019 to 0.006; and for the last clicked position, the variance of error is reduced from 0.015 to 0.001. In other words, the performance difference of the click models in this click prediction task is smoothed after incorporating user preferences.

On the other side, the performance ranking of the click models in this task is different from the ranking in perplexity. POS and LOG-r are among the worst by perplexity but have relatively low errors in click prediction. This inconsistency between the two evaluation methods simply indicates that different measures may lead to different evaluation results. However, it does not change the fact that all the reported models benefit from incorporating user preferences on both evaluation measures.

With the simulated clicks, we also draw the distribution of the first clicked position and the last clicked position in Figure 5. We use the top two models that have the largest performance gain for demonstration, along with the empirical ground truth. We observe that neither LOG-rd nor DIL matches the empirical distribution very well. But after incorporating user preferences, LOG-rd-user and DIL-user are significantly better matching the real distribution. For the click models that are not shown in Figure 5, small but not very significant improvements can be observed as well. This result further verifies the effectiveness of user preferences in improving a click model.

6. DISCUSSION

The idea of user preference was initially motivated by our observations in the eye-tracking study and click log analysis. Our experiment results showed that the user preferences were consistently helpful to improve the click model performance on multiple evaluation metrics. A good explanation is that the model assumptions with user preferences can better reflect the real situation.

The two evaluation measures used in this paper, i.e. the perplexity and the error of the predicted click position, produced different performance ranking of the models. For example, LOG-rd-user was the best performing model in perplexity but failed to lead in the click prediction task. Therefore, it seems tricky which evaluation measure one chooses when comparing two click models. However, this problem is not the main focus of this paper because our purpose is not to build a model that is better than all the other models on all evaluation metrics. Instead, we aim to provide a general approach that can be applied to most click models in order to consistently improve their performance on different metrics. For a variety of click models, we have shown consistent improvement by incorporating user preferences, using two different evaluation methods. This result supports our claim that the user preferences proposed in this paper are helpful in improving the click models.

We also note that the significance of the improvement differs across model. One possible reason might be the way how user preferences are integrated. The logistic function based models gained significant improvement from the inte-

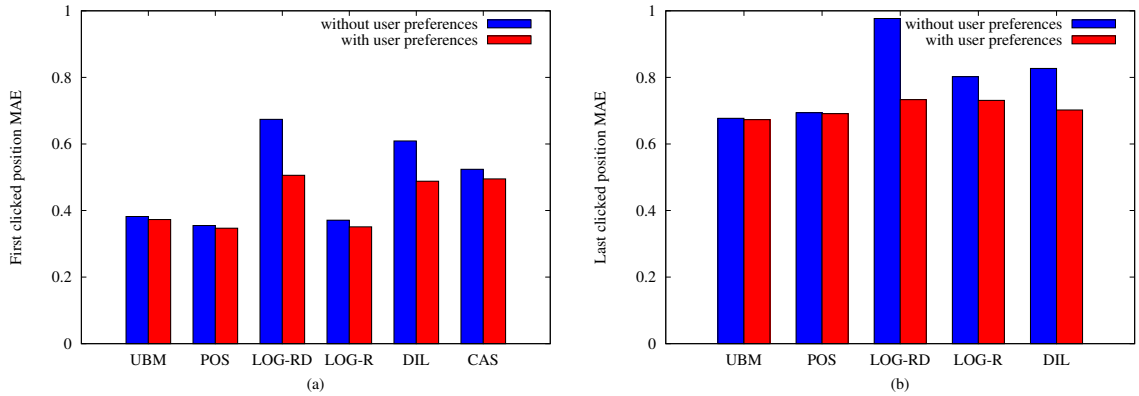

Figure 4: Mean absolute errors in predicting the first clicked position(a) and the last clicked position(b). In both results, the click models with user preferences incorporated have smaller errors (indicated by the red bars). The error variance among the models is also reduced after user preferences are incorporated

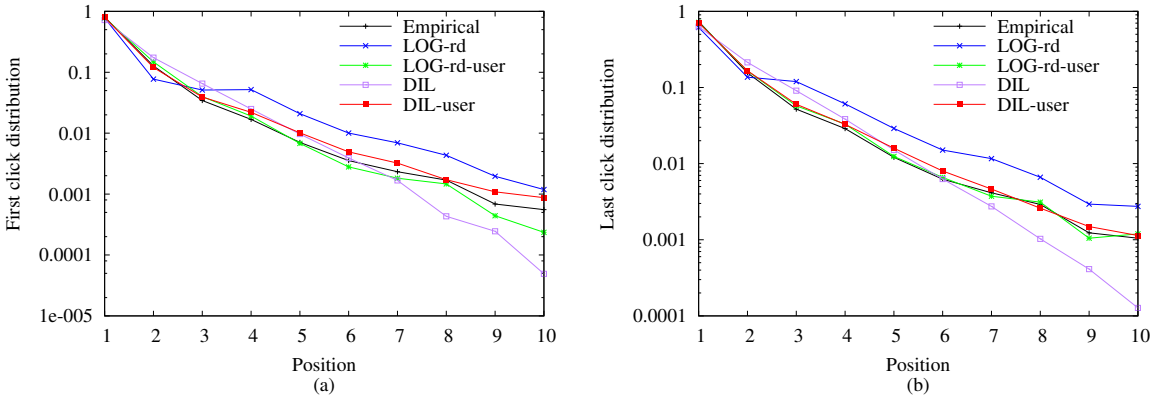

Figure 5: The distribution of the first click(a) and last click(b) predicted by click models. *Empirical* indicates the ground truth from the real click data. LOG-rd-user and DIL-user match the ground truth very well, while LOG-rd and DIL have relatively large gaps to the ground truth.

gration of user preferences, while the user browsing model and the position model gained less improvement. It may indicate that the integration method we used for the probabilistic models needs to be optimized. Since we have verified the effectiveness of incorporating user preferences, optimizing a specific click model would be an interesting direction for future work.

It is interesting to find that the dilution model (DIL), which has only three examination parameters in total, obtained comparable performance with the other models. With respect to perplexity, the dilution model is the second best (next to the user browsing model) among all the models without user preferences. It beat the models that have more examination parameters (POS and LOG-r both have ten examination parameters) and tied LOG-rd which has 55 examination parameters. With user preferences incorporated, it maintains the good performance. Having the most user-specific parameters in total, the dilution model did not suffer from over-fitting problems in our experiment. The reason is that we have filtered out users with few query sessions in our data set, combined with that the total number of user parameters are still much less than the number of relevance parameters. But in a real-world application scenario, we usually can not expect to have sufficient data for all users. In such case, the model can be further improved by using a hybrid mechanism: use global parameters for the users without sufficient data and learn personal parameters when enough data is collected.

7. CONCLUSION

Search users have considerable differences in search behavior. However, existing click models do not take this diversity into consideration and they usually assume that all users have the same examination and click preferences. Intuitively, this does not seem to be the case. In this paper, we carried out analyses on real click-through data, which confirmed that users have different click preferences. With an eye-tracking experiment on 21 human subjects, we also observed that the users' examination behavior differs a lot. Motivated by these observations, we proposed two user preference factors, namely the examination preference and the click preference, and incorporated them into multiple click models. Our proposed approach is general enough to be adopted in a variety of click models. In the experiments, we showed that by incorporating the proposed user preferences, the performance of the existing models and our proposed model are consistently improved on multiple evalua-

tion measures. The series of experiments confirmed that the two additional user-dependent elements can capture some common factors underlying the examination and click behavior of search users, which is thus a step further toward a better understanding of user behavior facing a search result.

8. ACKNOWLEDGEMENTS

This work was supported by Natural Science Foundation (61073071), National High Technology Research and Development (863) Program (2011AA01A205) of China. Part of the work has been done at the Tsinghua-NUS NExT Search Centre, which is supported by the Singapore National Research Foundation & Interactive Digital Media R&D Program Office, MDA under research grant (WBS:R-252-300-001-490).

9. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR* , pages 19–26. 2006.

[2] S. Bhavnani. Domain-specific search strategies for the effective retrieval of healthcare and shopping information. In *CHI'02 extended abstracts on Human factors in computing systems*, pages 610–611. 2002.

[3] S. Bochkanov and V. Bystritsky. Alglib (http://www.alglib.net).

[4] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th international conference on World wide web*, pages 1–10. ACM, 2009.

[5] H. Cheng and E. Cantú-Paz. Personalized click prediction in sponsored search. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 351–360, 2010.

[6] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the international conference on Web search and web data mining*, pages 87–94. ACM, 2008.

[7] G. Duggan and S. Payne. Knowledge in the head and on the web: using topic expertise to aid search. In *Proceedings of SIGCHI 08'*, pages 39–48. 2008.

[8] G. Dupret, V. Murdock, and B. Piwowarski. Web search engine evaluation using clickthrough data and a user model. In *WWW2007 workshop Query Log Analysis: Social and Technological Challenges*, 2007.

[9] G. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st annual international ACM SIGIR*, pages 331–338. 2008.

[10] F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y. Wang, and C. Faloutsos. Click chain model in web search. In *Proceedings of the 18th international conference on World wide web*, pages 11–20. 2009.

[11] F. Guo, C. Liu, and Y. Wang. Efficient multiple-click models in web search. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 124–131. ACM, 2009.

[12] B. Hu, Y. Zhang, W. Chen, G. Wang, and Q. Yang. Characterizing search intent diversity into click models. In *Proceedings of WWW '11*, pages 17–26, ACM. 2011.

[13] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. 2002.

[14] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR*, pages 154–161. 2005.

[15] S. Kullback and R. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[16] F. Radlinski and T. Joachims. Active exploration for learning rankings from clickthrough data. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 570–579, New York, NY, USA, 2007.

[17] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pages 521–530. 2007.

[18] S. Shen, B. Hu, W. Chen, and Q. Yang. Personalized click model through collaborative filtering. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 323–332, New York, NY, USA, 2012. ACM.

[19] R. White and D. Morris. Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th annual international ACM SIGIR* , pages 255–262. ACM, 2007.

[20] R. W. White and S. M. Drucker. Investigating behavioral variability in web search. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 21–30, New York, NY, USA, 2007.

[21] R. W. White, S. T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of WSDM '09*, pages 132–141, 2009.

[22] X. Zhang, H. Anghelescu, and X. Yuan. Domain knowledge, search behavior, and search effectiveness of engineering and science students: An exploratory study. *Information Research*, 10(2):10–2, 2005.

[23] C. Wang, Y. Liu, M. Zhang, S. Ma, M. Zheng, J. Qian, and K. Zhang. Incorporating Vertical Results into Search Click Models. In *Proceedings of SIGIR '13*, pages 503-512

[24] D. Xu, Y. Liu, M. Zhang, S. Ma. Incorporating Revisiting Behaviors into Click Models. In *Proceedings of WSDM '12*, pages 303-312.