Incorporating Non-sequential Behavior into Click Models

Chao Wang[†], Yiqun Liu[†], Meng Wang[‡], Ke Zhou[⋆], Jian-yun Nie[#], Shaoping Ma[†]
[†]Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science & Technology, Tsinghua University, Beijing, China
[‡]School of Computer and Information, HeFei University of Technology
[⋆]Yahoo Labs, London, U.K.
[#]Université de Montréal
yiqunliu@tsinghua.edu.cn

ABSTRACT

Click-through information is considered as a valuable source of users' implicit relevance feedback. As user behavior is usually influenced by a number of factors such as position, presentation style and site reputation, researchers have proposed a variety of assumptions (i.e. click models) to generate a reasonable estimation of result relevance. The construction of click models usually follow some hypotheses. For example, most existing click models follow the *sequential examination hypothesis* in which users examine results from top to bottom in a linear fashion. While these click models have been successful, many recent studies showed that there is a large proportion of non-sequential browsing (both examination and click) behaviors in Web search, which the previous models fail to cope with. In this paper, we investigate the problem of properly incorporating non-sequential behavior into click models. We firstly carry out a laboratory eye-tracking study to analyze user's non-sequential examination behavior and then propose a novel click model named Partially Sequential Click Model (PSCM) that captures the practical behavior of users. We compare PSCM with a number of existing click models using two real-world search engine logs. Experimental results show that PSCM outperforms other click models in terms of both predicting click behavior (perplexity) and estimating result relevance (NDCG and user preference test). We also publicize the implementations of PSCM and related datasets for possible future comparison studies.

Categories and Subject Descriptors

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

Keywords

click model; non-sequential behavior; eye-tracking

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

DOI: http://dx.doi.org/10.1145/2766462.2767712.

1. INTRODUCTION

Modern search engines record user interactions and use them to improve search quality. In particular, user's click-through has been successfully used to improve click-through rate (CTR), Web search ranking, query recommendation and suggestion, and so on.

Although click-through logs can provide implicit feedback of users' click preferences [1], it is difficult to derive accurate absolute relevance judgments due to the existence of click noises and behavior biases. Joachims et al. [19] worked on extracting reliable implicit feedback from user behaviors, and concluded that click logs are informative yet biased. Previous studies showed that users' clicking behaviors are biased towards many aspects such as "position" [8, 19] (user's attention decreases from top to bottom), "trust" [32] (Web site reputations will affect user's judgment), "presentation" [27] (different search results' display styles will influence users' attention allocation strategies) and so on. To address these problems, researchers have proposed a number of click models to describe user's practical browsing behavior and to obtain an unbiased estimation of result relevance [4, 11, 13].

In [20], users' examination behavior patterns are grouped into two categories according to the findings in eye-tracking studies: the depth-first strategy and the breadth-first strategy. The depth-first model assumes that a user examines result list from top to bottom and decides whether to click immediately after examining a result. The breadth-first strategy, however, draws a different picture: a user will look ahead at a series of results before clicking on the favorite results among them. Since the position bias can be easily incorporated into click models with the depth-first assumption, most existing click models [4, 11, 13] follow this assumption and assume that the user examines search results in a top-to-bottom fashion.

However, recent eye-tracking experiments [22] showed that only 34 % of search users' scan paths are linear while over 50 % of sessions contain revisiting behaviors (i.e. given a search engine result page (SERP), the user first clicks the result at position i and then clicks the one at position $j, j \leq i$) or skipping behaviors. Table 1 shows the non-sequential click proportion of multi-click query sessions (user clicked two or more results on one SERP) from two commercial search engine logs. We can see that nearly one-third of multi-click sessions contain non-sequential click actions. While most existing click models are based on ranking positions rather than action sequences, the click sequence information is usually ignored and non-sequential

Table 1: Non-sequential click proportions of multi-click query sessions in two different search behavior data sets (See data sets' detail in Sec. 5.1.2).

Search Engine	Sogou	Yandex
Percentages	27.9%	30.4%

clicking behaviors are not considered either. [10, 14] already showed that the last click in a search session may be more reliable than other clicks. However, the last click performed by a user is not necessarily the one at the lowest position, but the last one in the sequence of clicks. It is thus necessary to take the temporal sequential information into account.

Some existing click models [28–30] have tried to do so. These models relax the restrictions of user's examination sequence (e.g. [28] assumes that examination sequence can be arbitrary) to increase model's descriptive power. However, most of these methods actually abandon the prior knowledge of user browsing preference generated from other user behavior studies, which has been found useful. In practice, these models cannot achieve as good performance as other popular click models according to our experimental results in Section 6.

To better understand user's click and examination sequences, we design a laboratory study to analyze users' practical examination patterns. Our observations confirm clearly that the click behavior of users are non-sequential. On the other hand, the examinations of documents between two clicks usually follow one direction, but with possible skips. This observation shows that some of the assumptions used in the previous position-based models (e.g. the sequential examination assumption) are reasonable in local contexts (i.e. between two clicks). It is thus possible to build a new model upon the existing position-based models by adding new hypotheses. By this means, we not only inherit the framework which has already proved to be effective, but also combine sequential information to better capture user's preference on different search results.

Our contributions in this paper are four-fold:

- An eye-tracking experiment is carried out to analyze user's non-sequential examination and click behavior on search engine result pages (SERPs).
- A novel click model named Partially Sequential Click Model (PSCM) is proposed to incorporate non-sequential behavior.
- We show experimentally that the proposed PSCM model outperforms the existing models on two real-world commercial search engine datasets (one of which is publicly available).
- In addition, we make our implementation of the proposed model available as an open-source project (see Section 5.1).

The paper is organized as follows. Various click models are reviewed in Sec. 2. In Sec. 3, we outline insights of eye-tracking study on non-sequential behavior. In Sec. 4, we formally introduce PSCM. We report the experiments on PSCM and compare it with existing click models in Sec. 5. Finally, conclusions and future work are discussed in Sec. 6.

2. RELATED WORK

In this section, we review a number of important click models and introduce some preliminary assumptions shared by these models and PSCM. As this paper mainly focuses on the influence of click sequence, we separate existing click models into two classes: models following the depth-first assumption (i.e. position-based click models) which cannot take non-sequential click actions into account (Sec. 2.1); other models which are able to consider non-sequential click actions (Sec. 2.2).

Note that most click models follow the examination hypothesis [8]: a document being clicked ($C_i = 1$) should satisfy ($\rightarrow$) two conditions: it is examined ($E_i = 1$) and it is relevant ($R_i = 1$) (most click models assume $P(R_i = 1) = r_u$, which is the probability of the perceived relevance), and these two conditions are independent of each other.

$$C_i = 1 \rightarrow E_i = 1, R_i = 1 \tag{1}$$

$$E_i = 0 \rightarrow C_i = 0 \tag{2}$$

$$R_i = 0 \rightarrow C_i = 0 \tag{3}$$

Following this assumption, the probability of a document being clicked is determined as follows:

$$P(C_i = 1) = P(E_i = 1)P(R_i = 1) \tag{4}$$

2.1 Position-based Click Models

As this class of click models does not take click sequence into account, the click action is simply mapped to each search result's ranking position. Based on the assumption that a user examines from top position to bottom position, this kind of click models naturally takes position bias into account.

Craswell et al. [8] proposed the cascade model, which assumes that while a user examines the results from top to bottom sequentially, he/she immediately decides whether to click on a result. The cascade model is mostly suitable for single-click sessions. A number of succeeding models were proposed to improve both its applicability and performance. Based on the cascade hypothesis, the Dependency Click Model (DCM) [13] extends the cascade model in order to model user interactions within multi-click sessions. DCM assumes that a user may have a certain probability of examining the next document after clicking the current document, and this probability is influenced by the ranking position of the result.

Subsequently, the User Browsing Model (UBM) [11] further refines the examination hypothesis by assuming that the event of a document being examined depends on both the preceding click position and the distance between the preceding click position and the current one.

$$P(E_i = 1|C_{1\ldots i-1}) = \lambda_{r_i,d_i} \tag{5}$$

where r_i represents the preceding click position and d_i is the distance between the current rank and r_i.

The Dynamic Bayesian Network model (DBN) [4] is the first model to consider presentation bias due to snippet (rather than ranking position). This model distinguishes the actual relevance from the perceived relevance, where the perceived relevance indicates the relevance represented by titles or snippets in SERPs and the actual relevance is the relevance of the landing page.

$$P(R_i = 1) = r_u \tag{6}$$

$$P(S_i = 1|C_i = 1) = s_u \tag{7}$$

$$P(E_{i+1}|E_i = 1, S_i = 0) = \lambda \tag{8}$$

where S_i represents whether the user is satisfied with the i-th document, s_u is the probability of this event, r_u is the probability of the perceived relevance, and λ represents the probability of continuing the examination process.

Although some of these models have achieved great success in interpreting clicks and in predicting relevance, compared to our proposed PSCM, they cannot explain the situation where a user does not follow top-down click sequence and they ignore revisiting or duplicated clicks.

2.2 Click-sequence-based Click Models

To the best of our knowledge, only a few studies [28–30] have tried to take non-sequential behavior into consideration. Xu et al. first proposed a Temporal Click Model (TCM) [30] to model user click behavior for sponsored search. This model can only handle two results/ads in an SERP. The only non-sequential click action in this model is: the user first clicks the second result and then goes back to click the first result. This makes it impossible to cope with the whole ranked result list like other click models.

Wang et al. introduced the partially observable Markov Model (POM) [28] to model arbitrary examination orders. The POM model treats the user examination events as a partially observable stochastic process. Although POM can model non-sequential behaviors, it only considers the examination transition at each position (i.e. different users and different queries share the same examination sequence parameters). Therefore, this model cannot predict the click probability or relevance for a specific query and thus can hardly be used in a practical search environment. Due to this limitation, it cannot be compared with other state-of-the-art click models such as UBM and DBN which need to predict click probability and relevance for a specific query-URL pair. It also makes the first order examination assumption that the current examination behavior only depends on its previous examination step, which might not align with the real user behavior.

Xu et al. proposed a Temporal Hidden Click Model (THCM) [29] to cope with non-sequential click actions. They focused on the revisiting behavior and assumed that after clicking a search result, user has a probability to go back to examine previous results (bottom-up). However, their model is also based on one-order Markov examination assumption and supposes that users examine results one by one in examination process, which does not necessarily correspond to practical user behavior (see Sec. 3).

While the above three click models have the potential to take click sequence information into consideration, compared to our proposed PSCM model, their adopted methodology are less suitable to deal with practical search behavior of modern commercial search engines.[1] In comparison, our PSCM is inspired by the eye-tracking study on real users' non-sequential SERP behavior and therefore corresponds better to real-world user behavior.

[1]We actually adapt TCM and POM in Sec. 5.1 to enable them making click and relevance predictions. They will be used as baselines to compare against our model. In Sec. 5, we empirically demonstrate that these models cannot achieve better performance than the popular position-based click models.

3. NON-SEQUENTIAL USER BEHAVIOR

To investigate user's examination sequence during the search process, we carried out a laboratory study with 37 undergraduate students recruited from a university in China (18 males and 19 females with various self-reported Web search expertise). The number of subjects is similar to other Web search eye-tracking studies such as [9, 12].

Subjects were provided with a list of 25 search tasks. Each task was accompanied by a fixed query (with an explanation of the information need to avoid ambiguity) and a Chinese commercial search engine's first result page. We crawled and stored the corresponding SERP to ensure that all subjects saw the same page for each query. With this setup, each search task (query session) corresponds to one specific SERP. The queries for the search tasks were sampled from the NTCIR IMine task[2]. As different types of information need [2] may also affect the browsing behavior [12], the selected search tasks cover different types of search intent. In the query set, 5 of the queries are "Navigational" (e.g. "Meizu's official website"), 10 are "Informational" (e.g. "What is the sound card") and 10 are "Transactional" (e.g. "Web browser download").

With an eye-tracking device (Tobii X2-30), we recorded each subject's eye movement information on each result for each search task. For quality control purposes, each subject was asked to make an eye-tracking calibration before the experiment. The precision threshold of calibration was less than 1° for both vertical and horizontal directions. Subjects may perform the calibration several times before they meet the precision requriement. Behavior data from several query sessions were removed due to subjects' operation errors or software crashes. After removing data from these sessions, we finally collected 890 (out of 925) valid query sessions. When we look at the click-through behavior in the sessions, we found that there exist many query sessions (22.8%, 203 of 890) that contain non-sequential (revisiting or duplicate) click actions. This number confirms clearly the necessity of incorporating non-sequential behaviors into click models.

With the eye-tracking device, we collected two types of eye movement information: saccades and fixations. Saccade means fast eye movements from point to point in jerks, while fixation means that eyes stop for a short period of time [24]. As for the threshold of fixation, we adopt the one used in most previous works (200-500 ms as in [23, 25]) and set it to 250 ms. Because new information is mainly acquired during fixations, most existing studies [3, 16, 23] assumed that eye fixation is equivalent to user examination sequence. Although some recent study [21] showed that eye fixation does not necessary mean examination in many cases, it would be difficult to collect true examination information because this requires user's explicit feedback. Therefore, we still use the recorded fixation sequences to approximate subjects' examination sequences for simplicity. In this way, both click sequences and examination sequences could be restored.

With the data collected in the experiment, we want to find the answers to the following two questions about users' examination behavior on the SERPs.

[2]http://www.thuir.cn/imine/

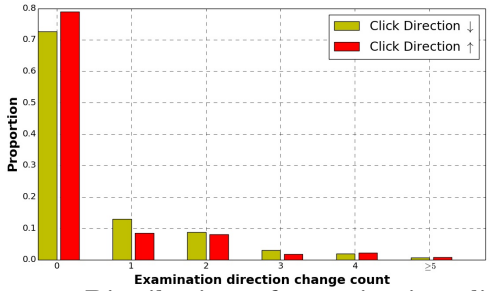

Figure 1: Distribution of examination direction change count for two types of adjacent clicks

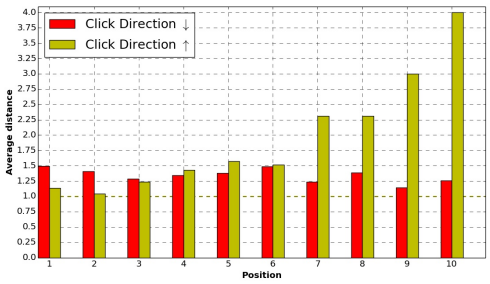

Figure 2: Average examination transition distance according to different examination transition start positions for two types of adjacent clicks.

RQ1: How often do users change the direction of examination between clicks?

RQ2: How far do users' eye gazes jump after examining the current clicked result?

By investigating these two questions, we aim to understand how users behave and to propose corresponding user behavior assumptions in order to model users' examination behavior in a more reasonable way. To simplify the notation, suppose that the first click is at position i and the next click is at position j, if $i < j$, it is a sequential action according to the depth-first assumption (this direction is referred to as "↓"). If $i \geq j$, it is a non-sequential click action according to the definition of revisiting behavior (this direction is referred to as "↑").

To answer the two research questions, we firstly divide all examination sequences into adjacent examination behavior pairs. For a given examination sequence $E =< E_1, E_2, ..., E_t, ..., E_T >$, it will be divided into $T-1$ pairs: $(E_1, E_2), (E_2, E_3), ..., (E_{T-1}, E_T)$. For each pair, similar with the definition of direction in adjacent clicks, we can define its direction as ↑ or ↓ according to whether the sequence of the examination pair follows a depth-first assumption or not.

To investigate ***RQ1***, we consider the examination sequence between ↑ and ↓ adjacent clicks separately. Intuitively, one may believe that the examination sequence between ↓ adjacent clicks should follow the depth-first assumption. In other words, the examination sequence would be consistent with the click sequence.

However, it is also possible that some parts in the examination sequence follow a non-sequential order. Similarly, the examination sequence between ↑ adjacent clicks may also contain ↓ adjacent examination pairs. To find out how often the examination direction change happens between adjacent clicks, we count the number of examination direction changes and the distributions are shown in Figure 1.

From this figure, we can see that no matter whether the click direction is ↑ or ↓, in most cases (72.7% for ↓ and 78.9% for ↑) the whole examination sequences follow the same direction as click direction without any direction changes. The percentage of sequences with direction changes between ↓ clicks is slightly larger than that between ↑ clicks. This phenomenon corresponds well to the behavior pattern in which users re-examine some higher-ranked results before moving to the lower-ranked ones. With this observation, we can formulate the following behavior assumption:

Locally Unidirectional Examination Assumption: Between adjacent clicks, users tend to examine search results in a single direction without changes, and the direction is usually consistent with that of clicks no matter it is ↑ or ↓.

To answer ***RQ2***, we look at the average examination transition distance within adjacent examination pairs. For a given adjacent examination pair (E_{t-1}, E_t), suppose that the first examination E_{t-1} is at position k while the next examination E_t is at position l, the transition distance can be calculated as $|k-l|$. Figure 2 shows the distribution of transition distance in different result positions.

We can see that all transition distances are around 1.25 when user follows top-down (↓) click sequences. While when user follows bottom-up (↑) click sequences, his/her eyes may skip several results to find a specific result.

In particular, we observe larger transition distances for bottom ranking positions, which tend to bring back to the middle positions (positions 5-6) in the list. As all the transition distances are statistically significantly larger than 1 ($p - value < 0.01$ for each position and each click direction based on t-test), we can make the following behavior assumption:

Non First-order Examination Assumption: although the examination behavior between adjacent clicks can be regarded as locally unidirectional, users may skip a few results and examine a result at some distance from the current one following a certain direction.

With the answers to these two research questions, we are able to draw a relatively clear picture of user's examination behavior between adjacent clicks. After a certain user clicks a result i, he/she may start examining results either in a ↑ or a ↓ direction. The user seldom changes the examination direction until he/she clicks another results located at position j (locally unidirectional examination assumption) but he/she may not examine all results on the examination path (non first-order examination assumption).

Compared to existing sequence-based click models such as POM, which assumes that the examination sequence within two clicks can be arbitrary, the actual user behavior shows much simpler patterns. It is thus possible for us to take advantage of the patterns so as to simplify model construction. Compared to THCM that assumes users examine results one by one, the observed user examination behavior demonstrates that user examination may include skips quite frequently. It is necessary for a click model to account for such behaviors.

4. PARTIALLY SEQUENTIAL CLICK MODEL

As we stated in our discussions, the existing click models are unable to correctly cope with the non-sequential behaviors of users we observed. A simple relaxation of the position bias is insufficient for the model to account for the observations we made in the previous section. Therefore, in this paper, we propose to incorporate click sequence information in a different way. At first, click sequence is divided into adjacent click pairs. Considering the two examination behavior assumptions proposed in Sec. 3, the examination process between adjacent clicks could be regarded as both unidirectional and non first-order. With those two assumptions, it is possible for us to employ traditional position-based models in these click sub-sequences. By this means, we can combine the findings in practical user behavior with existing position-based hypotheses consistently.

4.1 Model and Hypotheses

We first introduce some definitions and notations. Suppose that there are N sessions, each of which records certain user interactions with top-M results (M is usually set to 10 in most existing click model resesarches). The result list can be represented as an impression sequence: $D =< d_1, d_2, ..., d_i, ..., d_M >$, where i corresponds to the ranking position (from 1 to M) and d_i is ranked higher than d_j if $i < j$. The relevance of each result is represented by: $R =< R_1, R_2, ..., R_i, ..., R_M >$. With the timestamp information recorded in the logs, we organize the click sequence as $C =< C_1, C_2, ..., C_t, ..., C_T >$, where t is the relative temporal order of a click and C_t records the result position of the t-th click ($1 \leq C_t \leq M$).

The *First-order Click Hypothesis* is usually accepted in most click models such as DBN and UBM. We do the same in this work. It supposes that the click event at time $t+1$ is only determined by the click event at time t. According to this hypothesis, user's click action $C =< C_1, C_2, ..., C_t, ..., C_T >$ can be independently separated to $T+1$ adjacent click pairs: $< C_0, C_1 >, ..., < C_{t-1}, C_t >, ..., < C_T, C_{T+1} >$ (C_0 represents the begining of search process and C_{T+1} represents the end of search process). This makes it possible for us to divide a click sequence into sub-sequences (adjacent click pairs).

According to the *Locally Unidirectional Examination Assumption*, given an observation of adjacent clicks at time t: $O = \{< C_{t-1} = m, C_t = n >\}$, users tend to examine the results on the path from m to n without any direction changes. Then the examination and click sequence between C_{t-1} and C_t can be noted as $< \bar{E}_m, ..., \bar{E}_j, ..., \bar{E}_n >$ and $< \bar{C}_m, ..., \bar{C}_j, ..., \bar{C}_n >$, respectively. Note that different from C_t which is used to record the position of click event, $\bar{E}_j$ and $\bar{C}_j$ ($m \leq j \leq n$ *or* $n \leq j \leq m$) are all binary variables representing whether examination or click behavior happens (=1) or not (=0) on the corresponding result position. In addition, we can also deduce that in the click sequence, only $\bar{C}_m$ and $\bar{C}_n$ have value 1 and the other positions on the path have value 0.

The proposed Partially Sequential Click Model (PSCM) adopt these two assumptions. It is then described as follows:

$$P(C_t|C_{t-1}, ..., C_1) = P(C_t|C_{t-1}) \tag{9}$$

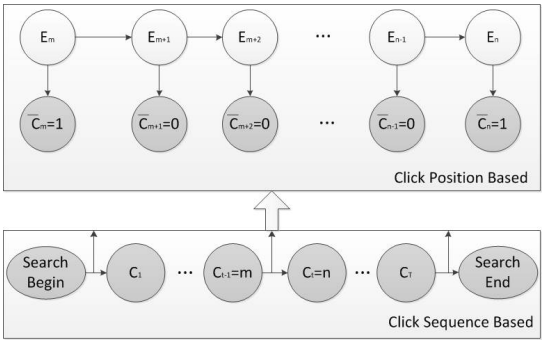

Figure 3: Sketch of Partially Sequential Click Model. Click actions are listed according to their click timestamps. For each adjacent click pair, a position-based framework is constructed based on their click positions.

$$P(C_t = n|C_{t-1} = m) = P(\bar{C}_m = 1, ..., \bar{C}_i = 0, ..., \bar{C}_n = 1) \tag{10}$$

$$P(\bar{E}_i = 1|C_{t-1} = m, C_t = n) = \begin{cases} \gamma_{imn}, m \leq i \leq n \text{ or } n \leq i \leq m \\ 0, other \end{cases} \tag{11}$$

$$\bar{C}_i = 1 \Leftrightarrow \bar{E}_i = 1, R_i = 1 \tag{12}$$

$$P(R_i = 1) = \alpha_{uq} \tag{13}$$

Equation (9) encodes the *first-order click hypothesis* while Equation (10) encodes the *locally unidirectional examination assumption* by restricting the examination process to one-way from m to n. We define the examination probability of $\bar{E}_i$ as Equation (11) because according to Figure 2, the examination behavior between adjacent clicks may not follow cascade assumptions (non first-order examination assumption). The probability of examination depends on the positions of the clicks. This is similar to UBM, which also allow skips, but only within sequential behaviors. PSCM also follows examination hypothesis described in Equation (12) as in most existing click models. α_{uq} corresponds to the relevance of the document URL u at position i to the specific query q.

Figure 3 shows the framework of the PSCM model. Unlike previous position-based models (such as UBM or DBN) which suppose user examine results top-down sequentially, PSCM allows non-sequential interactions. A user may click on a lower position (m) and then a higher position ($n > m$), with all the documents between them having some probability to be examined. Such a behavior is modeled by Equation (10) and Equation (11). Compared to the existing click-sequence-based models, PSCM is more flexible than THCM as it no longer makes first-order examination assumption, and is more controlled than POM as it does not allow arbitrary examination. Such a compromised position is justified by the observations we made on user behaviors. A too rigid model such as THCM would be unable to account for the non-sequential behaviors, while a too flexible model such as POM would give the model too much freedom to be correctly parameterized in practice.

Our model uses two groups of parameters: $\{\alpha_{uq}\}$ represents the probability of being relevant for each query-result pair, $\{\gamma_{imn}\}$ represents the examination

transition probability in either $\downarrow$ and $\uparrow$ directions. Note that γ are global parameters just as those in UBM. Although setting this parameter according to query (or user) may be helpful, we do not do it in this paper in order to simplify the model. This is a problem we will investigate in the future.

Table 2: Conditional probability for $\bar{C}_i$, $\bar{E}_i$, R_i given parameter $\theta^{(v)}$ and C_t (use Λ as the abbreviation of $(\theta^{(v)}, C_{t-1} = m, C_t = n)$).

Conditional probability	value
$P(\bar{C}_i = 0, R_i = 0, \bar{E}_i = 0\|\Lambda)$	$= (1 - \alpha_{uq}^{(v)})(1 - \gamma_{imn}^{(v)})$
$P(\bar{C}_i = 0, R_i = 0, \bar{E}_i = 1\|\Lambda)$	$= (1 - \alpha_{uq}^{(v)})\gamma_{imn}^{(v)}$
$P(\bar{C}_i = 0, R_i = 1, \bar{E}_i = 0\|\Lambda)$	$= \alpha_{uq}^{(v)}(1 - \gamma_{imn}^{(v)})$
$P(\bar{C}_i = 0, R_i = 1, \bar{E}_i = 1\|\Lambda)$	$= 0$
$P(\bar{C}_i = 1, R_i = 0, \bar{E}_i = 0\|\Lambda)$	$= 0$
$P(\bar{C}_i = 1, R_i = 0, \bar{E}_i = 1\|\Lambda)$	$= 0$
$P(\bar{C}_i = 1, R_i = 1, \bar{E}_i = 0\|\Lambda)$	$= 0$
$P(\bar{C}_i = 1, R_i = 1, \bar{E}_i = 1\|\Lambda)$	$= \alpha_{uq}^{(v)}\gamma_{imn}^{(v)}$

As in most existing studies, we assume that the document relevance $\{\alpha_{uq}\}$ and the click events of different sessions are independent of each other. Based on this assumption, we discuss the inference of document relevance.

4.2 Model Inference for PSCM

In our model, two groups of parameters ($\{\alpha_{uq}\}$ and $\{\gamma_{imn}\}$) need to be inferred from click logs. The Expectaction-Maximization (EM) algorithm [15] is used to find the maximum likelihood estimate of the variables $\{\alpha_{uq}\}$ and $\{\gamma_{imn}\}$.

The observation of our model is click sequence ($Y = \{C\}$), the hidden variables are query-result relevance and user examination information ($Z = \{R, E\}$), and the parameters are $\theta = \{\alpha_{uq}, \gamma_{imn}\}$. Therefore, given one specific query-session, the marginal likelihood is:

$$P(Y, Z|\theta) = P(C, E, R|\theta) = \prod_{t=1}^{T} P(C_t, E, R|C_{t-1}, \theta) \quad (14)$$

According to Equation (10) and Equation (12) (omit θ for conciseness):

$$\begin{aligned} &P(C_t = n, E, R|C_{t-1} = m) = \\ &\{\prod_{i=m+1}^{n-1} P(\bar{C}_i = 0|\bar{E}_i, R_i)P(R_i)P(\bar{E}_i|C_{t-1} = m, C_t = n)\} \\ &\cdot \{P(\bar{C}_n = 1|R_n, \bar{E}_n)P(R_n)P(\bar{E}_n|C_{t-1} = m, C_t = n)\} \end{aligned} \quad (15)$$

The conditional expected log-likelihood (Q-function) can be written as (suppose that the parameter at iteration v is $\theta^{(v)}$):

$$Q = E_{E,R|C,\theta^{(v)}}[\log P(C, E, R|\theta)] \quad (16)$$

The values of the required conditional probabilities are given in Table 2. Based on these values, the posterior distributions of $\bar{E}_i$, R_i can be easily calculated. Given N query session and M results for each query, in iteration round v, the formulation of parameter α_{uq} corresponding to a specific query q and result u in Q-function is:

$$\begin{aligned} Q_{\alpha_{uq}} = &\sum_{j=1}^{N}\sum_{\bar{t}}^{T^j}\{I_{mn} \cdot [I_{\neq} \cdot \frac{1 - \alpha_{uq}^{(v)}}{1 - \alpha_{uq}^{(v)}\gamma_{imn}^{(v)}} \cdot \log(1 - \alpha_{uq}) \\ &+ I_{\neq} \cdot \frac{\alpha_{uq}^{(v)}(1 - \gamma_{imn}^{(v)})}{1 - \alpha_{uq}^{(v)}\gamma_{imn}^{(v)}} \cdot \log(\alpha_{uq}) + I_{=} \cdot 1 \cdot \log(\alpha_{uq})]\} \end{aligned} \quad (17)$$

where j is the j-th session in N, T^j is the click sequence length in this session, $\bar{t}$ corresponds to the t-th adjacent click pair $\{t, C_{t-1} = m, C_t = n\}$, $I(\cdot)$ represents the indicator function, I_{mn} is the abbreviation of $I(m \leq i \leq n \ or \ n \leq i \leq m)$, $I_{=}$ is the abbreviation of $I(d_i^j = u, q^j = q, i = n)$ and $I_{\neq}$ is the abbreviation of $I(d_i^j = u, q^j = q, i \neq n)$).

The formulation of parameter γ_{imn} corresponding to a specific position i (the adjacent clicks are m and n) in Q-function is:

$$\begin{aligned} Q_{\gamma_{imn}} = &\sum_{j=1}^{N}\sum_{\bar{t}}^{T^j}\{I_{mn} \cdot [I_{\neq} \cdot \frac{1 - \gamma_{imn}^{(v)}}{1 - \alpha_{uq}^{(v)}\gamma_{imn}^{(v)}} \cdot \log(1 - \gamma_{imn}) \\ &+ I_{\neq} \cdot \frac{\gamma_{imn}^{(v)}(1 - \alpha_{uq}^{(v)})}{1 - \alpha_{uq}^{(v)}\gamma_{imn}^{(v)}} \cdot \log(\gamma_{imn}) + I_{=} \cdot 1 \cdot \log(\gamma_{imn})]\} \end{aligned} \quad (18)$$

By separately taking derivation of α_{uq} on Equation (17) and γ_{imn} on Equation (18), we can generate the corresponding updating formulation for $\alpha_{uq}^{(v+1)}$ and $\gamma_{imn}^{(v+1)}$ in iteration round (v):

$$\begin{aligned} A_1^{(v)} &= \sum_{j=1}^{N}\sum_{\bar{t}}^{T^j}\{I_{mn} \cdot I_{\neq} \cdot \frac{1 - \alpha_{imn}^{(v)}}{1 - \alpha_{uq}^{(v)}\gamma_{imn}^{(v)}}\} \\ A_2^{(v)} &= \sum_{j=1}^{N}\sum_{\bar{t}}^{T^j}\{I_{mn} \cdot I_{\neq} \cdot \frac{\alpha_{imn}^{(v)}(1 - \gamma_{uq}^{(v)})}{1 - \alpha_{uq}^{(v)}\gamma_{imn}^{(v)}}\} \\ A_3^{(v)} &= \sum_{j=1}^{N}\sum_{\bar{t}}^{T^j}\{I_{mn} \cdot I_{=}\} \\ \alpha_{uq}^{(v+1)} &= \frac{A_2^{(v)} + A_3^{(v)}}{A_1^{(v)} + A_2^{(v)} + A_3^{(v)}} \end{aligned} \quad (19)$$

$$\begin{aligned} G_1^{(v)} &= \sum_{j=1}^{N}\sum_{\bar{t}}^{T^j}\{I_{mn} \cdot I_{\neq} \cdot \frac{1 - \gamma_{imn}^{(v)}}{1 - \alpha_{uq}^{(v)}\gamma_{imn}^{(v)}}\} \\ G_2^{(v)} &= \sum_{j=1}^{N}\sum_{\bar{t}}^{T^j}\{I_{mn} \cdot I_{\neq} \cdot \frac{\gamma_{imn}^{(v)}(1 - \alpha_{uq}^{(v)})}{1 - \alpha_{uq}^{(v)}\gamma_{imn}^{(v)}}\} \\ G_3^{(v)} &= \sum_{j=1}^{N}\sum_{\bar{t}}^{T^j}\{I_{mn} \cdot I_{=}\} \\ \gamma_{imn}^{(v+1)} &= \frac{G_2^{(v)} + G_3^{(v)}}{G_1^{(v)} + G_2^{(v)} + G_3^{(v)}} \end{aligned} \quad (20)$$

5. EXPERIMENTS AND DISCUSSIONS

To test the effectiveness of the proposed PSCM model, we compare its performance with a number of existing click models for click prediction and relevance estimation. We choose two of the most popular position-based click models (UBM [11] and DBN [4]) as our first baselines. As discussed in Section 2.2, there exists a number of sequence-based click models (POM [28], THCM [29] and TCM [30]) which can

Table 3: Two large-scale commercial search logs (different languages) used to evaluate click models ("#" represents "number of").

Data	Data-C	Data-Y
Description	Sogou's logs	Yandex's logs
#Distinct Queries	406,345	20,588,928
#Sessions	11,813,260	65,172,853
#Click Sessions	7,951,495	38,288,389
#Multi-click Sessions	2,358,648	11,383,886
Experiments	Click Perplexity NDCG User Preference	Click Perplexity

also take non-sequential behavior into consideration. These models are also used as our baselines.

Three types of experiments are performed to validate our model. First we evaluate the click model in terms of predicting click probabilities (click perplexity) from search logs, which is a widely adopted metric to evaluate click models' performances [4, 11, 29]. After that, we use the predicted relevance as a signal for document ranking and evaluate each click model's ranking performance with traditional IR metrics (in this paper we use NDCG metric [17]). Finally, since professional assessors' relevance annotation may not always agree with users' actual judgments [6], we also conduct a user preference test [26] in which participants are asked to label their preferences on parellaled result lists generated by different models.

5.1 Experimental Setups

As discussed in Sec. 2, TCM and POM are not directly comparable to other click models due to different behavior assumptions. Therefore, we first describe how we address the limitations and adapt them for performance comparison. As for other baseline models, we refer to the implementations from [7]. Our code implementations and evaluation data set are publicly available at https://github.com/THUIR/PSCMModel.

5.1.1 Baseline Model Adaptation

(TCM) As we have mentioned in Sec. 2.2, this model can only handle result lists containing exactly two results. As this model enumerates all possible click sequences for a specific ranking list (5 possible situation for two results [30]), it faces the exponential explosion problem when the number of results becomes large. Therefore, we cannot expand this model to M results in one SERP (M equals to 10 in our data set). In order to compare this model with other existing click models which can handle arbitrary number of results in an SERP, we made an trivial expansion of TCM model: we separate these results into $M/2$ pairs ($< 1,2 >, < 3,4 >, ..., < M-1, M >$) and implement TCM model for each pair separately. Then, from each pair we can deduce two results' relevance and click probability. We thus combine $M/2$ pairs together to generate click prediction and relevance prediction for the whole result list.

(POM) Although POM can model non-sequential behaviors in user interactions, this model is not designed to predict the click probability or result relevance for a specific query, as we discussed in Sec. 2. It is unfair to compare POM with other models. To make POM more suitable for the click and relevance prediction tasks, we modify the original POM model by setting a relevance score for each specific document-query pair.

According to search logs, clicks can be re-organized as a temporal sequence of behaviors by record timestamps: $E =< E_1, E_2, ..., E_t, ..., E_T >$, where t represents the events' relative order, E_t represents the corresponding ranking of the result being examined at time t and $C =< C_1, C_2, ..., C_t, ..., C_T >$, where C_t represents the corresponding result is clicked or not. From search logs, we can only observe which results are clicked by users. Based on the assumption that user must examine a result before clicking on it (examination hypothesis [8]), we can infer that the clicked results must be examined. Therefore, user may examine some results in his/her browsing process but not click them given a click sequence observation $O = \{(E_1 = e_1, C_1 = 1), ..., (E_T = e_T, C_T = 1)\}$. So an arbitrary $O' = \{(E'_1, C'_1), ..., (E'_k, C'_k), ...(E'_K, C'_K)\}$ can be generated based on the original observation O where $O \subseteq O'$. The POM model assumes that the probability of original observation is the summation of the probabilities of all compatible examination sequences. Furthermore the POM model makes first order assumption that current examining result only depends on previous examination. So the POM model can be represented as follows:

$$P(O) = \sum_{O'} P(O') = \sum_{O'} \prod_{i=1}^{K} P(C_i|E_i)P(E_i|E_{i-1}) \quad (21)$$

$$P(C_i = 1|E_i = m) = c_m \quad (22)$$

$$P(E_i = n|E_{i-1} = m) = e_{mn} \quad (23)$$

where E_0 represents the submitted query received at the begining of a search session, c_m is the click probability of rank m, e_{mn} is the examination transition probability. According to the formulations above, POM model can model arbitrary examination orders. As a matter of fact, it can describe the non-sequential click behavior during search process.

However, in the original POM model, given the examination of a result, the click probability is only dependent on the result position (Equation (22)). Therefore, we simply adopt the examination hypothesis that given the examination of a result, the click probability is dependent on the result's relevance. Therefore, the Equation (22) is revised as:

$$P(C_i = 1|E_i = m) = \alpha_{uq} \quad (24)$$

where α_{uq} is the relevance of query-document pair. So the click probability no longer depends on the rank position but depends on the search query. Once we get α_{uq}, we can compare POM with other click models in terms of click perplexity and NDCG. The parameter estimation formulation is similar to the original model [28] by using Expectation-Maximization (EM) algorithm.

5.1.2 Data Sets

To evaluate the click models, we utilize two real-world large-scale data sets collected by Sogou from China and Yandex[3] from Russia. The detailed statistics of the two

[3]The Yandex dataset is publicly available at https://www.kaggle.com/c/yandex-personalized-web-search-challenge/data.

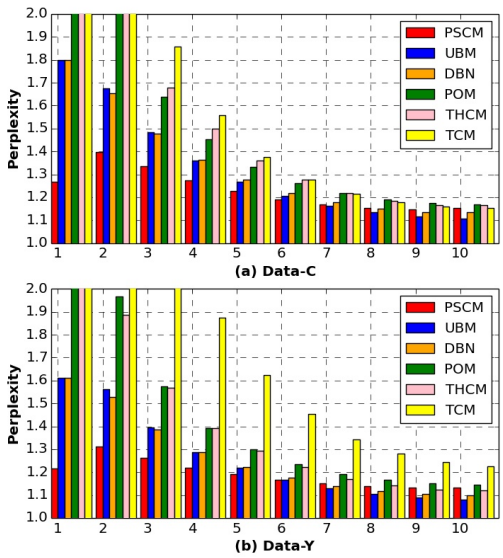

Figure 4: Click perplexities of different positions on Data-C and Data-Y.

Table 4: Overall click perplexity of each model on Data-C and Data-Y (all improvements are statistically significant according to t-test with $p-value < 10^{-5}$).

Model	Data-C	PSCM Impr.	Data-Y	PSCM Impr.
PSCM	1.232	-	1.192	-
UBM	1.332	30.1%	1.265	27.4%
DBN	1.339	31.6%	1.267	27.9%
POM	1.782	70.4%	1.826	76.7%
THCM	1.583	60.3%	1.545	64.7%
TCM	2.435	83.8%	2.691	88.6%

datasets can be found in Table 3. Specifically, we evaluate the click models on click perplexity for both of the whole datasets. Then for evaluating relevance prediction of click models (with respect to either NDCG or user preference), we use a subset of Data-C for testing relevance prediction because relevance annotation is not available for the Yandex set.

5.2 Evaluation of Click Prediction

We use two search logs (see Table 3) to compute the click perplexity of each model. For each dataset, we split all query sessions into the training and testing sets at a ratio of 70% : 30% as previous studies did [5, 27].

Click perplexity [27] measures the probability of the actual click events occurring for each session and each position. It indicates how well a model can predict the clicks. A smaller perplexity value indicates a better modeling performance, and the value reaches 1 in the ideal case. The improvement of click perplexity CP_1 over CP_2 is calculated through $\frac{CP_2-CP_1}{CP_2-1} * 100\%$ [5, 27].

Table 4 illustrates the overall perplexity of each model and Figure 4 shows perplexities in different positions. We can see that PSCM achieves the best overall results among all click models. According to Table 4, existing sequence-based models (POM, THCM and TCM) cannot achieve as good performance as position-based models (UBM and DBN). This suggests that the assumptions on the examination and click sequences are either too strict (e.g. restrict one-by-one examination in THCM) or too flexible (e.g. allowed at any position in POM). As we observed, user behaviors basically follow the same direction, but with occasional changes of direction and jumps. Our model is built on these observations. As we can see in Table 4, our model can better predict clicks than all the other models. This is a strong indication that the sequence of user behaviors is better coped with in our model.

From Figure 4 we can also see that PSCM generates very good results in the top positions (1-5). However, it achieves slightly worse performance than position-based models (UBM and DBN) in the bottom positions (8-10). This phenomenon can be explained by the fact that PSCM has taken more possible examination sequences (compared to position-based models) into account and thus it gives slightly higher click probability in bottom positions. While in most query sessions, a user usually skips results in bottom positions. Therefore, PSCM will receive more penalties in perplexity than UBM and DBN. However, these degradations on bottom positions are much less significant than the improvements on top positions because 1) the scale of the degradations are far less than that of the improvements; and 2) users often care more about the top positions than bottom positions.

5.3 Evaluation of Relevance Estimation

As a click model also provides a prediction on the relevance of a document for a query - α_{uq}, we can rank documents according to this value. The ranking results can be measured using NDCG [17]. This evaluation is performed only on Data-C for which human evaluators can be recruited to judge document relevance. The same evaluation cannot be done on Data-Y because the data has been encoded to unreadable codes, and no relevance information is available.

For a random sample of 1,187 queries in Data-C, several professional assessors (from Sogou.com, without knowing any information about this work) annotated a number of results' relevance scores for each query. The annotation is performed with 5 grades ("Perfect", "Excellent", "Good", "Fair" and "Bad") as in most existing studies such as [31]. Majority voting is adopted to decide the relevance score if there are conflicts (at least 3 assessors were involved in each query-result pair annotation). Due to limited human resources, top five results for 460 queries were annotated while only top 3 results are annotated for the other 727 queries. With the annotation results, we calculated the NDCG@N (N=3,5) scores for different click models and the results are shown in Figure 5.

We can see that PSCM achieves statistically significant improvement over the other click models. We can also see that position-based click models (UBM and DBN) still achieve better relevance prediction than sequence-based click models (POM, THCM and TCM) as in click prediction. This confirms our assumption in Sec. 2 that existing click-sequence-based click models are less suitable to cope with sequences of user interactions.

5.4 User Preference Test

Besides the evaluation in relevance estimation, we also want to find out whether the ranking lists provided by PSCM are preferred by real users than other click models.

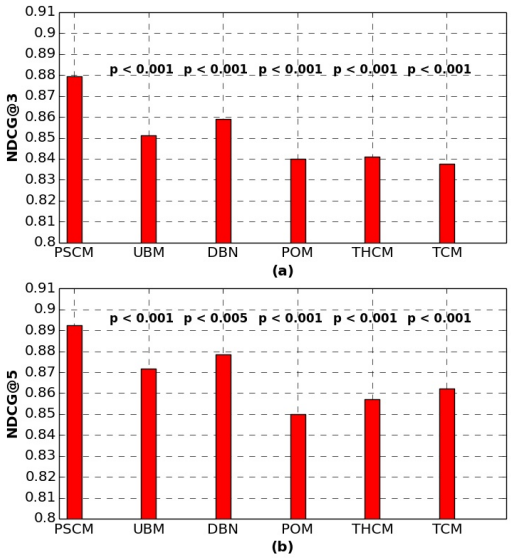

Figure 5: Relevance estimation performances in terms of NDCG@3 (1,187 queries) and NDCG@5 (460 queries) for Data-C. The Improvements of PSCM compared with other models are all significant (paired t-test $p - values$ are also shown in the figure)

Therefore, we conducted a side-by-side user preference test of PSCM against the ranking list provided by DBN (second best performance according to NDCG) as is done in [18].

We randomly sampled 200 queries and recruited 22 human evaluators (all university students with a variety of majors and self-reported search expertise) to label their preferences. For each query, the evaluators were shown two result lists produced by PSCM and DBN respectively. They were required to label a preference degree after reviewing the two lists with 9 levels: "Left results list is better {+4, +3, +2, +1}", "Tie {0}" and "Right results list is better {+1, +2, +3, +4}". The evaluators did not know which list was generated from one specific model as the display side was randomly chosen. Each evaluator annotated at least 50 result list pairs to make sure that each pair was labeled by at least 5 different evaluators.

Figure 6 shows the label distribution of all evaluators. We can see that most of them prefer PSCM (46.5%), and only 27.5% labels prefer DBN. To summarize the user preference for each query, we also use the majority voting method to merge labels from different evaluators. The experimental results are illustrated in Figure 7. We can also see that evaluators show clear preferences to PSCM in nearly half of all queries (47.5%). The proportion of "PSCM better" is statistically significantly larger than the proportion of "Tie" and "DBN better" according to Pearson's chi-squared test ($p - value < 0.001$). This result is consistent with the distribution of labels, which indicates that PSCM can produce better ranking lists than DBN.

With the evaluation results in click perplexity, NDCG and user preference test, we can conclude that our model can better describe user's actual examining and clicking behavior. In addition, this model can also provide more accurate estimations in query-result relevance, and thus generate better result ranking lists.

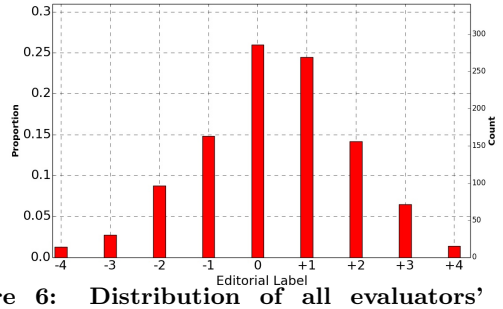

Figure 6: Distribution of all evaluators' labels ("$+1 \sim +4$" represents user prefers PSCM, "$-1 \sim -4$" represents user prefers DBN and "0" represents tie).

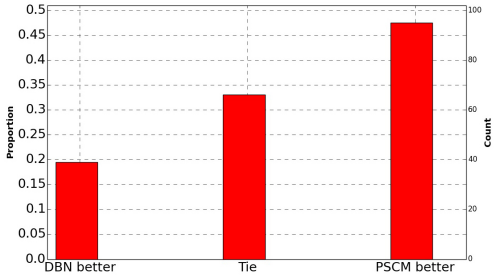

Figure 7: Editorial evaluation for PSCM and DBN in the user preference test.

6. CONCLUSIONS

In this paper, we address the problem of properly incorporating click sequence information into click models. First, we carried out a laboratory eye-tracking experiment to analyze search users' examination behaviors. From the observations, we formulated two assumptions: the locally unidirectional assumption and non-first-order examination assumption. Based on these findings, we proposed a new click model named PSCM, which incorporates non-sequential click behaviors into click models while following the two assumption on examinations between two clicks. The experimental results on large-scale click-through data showed that our model outperforms existing models in click prediction. We also conducted tests on query-result relevance estimation and user preference of ranking lists. The experimental results show that PSCM outperform existing models in both relevance evaluation (NDCG) and user preference test. This study shows the importance for a click model to correct cope with user's interaction sequences. Compared to the previous models, the assumptions made in our model are more realistic and correspond better to the observations in practice. The proposed model can be further improved on several aspects. For example, as click dwell time has been proved to be a very useful signal for relevance prediction, we plan to combine click dwell time information with click sequence information together to better model users' search behaviors in future work.

7. ACKNOWLEDGMENTS

This work was supported by National Key Basic Research Program (2015CB358700), Tsinghua-Samsung Joint Lab, Tsinghua University Initiative Scientific Research Program (2014Z21032), and Natural Science Foundation (61472206) 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).

8. REFERENCES

[1] E. Agichtein, E. Brill, S. T. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. *SIGIR'06*, pages 3–10, Aug. 2006.

[2] A. Broder. A taxonomy of web search. In *SIGIR'02*, volume 36, pages 3–10. ACM, 2002.

[3] G. Buscher, S. White, Ryen W, and J. Huang. Large-scale analysis of individual and task differences in search result page examination strategies. In *WSDM'12*, pages 373–382. ACM, 2012.

[4] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. *WWW'09*, pages 1–10, Apr. 2009.

[5] D. Chen, W. Chen, H. Wang, Z. Chen, and Q. Yang. Beyond ten blue links: enabling user click modeling in federated web search. *WSDM'12*, pages 463–472, Feb. 2012.

[6] F. Chen, Y. Liu, Z. Dou, K. Xu, Y. Cao, M. Zhang, and S. Ma. Revisiting the evaluation of diversified search evaluation metrics with user preferences. In *Information Retrieval Technology*, pages 48–59. Springer, 2014.

[7] A. Chuklin, P. Serdyukov, and M. De Rijke. Click model-based information retrieval metrics. In *SIGIR'13*, pages 493–502. ACM, 2013.

[8] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *WSDM'08*, pages 87–94. ACM, 2008.

[9] E. Cutrell and Z. Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *CHI 07*, pages 407–416. ACM, 2007.

[10] G. Dupret and C. Liao. A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. In *WSDM'10*, pages 181–190. ACM, 2010.

[11] G. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. *SIGIR'08*, pages 331–338, July 2008.

[12] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. In *SIGIR 04*, pages 478–479. ACM, 2004.

[13] F. Guo, C. Liu, and Y. M. Wang. Efficient multiple-click models in web search. *WSDM'09*, pages 124–131, Feb. 2009.

[14] Q. Guo, D. Lagun, D. Savenkov, and Q. Liu. Improving relevance prediction by addressing biases and sparsity in web search click data. In *WSCD'12*, pages 71–75, 2012.

[15] M. R. Gupta and Y. Chen. *Theory and use of the EM algorithm*. Now Publishers Inc, 2011.

[16] J. Huang, R. W. White, and S. Dumais. No clicks, no problem: using cursor movements to understand and improve search. In *CHI's 11*, pages 1225–1234. ACM, 2011.

[17] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[18] L. Jie, S. Lamkhede, R. Sapra, E. Hsu, H. Song, and Y. Chang. A unified search federation system based on online user feedback. In *SIGKDD'13*, pages 1195–1203. ACM, 2013.

[19] T. Joachims, L. A. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting click through data as implicit feedback. *SIGIR'05*, pages 154–161, July 2005.

[20] K. Klockner, N. Wirschum, and A. Jameson. Depth- and breadth-first processing of search result lists. *CHI'04*, pages 1539–1539, Apr. 2004.

[21] Y. Liu, C. Wang, K. Zhou, J. Nie, M. Zhang, and S. Ma. From skimming to reading: A two-stage examination model for web search. In *CIKM'14*, pages 849–858. ACM, 2014.

[22] L. Lorigo, B. Pan, H. Hembrooke, T. Joachims, L. A. Granka, and G. Gay. The influence of task and gender on search and evaluation behavior using google. *Information Processing and Management*, 42(4):1123–1131, 2006.

[23] R. Navalpakkam, Vidhya, S. Ravi, A. Ahmed, and A. Smola. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *WWW 13*, pages 953–964, 2013.

[24] K. Rayner. Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology*, 62(8):1457–1506, 2009.

[25] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 71–78. ACM, 2000.

[26] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *SIGIR'10*, pages 555–562. ACM, 2010.

[27] C. Wang, Y. Liu, M. Zhang, S. Ma, M. Zheng, J. Qian, and K. Zhang. Incorporating vertical results into search click models. In *SIGIR'13*, pages 503–412. ACM, 2013.

[28] K. Wang, N. Gloy, and X. Li. Inferring search behaviors using partially observable markov(pom) model. *WSDM'10*, pages 211–220, Feb. 2010.

[29] D. Xu, Y. Liu, M. Zhang, S. Ma, and L. Ru. Incorporating revisiting behaviors into click models. In *WSDM'12*, pages 303–312. ACM, 2012.

[30] W. Xu, E. Manavoglu, and E. Cantu-Paz. Temporal click model for sponsored search. In *SIGIR'10*, pages 106–113. ACM, 2010.

[31] H. Yang, A. Mityagin, K. M. Svore, and S. Markov. Collecting high quality overlapping labels at low cost. In *SIGIR'10*, pages 459–466. ACM, 2010.

[32] Y. Yue, R. Patel, and H. Roehrig. Beyond position bias: examining result attractiveness as a source of presentation bias in click through data. *WWW'10*, pages 1011–1018, Apr. 2010.