

## Time-Aware Click Model

YIQUN LIU, XIAOHUI XIE, and CHAO WANG, Tsinghua University  
 JIAN-YUN NIE, Université de Montréal  
 MIN ZHANG and SHAOPIING MA, Tsinghua University

Click-through information is considered as a valuable source of users' implicit relevance feedback for commercial search engines. As existing studies have shown that the search result position in a search engine result page (SERP) has a very strong influence on users' examination behavior, most existing click models are position based, assuming that users examine results from top to bottom in a linear fashion. Although these click models have been successful, most do not take temporal information into account. As many existing studies have shown, click dwell time and click sequence information are strongly correlated with users' perceived relevance and search satisfaction. Incorporating temporal information may be important to improve performance of user click models for Web searches. In this article, we investigate the problem of properly incorporating temporal information into click models. We first carry out a laboratory eye-tracking study to analyze users' examination behavior in different click sequences and find that the user common examination path among adjacent clicks is linear. Next, we analyze the user dwell time distribution in different search logs and find that we cannot simply use a click dwell time threshold (e.g., 30 seconds) to distinguish relevant/irrelevant results. Finally, we propose a novel time-aware click model (TACM), which captures the temporal information of user behavior. We compare the TACM to several existing click models using two real-world search engine logs. Experimental results show that the TACM outperforms other click models in terms of both predicting click behavior (perplexity) and estimating result relevance (NDCG).

CCS Concepts: • **Information systems** → **Web searching and information discovery**; **Retrieval models and ranking**;

Additional Key Words and Phrases: Click model, click sequence, click dwell time

### ACM Reference Format:

Yiqun Liu, Xiaohui Xie, Chao Wang, Jian-Yun Nie, Min Zhang, and Shaoping Ma. 2016. Time-aware click model. *ACM Trans. Inf. Syst.* 35, 3, Article 16 (December 2016), 24 pages.

DOI: <http://dx.doi.org/10.1145/2988230>

## 1. INTRODUCTION

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

---

This article is an extension of Wang et al. [2015]. Compared to the previous conference version, it introduces a new time-aware click model that incorporates click dwell time information. It also includes an extensive experimental assessment of the new model and compares the performance to several existing models, including the partially sequential click model. This work was supported by the Tsinghua University Initiative Scientific Research Program (2014Z21032), the National Key Basic Research Program (2015CB358700), and the Natural Science Foundation of China (61622208, 61532011).

Authors' addresses: Y. Liu, X. Xie, C. Wang, M. Zhang, and S. Ma, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing, China, 100084; emails: {yiqunliu, z-m, msp}@tsinghua.edu.cn, chaowang0707@gmail.com, xiexh\_thu@163.com; J.-Y. Nie, Département d'informatique et recherche opérationnelle, Université de Montréal, C.P. 6128, Succ Centre-Ville Montréal, Québec, Canada, H3C 3J7; email: nie@iro.umontreal.ca.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or [permissions@acm.org](mailto:permissions@acm.org).

© 2016 ACM 1046-8188/2016/12-ART16 \$15.00

DOI: <http://dx.doi.org/10.1145/2988230>

Although click-through logs can provide implicit feedback of users' click preferences [Agichtein et al. 2006b], it is difficult to derive accurate absolute relevance judgments owing to the existence of click noises and behavior biases. Joachims et al. [2005] 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 toward many aspects, such as "position" [Craswell et al. 2008; Joachims et al. 2005], "trust" [Yue et al. 2010], and "presentation" [Wang et al. 2013]. To address these problems, researchers have proposed several click models to describe users' practical browsing behavior and to obtain an unbiased estimation of result relevance [Guo et al. 2009a; Dupret and Piwowarski 2008; Chapelle and Zhang 2009].

Most click models follow the findings from Craswell et al. [2008] and Joachims et al. [2005] that users' attention decreases from top to bottom and assume that users' potential examination/click paths are unique: the examination/click sequence is consistent with the result position. Therefore, these models do not actually take practical temporal information into account. As modern search logs contain a time stamp for each user interaction (querying, clicking, etc.), we can obtain two important messages from this time stamp: the sequence of user clicks and the dwell time after each user click.

For the click sequence information, eye-tracking experiments [Lorigo et al. 2006] showed that only 34% of search users' scan paths are linear, whereas more than 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. We counted the nonsequential click proportion of multiclick query sessions (when a user clicked two or more results on one SERP) from two commercial search engine logs (Sogou and Yandex; see details in later in Table I). We find that nearly one third (27.9% for Sogou and 30.4% for Yandex) of multiclick sessions contain nonsequential click actions. Although most existing click models are based on ranking positions rather than action sequences, the click sequence information is usually ignored, and nonsequential clicking behaviors are not considered. Dupret and Liao [2010] and Guo et al. [2012] 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 click sequence information into account.

As for click dwell time information, existing studies [Kim et al. 2014; Fox et al. 2005] showed that dwell time on the landing page led by user clicks (click dwell time) is a very strong indicator for user-perceived result relevance and user-perceived search satisfaction. Fox et al. [2005] showed that users are more willing to spend longer durations of time on pages that are interesting and relevant. Kim et al. [2014] also showed that the longer the dwell time, the more satisfied the user will be, and the more relevant the search result tends to be. Therefore, click dwell time information will be very helpful for us to better understand users' click behavior and make an accurate relevance estimation.

Some existing click models [Xu et al. 2010; Wang et al. 2010; Xu et al. 2012] have tried to cope with click sequence information. These models relax the restrictions on users' examination sequences (e.g., Wang et al. [2010] assumes that examination sequences can be arbitrary) to increase models' descriptive power. However, most of these methods abandon the prior knowledge of user examination preference generated from other user behavior studies, which has been found useful. In practice, these models cannot achieve performance that is comparable to other popular click models according to our experimental results.

To better understand users' search interaction processes, we design a laboratory study to analyze users' practical examination patterns. Our observations confirm clearly that many click behaviors are nonsequential. 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 onto the existing position-based models by adding new hypotheses. By this means, we not only inherit a framework that has already proved to be effective but also combine sequential information to better capture users' preferences for different search results.

To better use click dwell time information, we analyze the dwell time distribution for different search logs (Sogou and Yandex). We verified the previous findings in Agichtein et al. [2006a] that clicks with dwell time longer than a certain threshold (e.g., 30 seconds) are good indicators of users' perceived relevance. We also find that the dwell time distribution in different search engines may be rather different, which means that we must take the distribution factor into consideration to better model user behavior. Combining our findings with the previous conclusions from Kim et al. [2014], we design different mapping functions to model user satisfaction based on click dwell time and further introduce the satisfaction factor into click models. Although a few existing models [Chapelle and Zhang 2009] have attempted to take user satisfaction into account, this is the first time that click dwell time is used as a satisfaction indicator in click models.

Our contributions of this article are as follows:

- An eye-tracking experiment is carried out to analyze users' nonsequential examination and click behavior on SERPs.
- The novel time-aware click model (TACM) is proposed to incorporate click sequence information and click dwell time information.
- We show experimentally that the proposed TACM outperforms the existing models on two real-world commercial search engine datasets (one of which is publicly available).

This article is organized as follows. Various click models are reviewed in Section 2. In Section 3, we outline insights of studies on examination/click sequences and click dwell time. In Section 4, we formally introduce the TACM and compare it to the partially sequential click model (PSCM). We report experiments on the TACM and compare it to existing click models in Section 5. Finally, conclusions and future work are discussed in Section 6.

## 2. RELATED WORK

In this section, we introduce related work on click sequences and click dwell time information. We first introduce some basic click models for Web search [Chuklin et al. 2015] to show the essential ideas and assumptions of the click model and then introduce some existing click models that can partially handle temporal information.

### 2.1. Basic Click Models

Most click models follow the examination hypothesis [Craswell et al. 2008], in which 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 that  $P(R_i = 1) = r_u$ , which is the probability of the perceived relevance), and these two conditions are independent of each other.

$$C_i = 1 \rightarrow E_i = 1, R_i = 1 \quad (1)$$

$$E_i = 0 \rightarrow C_i = 0 \quad (2)$$

$$R_i = 0 \rightarrow C_i = 0 \quad (3)$$

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

$$P(C_i = 1) = P(E_i = 1)P(R_i = 1). \quad (4)$$

The click action is simply mapped to each search result's ranking position. Based on the assumption that a user examines from the top position to the bottom position, this kind of click model naturally takes position bias into account.

Craswell et al. [2008] proposed the cascade model, which assumes that while a user examines results from top to bottom sequentially, he or she immediately decides whether to click on a result. The cascade model is mostly suitable for single-click sessions. Several succeeding models were proposed to improve both its applicability and performance.

$$P(E_1) = 1 \quad (5)$$

$$P(E_{i+1} = 1 | E_i = 1, C_i) = 1 - C_i \quad (6)$$

Here the examination of the  $(i + 1)$ th result indicates that the  $i$ th result has been examined but not clicked. Although the cascade model performs well in predicting CTRs, this model is only suited for a single-click scenario.

Based on the cascade hypothesis, the dependency click model (DCM) [Guo et al. 2009a] extends the cascade model to model user interactions within multiclick sessions. The 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. The DCM is characterized as follows:

$$P(E_{i+1} = 1 | E_i = 1, C_i = 0) = 1, \quad (7)$$

$$P(E_{i+1} = 1 | E_i = 1, C_i = 1) = \lambda_i, \quad (8)$$

where  $\lambda_i$  represents the preservation probability<sup>1</sup> of the position  $i$ .

Subsequently, the user browsing model (UBM) [Dupret and Piwowarski 2008] further refined 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:

$$P(E_i = 1 | C_{1\dots i-1}) = \lambda_{r_i, d_i}, \quad (9)$$

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) [Chapelle and Zhang 2009] is the first model to consider presentation bias due to a 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:

$$P(R_i = 1) = r_u, \quad (10)$$

$$P(S_i = 1 | C_i = 1) = s_u, \quad (11)$$

$$P(E_{i+1} | E_i = 1, S_i = 0) = \lambda, \quad (12)$$

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  $\lambda$  represents the probability of continuing the examination process.

<sup>1</sup>The probability of the  $(i + 1)$ th result being examined when the  $i$ th document is clicked.

Subsequently, the click chain model (CCM) [Guo et al. 2009b] uses Bayesian inference to obtain the posterior distribution of the relevance. In contrast to other existing models, this model introduces skipping behavior. The CCM is scalable for large-scale click-through data, and the experimental results show that it is effective for low-frequency (also known as long-tail) queries.

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

## 2.2. Click Dwell Time

Click dwell time measures how long it takes for someone to return to a SERP after clicking on a result. Usually, it is recorded in the search engine's behavior log data, which makes it practical for the person to make use of this kind of information.

Kim et al. [2014] conducted an experiment to estimate click dwell time distributions for SAT (satisfied) or DSAT (dissatisfied) clicks for different click segments. The experimental results showed that the longer the dwell time, the more satisfied the user will be, and the more relevant the search result. In Kim et al. [2014], dwell time was measured as time between the click and the next observed click or query, which is the same as our method in this work. Fox et al. [2005] also made the conclusion that users are more willing to spend longer times on pages that are interesting and related to their focus. Smucker and Clarke [2012] analyzed the correlation between click dwell time and user information gain, and found that the correlation is positive but not linear.

Borisov et al. [2016a] was among the first to propose that time elapsed between a pair of user actions depends on the context of behaviors. They further constructed a context-aware model to predict the time between user actions in contexts. Their work shows that the dwell time of user clicks is affected by many different factors, and incorporating such information may help the behavior model to better correlate with users' practical actions.

The TACM is based on these existing findings and tries to incorporate dwell time into the click modeling process. By doing so, it is our hope to make better use of the feedback information provided by the dwell time to improve model performance.

## 2.3. Temporal Click Models

Several studies [Xu et al. 2010; Wang et al. 2010; Xu et al. 2012] have tried to take temporal click information into consideration.

Xu et al. [2010] first proposed a temporal click model (TCM) to model user click behavior for sponsored searches. They enumerate all possible permutation of click sequences for search results. This model can only handle two results/ads in a SERP. This makes it impossible to cope with the whole ranked result list like in other click models.

Wang et al. [2010] introduced a partially observable Markov model (POM) to model arbitrary click orders. The POM treats user examination events as a partially observable stochastic process. Although the POM can model nonsequential 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 to other state-of-the-art click models such as the 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 real user behavior.

Xu et al. [2012] proposed a temporal hidden click model (THCM) to cope with non-sequential click actions. They focused on revisiting behavior and assumed that after clicking a search result, the user has a probability of going back to examine previous results (bottom-up). However, their model was also based on a one-order Markov examination assumption and supposes that users examine results one by one in the examination process, which does not necessarily correspond to practical user behavior (see Section 3).

Although the preceding three click models have the potential to take click sequence information into consideration, compared to our proposed PSCM [Wang et al. 2015], their adopted methodologies are less suitable for dealing with practical search behavior in modern commercial search engines. The PSCM is inspired by an eye-tracking study on real users' nonsequential SERP behavior and therefore corresponds better to real-world user behavior.

Zhang et al. [2014] proposed a click model based on recurrent neural networks (RNNs) for sponsored search. They directly model the dependency on users' sequential behaviors into the click prediction process through the recurrent structure in the RNN. Borisov et al. [2016b] also proposed an RNN-based click model to model users' sequential click behaviors. These models only take click sequence information into account and ignore the influence of different click dwell times among click actions.

As related studies showed that the click dwell time has a positive correlation with user satisfaction, we tried to design some functions that map click dwell time to user satisfaction and incorporated this into our click models. We designed two different mapping functions: one is a linear mapping function, and the other is an exponential function based on Smucker and Clarke [2012]. In Section 5, we implement these different mapping functions and compare them. As the PSCM showed better performance compared to other click sequence-based models, we choose the PSCM as the basic framework for our new model and try to add click dwell time information into this framework.

### 3. USER BEHAVIOR ANALYSIS

#### 3.1. Click Sequence Analysis

To investigate users' examination sequences 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 was similar to other Web search eye-tracking studies, such as those of Granka et al. [2004] and Cutrell and Guan [2007].

Subjects were provided with a list of 25 search tasks. Each task was accompanied by a fixed query (with an explanation of the information needed to avoid ambiguity) and a Chinese commercial search engine's first result page. We crawled and stored the corresponding SERPs to ensure that all subjects saw the same page for each query. With this setup, each search task (query session) corresponded to one specific SERP. The queries for the search tasks were sampled from the NTCIR IMine task.<sup>2</sup> As different types of information needs [Broder 2002] may also affect browsing behavior [Granka et al. 2004], the selected search tasks covered different types of search intents. In the query set, 5 of the queries were "Navigational" (e.g., "Meizu's official Website"), 10 were "Informational" (e.g., "What is the sound card?"), and 10 were "Transactional" (e.g., "Web browser download").

<sup>2</sup><http://www.thuir.cn/imine/>.

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

With the eye-tracking device, we collected two types of eye movement information: saccades and fixations. A *saccade* refers to a fast eye movement from point to point in jerks, whereas *fixation* means that the eyes stop moving for a short period of time [Rayner 2009]. As for the threshold of fixation, we adopted the one used in most previous works (200 to 500 ms, as in Navalpakkam et al. [2013] and Salvucci and Goldberg [2000]) and set it to 250ms. Because new information is mainly acquired during fixation, most existing studies [Huang et al. 2011; Buscher et al. 2012; Navalpakkam et al. 2013] assumed that eye fixation is equivalent to the user's examination sequence. Although some recent studies [Liu et al. 2014] showed that eye fixation does not necessary mean examination in many cases, it would be difficult to collect true examination information because this requires users' explicit feedback. Therefore, we still used the recorded fixation sequences to approximate subjects' examination sequences for simplicity. In this way, both click sequences and examination sequences could be reconstructed.

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

*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 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 " $\downarrow$ "). If  $i \geq j$ , it is a nonsequential click action according to the definition of revisiting behavior (this direction is referred to as " $\uparrow$ ").

To answer the preceding two research questions, we first divide all examination sequences into adjacent examination behavior pairs. For a given examination sequence  $E = \langle E_1, E_2, \dots, E_t, \dots, E_T \rangle$ , it will be divided into  $T - 1$  pairs:  $(E_1, E_2), (E_2, E_3), \dots, (E_{T-1}, E_T)$ . For each pair, similar to the definition of direction in adjacent clicks, we can define its direction as  $\uparrow$  or  $\downarrow$  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  $\uparrow$  and  $\downarrow$  adjacent clicks separately. Intuitively, one may believe that the examination sequence between  $\downarrow$  adjacent clicks should follow the depth-first assumption—in other words, that the examination sequence would be consistent with the click sequence.

However, it is also possible that some parts of the examination sequence follow a nonsequential order. Similarly, the examination sequence between  $\uparrow$  adjacent clicks may also contain  $\downarrow$  adjacent examination pairs. To find out how often examination direction changes occur between adjacent clicks, we counted the number of examination direction changes; their distributions are shown in Figure 1.

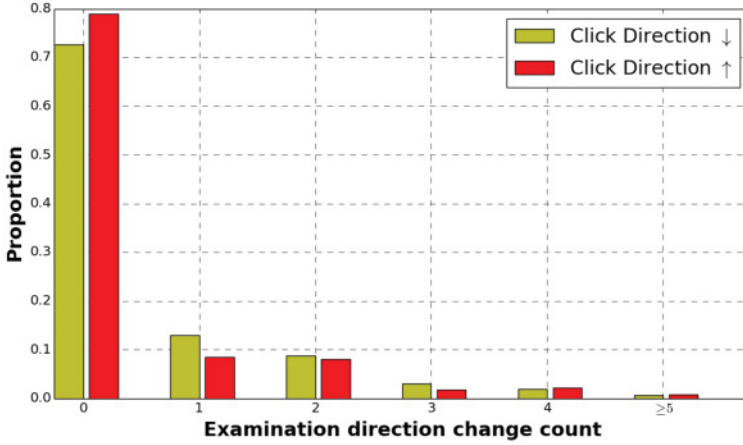


Fig. 1. Distribution of the number of examination direction changes for two types of adjacent clicks.

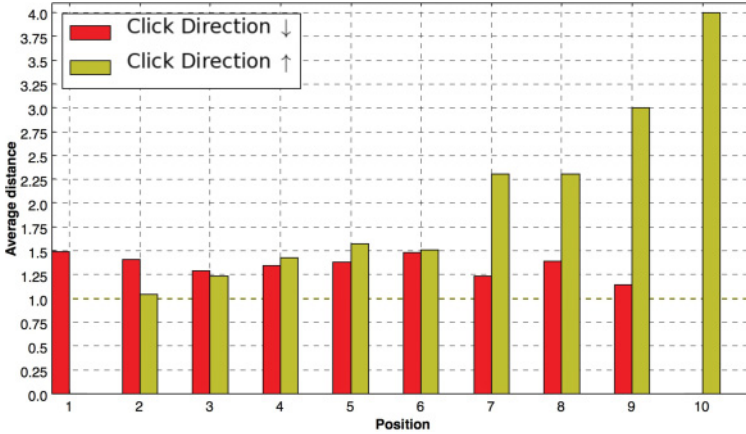


Fig. 2. Average examination transition distance according to different examination transition start positions for two types of adjacent clicks.

From this figure, we can see that regardless of whether the click direction is  $\uparrow$  or  $\downarrow$ , in most cases (72.7% for  $\downarrow$  and 78.9% for  $\uparrow$ ), the whole examination sequences follow the same direction as the click direction without any direction changes. The percentage of sequences with direction changes between  $\downarrow$  clicks is slightly larger than that between  $\uparrow$  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 whether it is  $\uparrow$  or  $\downarrow$ .

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  $k$ , whereas 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 resulting positions.



We can see that all transition distances are around 1.25 when a user follows a top-down ( $\downarrow$ ) click sequence. Meanwhile, when a user follows a bottom-up ( $\uparrow$ ) click sequence, his or her eyes may skip several results to find a specific result.

In particular, we observe larger transition distances for bottom-ranked positions, which tend to bring focus back to the middle positions (positions 5 and 6) in the list. As all of the transition distances are statistically significantly larger than 1 ( $p$ -value < 0.01 for each position and each click direction based on the  $t$ -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 or she may start examining results either in a  $\uparrow$  or a  $\downarrow$  direction. The user seldom changes the examination direction until he or she clicks another result located at position  $j$  (the locally unidirectional examination assumption), but he or she may not examine all results on the examination path (the non-first-order examination assumption).

Compared to existing sequence-based click models such as the POM, which assume that the examination sequence within two clicks can be arbitrary, 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 the THCM, which assumes that 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.

### 3.2. Click Dwell Time Analysis

Among many implicit measures, click dwell time (the time that the user spends on a clicked result) is one of the most important features because it is clearly correlated with result-level satisfaction or document relevance [Buscher et al. 2009; Smucker and Clarke 2012; Fox et al. 2005]. Longer dwell time on a clicked page has traditionally been used to identify satisfied (SAT) clicks. Although click-through statistics can sometimes be misleading owing to order and caption biases, the click dwell time is a more robust measure.

Click dwell time has been successfully used in several retrieval applications (e.g., implicit relevance feedback [White and Kelly 2006] and reranking [Agichtein et al. 2006a]). In those applications, SAT clicks are simply identified by some predefined time threshold (i.e., a click is SAT if its dwell time equals or exceeds that threshold). A dwell time equal to or exceeding 30 seconds, as proposed in Fox et al. [2005], has typically been used to identify clicks with which searchers are satisfied.

As the click dwell time is very important feedback, we want to add this information into the click model's framework. Hence, we first choose the SAT click indicator (30 seconds) as our first information gain mapping reference. However, the dwell time depends on page content and has been shown to vary based on other factors, such as the search task and the user. A more robust interpretation of click dwell time is therefore needed.

We use the click dwell time distribution in the Sogou and Yandex dataset. According to Figure 3, we can see that the dwell time distribution is very long tailed. Although in greater than half (50.4%) of situations the users spend less than 30 seconds on each click, many clicks still cost more than 100 seconds. Moreover, we can also see that

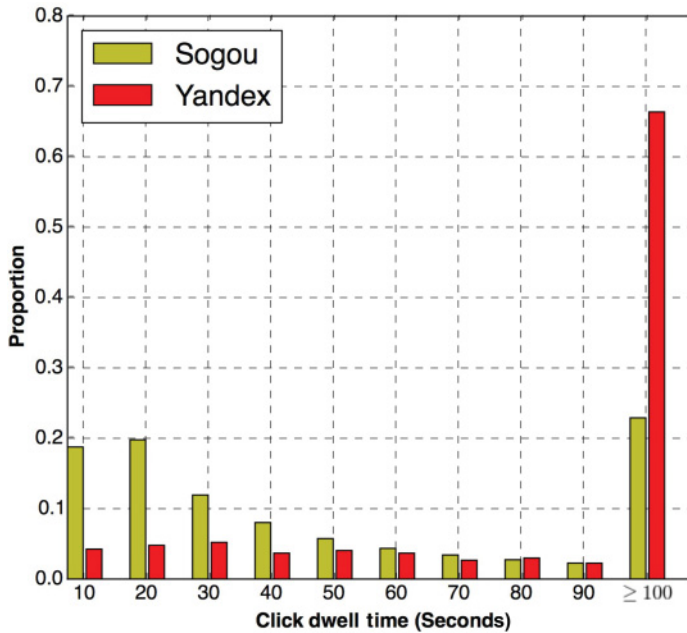


Fig. 3. Dwell time distribution in different search logs.

user behavior is very different in the Sogou and Yandex search logs. The click dwell time in the Yandex dataset tends to be much longer than in the Sogou dataset. This may be caused by the language differences, culture differences, or differences in the network environment. Although it still shows that a single dwell time threshold (e.g. 30 seconds) may not correctly indicate user satisfaction, we should take dwell time information into account in our model framework.

#### 4. TIME-AWARE CLICK MODEL

According to the previous section, we may assume that the click dwell time can provide valuable feedback information that we cannot obtain from click actions and click sequence information. An intuitive idea is that if a user prefers a certain result, he or she may stay at the corresponding landing page for a longer time than a result that he or she dislikes. Therefore, we should use click dwell time information to help us infer these preferences.

To incorporate the dwell time information into the modeling process, we inherit the assumption from the DBN [Chapelle and Zhang 2009], which assumes that when a user achieves a satisfied state, he or she will not continue the search process. We use a function that maps the click dwell time to the satisfaction state. As our goal is to test the effectiveness of introducing click dwell time information into click models, we choose to build a new click model based on some existing click models. As the PSCM [Wang et al. 2015] is a newly proposed click model that shows good performance and already takes click sequence information into account compared to other position-based click models, we choose the PSCM as our basic model to further add click dwell time information to build the TACM.

In the following sections, we first propose the PSCM and the TACM. After that, we compare these two models and introduce the inference process of the TACM.

#### 4.1. Partially Sequential Click Model

We first make some definitions and notations. Suppose that there are  $N$  sessions, each of which records certain user interactions with the top  $M$  results ( $M$  is usually set to 10 in most existing click model research). The results list can be represented as an impression sequence,  $D = \langle d_1, d_2, \dots, d_i, \dots, d_M \rangle$ , 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  $R = \langle R_1, R_2, \dots, R_i, \dots, R_M \rangle$ . With the timestamp information recorded in the logs, we organize the click sequence as  $C = \langle C_1, C_2, \dots, C_t, \dots, C_T \rangle$ , 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 the DBN and UBM. We do the same in this work. This supposes that the click event at time  $t + 1$  is only determined by the click event at time  $t$ . According to this hypothesis, a user's click action  $C = \langle C_1, C_2, \dots, C_t, \dots, C_T \rangle$  can be independently separated to  $T + 1$  adjacent click pairs:  $\langle C_0, C_1 \rangle, \dots, \langle C_{t-1}, C_t \rangle, \dots, \langle C_T, C_{T+1} \rangle$  ( $C_0$  represents the beginning of the search process, and  $C_{T+1}$  represents the end of the search process). This makes it possible for us to divide a click sequence into subsequences (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  $\langle \bar{E}_m, \dots, \bar{E}_j, \dots, \bar{E}_n \rangle$  and  $\langle \bar{C}_m, \dots, \bar{C}_j, \dots, \bar{C}_n \rangle$ , respectively. Note that in contrast to  $C_t$ , which is used to record the position of the 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 ( $=0$ ) at the corresponding result position. In addition, we can also deduce that in the click sequence, only  $\bar{C}_m$  and  $\bar{C}_n$  have values of 1 and the other positions on the path have values of 0.

The proposed PSCM adopts these two assumptions. It is then described as follows:

$$P(C_t | C_{t-1}, \dots, C_1) = P(C_t | C_{t-1}), \quad (13)$$

$$\begin{aligned} &P(C_t = n | C_{t-1} = m) \\ &= P(\bar{C}_m = 1, \dots, \bar{C}_i = 0, \dots, \bar{C}_n = 1), \end{aligned} \quad (14)$$

$$\begin{aligned} &P(\bar{E}_i = 1 | C_{t-1} = m, C_t = n) \\ &= \begin{cases} \gamma_{im}, m \leq i \leq n \text{ or } n \leq i \leq m \\ 0, \text{ other} \end{cases}, \end{aligned} \quad (15)$$

$$\bar{C}_i = 1 \Leftrightarrow \bar{E}_i = 1, R_i = 1, \quad (16)$$

$$P(R_i = 1) = \alpha_{uq}. \quad (17)$$

Equation (13) encodes the first-order click hypothesis, whereas Equation (14) 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 (15) because, according to Figure 2, the examination behavior between adjacent clicks may not follow cascade assumptions (the non-first-order examination assumption). The probability of examination depends on the positions of the clicks. This is similar to the UBM, which also allows skips, but only within sequential behavior. The PSCM also follows the examination hypothesis described in Equation (16) as in most existing click models.  $\alpha_{uq}$  corresponds to the relevance of the document URL  $u$  at position  $i$  for the specific query  $q$ .

#### 4.2. Time-Aware Click Model

To add click dwell time information, we introduce a new hidden state (satisfaction state) into this model,  $S = \langle S_0, S_1, S_2, \dots, S_t, \dots, S_T \rangle$ , where each  $S_t = 1$  represents that, after a user's  $t$ -th click, the user has already obtained enough information and prepares to finish his or her search process. This hidden state is inspired by the DBN [Chapelle and Zhang 2009], which assumes that a user may achieve a satisfaction state and stop browsing after reading some results. As Fox et al. [2005] and Smucker and Clarke [2012] showed that users are more willing to spend longer times on pages that are related and that the correlation between click dwell time and user information gain is positive, we want to use click dwell time information to represent the user's information gain.

The proposed TACM is then described as follows:

$$P(C_t | C_{t-1}, \dots, C_1, S_t - 1, \dots, S_1) = P(C_t | C_{t-1}, S_{t-1}), \quad (18)$$

$$S_{t-1} = 1 \rightarrow C_t = 0, \quad (19)$$

$$P(C_t = n | C_{t-1} = m) \quad (20)$$

$$= P(\bar{C}_m = 1, \dots, \bar{C}_i = 0, \dots, \bar{C}_n = 1),$$

$$P(\bar{E}_i = 1 | C_{t-1} = m, C_t = n)$$

$$= \begin{cases} \gamma_{im}, m \leq i \leq n \text{ or } n \leq i \leq m \\ 0, \text{ other} \end{cases}, \quad (21)$$

$$\bar{C}_i = 1 \Leftrightarrow \bar{E}_i = 1, R_i = 1, \quad (22)$$

$$P(R_i = 1) = \alpha_{uq}, \quad (23)$$

$$P(S_t = 1) = P(R_t = 1) \times F(DwellTime_t). \quad (24)$$

We can see that Equation (18) and Equation (20) still follow the first-order click hypothesis and locally unidirectional examination assumption proposed in the PSCM. In Equation (18), we also add the influence of the user's satisfaction factor, and Equation (19) shows that the user may stop the browsing process if he or she feels satisfied. The TACM also follows the examination hypothesis described in Equation (22) as in most existing click models.  $\alpha_{uq}$  corresponds to the relevance of the document URL  $u$  at position  $i$  for the specific query  $q$ .

Equation (24) describes the usage of click dwell time information. After each click, the user will obtain an information gain based on the dwell time and the result relevance. We use the following four different mapping functions.

*Linear mapping function.* According to Fox et al. [2005], a dwell time equal to or exceeding 30 seconds has typically been used to identify clicks with which searchers are satisfied. Therefore, we assume that if a user spends more than 30 seconds on a result, he or she will completely obtain the information gain from this result, and the obtaining process is linear:

$$F(DwellTime_t) = \frac{\min(DwellTime + \delta, 30 - \delta)}{30}. \quad (25)$$

Here,  $\delta > 0$  is a small positive number to make sure that the probability will never be 0, which may cause errors in logarithmic terms.

*Quadratic mapping function.* To verify whether increasing the order of the polynomial interpolation will enhance the fitting degree of mapping function, we also test the quadratic mapping function by simply squaring the linear mapping function.

*Exponential mapping function.* Smucker and Clarke [2012] proposed an exponential function to fit the time-based gain density function:

$$F(DwellTime_t) = e^{-DwellTime \times \frac{\ln 2}{h}}, \quad (26)$$

where  $h$  is the time at which half of the users have stopped scanning the result list. According to our analysis of two real-world large-scale datasets (Sogou from China and Yandex from Russia), results show that the value  $h$  for Sogou is 68.96 seconds and the value  $h$  for Yandex is 2,110.56 seconds.

*Rayleigh mapping function.* Liu et al. [2010] utilized the Weibull distribution to analyze dwell time on Web-browsing behaviors. The Rayleigh distribution is a special case of the Weibull distribution when parameter  $k$  in the Weibull distribution equals 2. The time-based gain density function of the Rayleigh distribution has the following format:

$$F(DwellTime_t) = \frac{2 \times DwellTime}{h^2} \times e^{-\left(\frac{DwellTime}{h}\right)^2}, \quad (27)$$

where  $h$  is also the time at which half of the users have stopped scanning the result list.

### 4.3. Model Inference for the TACM

According to the description of the TACM and PSCM in previous sections, we can see that the major differences of these two models are as follows. In the TACM, we try to emphasize the influence of different click dwell times on clicked results; therefore, we introduce a new group of hidden state  $S$  to represent users' degree of satisfaction (represents the probability of stopping the search process). According to existing studies [Fox et al. 2005; Smucker and Clarke 2012; Liu et al. 2010], we assume that the stopping probability is related to the result relevance ( $\alpha_{uq}$ ) and the dwell time user costs on it ( $F(DwellTime_t)$ ).

According to the definitions of different dwell time mapping functions, we do not introduce any new hidden parameters from these mapping functions. Therefore, the hidden parameters of the TACM are the same as the PSCM ( $\{\alpha_{uq}\}$  and  $\{\gamma_{imm}\}$ ). As the different dwell times will change the relevance estimation for  $\{\alpha_{uq}\}$  according to Equation (24), and the examination parameters  $\{\gamma_{imm}\}$  are global parameters shared among different results, all of these hidden parameters in the TACM will be different from the PSCM. As we do not introduce any new parameters for the TACM, we can test the effectiveness of using the click dwell time in the click model by comparing the TACM to the PSCM.

We use the expectation-maximization (EM) algorithm [Gupta and Chen 2011] to find the maximum likelihood estimate of the variables  $\{\alpha_{uq}\}$  and  $\{\gamma_{imm}\}$ . We first introduce some notations. Suppose that we have  $N$  query sessions and  $M$  results for each query,  $j$  is the  $j$ th query session in  $N$ ,  $T^j$  is the click sequence length in this session,  $d_i^j = u$  means that the  $i$ th document's URL is  $u$  in the  $j$ th query session,  $q^j = q$  means that the query is  $q$  in the  $j$ th query session,  $\bar{t}$  corresponds to the  $t$ th adjacent click pair  $\{t, C_{t-1} = m, C_t = n\}$ ,  $C_t = n$  means that the  $t$ -th click position is  $n$ ,  $I(\cdot)$  represents the indicator function,  $I_{mn}$  is the abbreviation of  $I(m \leq i \leq n \text{ or } n \leq i \leq m)$ ,  $I_{-}$  is the abbreviation for  $I(d_i^j = u, q^j = q, i = n)$ , and  $I_{\neq}$  is the abbreviation for  $I(d_i^j = u, q^j = q, i \neq n)$ . The observation of our model is the click sequence ( $Y = \{C\}$ ), the hidden variables are the query result relevance and user examination information ( $Z = \{R, E, S\}$ ), and the parameters are  $\theta = \{\alpha_{uq}, \gamma_{imm}\}$ . Therefore, given one specific query session, the marginal

likelihood is as follows:

$$\begin{aligned}
 P(Y, Z|\theta) &= P(C, E, R, S|\theta) = \prod_{t=1}^T P(C_t, E, R|C_{t-1}, \theta) P(S_{t-1} = 0|\theta) \\
 &= \prod_{t=1}^T P(C_t, E, R|C_{t-1}, \theta) \times \prod_{t=1}^T (1 - P(R_t = 1)F(DwellTime_t)).
 \end{aligned} \tag{28}$$

According to Equations (14) and (16) (omit  $\theta$  for conciseness),

$$\begin{aligned}
 &P(C_t = n, E, R|C_{t-1} = m) \\
 &= \left\{ \prod_{i=m+1}^{n-1} P(\bar{C}_i = 0|\bar{E}_i, R_i)P(R_i)P(\bar{E}_i|C_{t-1} = m, C_t = n) \right\} \\
 &\cdot \{P(\bar{C}_n = 1|R_n, \bar{E}_n)P(R_n)P(\bar{E}_n|C_{t-1} = m, C_t = n)\}.
 \end{aligned} \tag{29}$$

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

$$Q = E_{E,R,S|C,\theta^{(v)}} [\log P(C, E, R, S|\theta)]. \tag{30}$$

In iteration  $v$ , the formulation of parameter  $\alpha_{uq}$  corresponding to a specific query  $q$  and result  $u$  in the Q-function is the following:

$$\begin{aligned}
 Q_{\alpha_{uq}} &= \sum_{j=1}^N \sum_{t}^{T^{j-1}} \left\{ I_{=} \cdot \log(1 - \alpha_{uq}^{(v)} F(DwellTime)) + \sum_{j=1}^N \sum_{t}^{T^j} \{ I_{mn} \cdot [ I_{=} \cdot 1 \cdot \log(\alpha_{uq}) \right. \\
 &\quad \left. + I_{\neq} \cdot \frac{1 - \alpha_{uq}^{(v)}}{1 - \alpha_{uq}^{(v)} \gamma_{imn}^{(v)}} \cdot \log(1 - \alpha_{uq}) + I_{\neq} \cdot \frac{\alpha_{uq}^{(v)}(1 - \gamma_{imn}^{(v)})}{1 - \alpha_{uq}^{(v)} \gamma_{imn}^{(v)}} \cdot \log(\alpha_{uq}) \right] \}.
 \end{aligned} \tag{31}$$

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

$$\begin{aligned}
 Q_{\gamma_{imn}} &= \sum_{j=1}^N \sum_{t}^{T^j} \left\{ I_{mn} \cdot \left[ I_{\neq} \cdot \frac{1 - \gamma_{imn}^{(v)}}{1 - \alpha_{uq}^{(v)} \gamma_{imn}^{(v)}} \cdot \log(1 - \gamma_{imn}) \right. \right. \\
 &\quad \left. \left. + I_{\neq} \cdot \frac{\gamma_{imn}^{(v)}(1 - \alpha_{uq}^{(v)})}{1 - \alpha_{uq}^{(v)} \gamma_{imn}^{(v)}} \cdot \log(\gamma_{imn}) + I_{=} \cdot 1 \cdot \log(\gamma_{imn}) \right] \right\}.
 \end{aligned} \tag{32}$$

For  $\gamma_{imn}$  in Equation (32), we can take the derivative and generate the corresponding updating formulation in iteration round  $(v)$ .

By separately taking the derivatives of  $\alpha_{uq}$  in Equation (31) and  $\gamma_{imn}$  in Equation (32), we can generate the corresponding updating formulation for  $\alpha_{uq}^{(v+1)}$  and  $\gamma_{imn}^{(v+1)}$  in iteration

round ( $v$ ):

$$\begin{aligned}
G_1^{(v)} &= \sum_{j=1}^N \sum_{\bar{t}}^{T^j} \left\{ I_{mn} \cdot I_{\neq} \cdot \frac{1 - \gamma_{imn}^{(v)}}{1 - \alpha_{uq}^{(v)} \gamma_{imn}^{(v)}} \right\} \\
G_2^{(v)} &= \sum_{j=1}^N \sum_{\bar{t}}^{T^j} \left\{ I_{mn} \cdot I_{\neq} \cdot \frac{\gamma_{imn}^{(v)} (1 - \alpha_{uq}^{(v)})}{1 - \alpha_{uq}^{(v)} \gamma_{imn}^{(v)}} \right\} \\
G_3^{(v)} &= \sum_{j=1}^N \sum_{\bar{t}}^{T^j} \{ I_{mn} \cdot I_{=} \} \\
\gamma_{imn}^{(v+1)} &= \frac{G_2^{(v)} + G_3^{(v)}}{G_1^{(v)} + G_2^{(v)} + G_3^{(v)}}.
\end{aligned} \tag{33}$$

Meanwhile, for  $\alpha_{uq}$  in Equation (31), as for Equation (31), it can be written as follows:

$$\begin{aligned}
Q_{\alpha_{uq}} &= \sum_i a_i \log(\alpha_{uq}) + \sum_i b_i \log(1 - \alpha_{uq}) + \sum_i c_i \log(1 - F(DwellTime)\alpha_{uq}) \\
&= \sum_i w_i \times f(\alpha_{uq}).
\end{aligned} \tag{34}$$

We can use the stochastic gradient descent method to find the updating value. The initial value is the close-formed updating formulation of  $a_i$  and  $b_i$ :

$$\begin{aligned}
A_1^{(v)} &= \sum_{j=1}^N \sum_{\bar{t}}^{T^j} \left\{ I_{mn} \cdot I_{\neq} \cdot \frac{1 - \alpha_{imn}^{(v)}}{1 - \alpha_{uq}^{(v)} \gamma_{imn}^{(v)}} \right\} \\
A_2^{(v)} &= \sum_{j=1}^N \sum_{\bar{t}}^{T^j} \left\{ I_{mn} \cdot I_{\neq} \cdot \frac{\alpha_{imn}^{(v)} (1 - \gamma_{uq}^{(v)})}{1 - \alpha_{uq}^{(v)} \gamma_{imn}^{(v)}} \right\} \\
A_3^{(v)} &= \sum_{j=1}^N \sum_{\bar{t}}^{T^j} \{ I_{mn} \cdot I_{=} \} \\
\alpha_{uq}^{initial} &= \frac{A_2^{(v)} + A_3^{(v)}}{A_1^{(v)} + A_2^{(v)} + A_3^{(v)}} \\
\alpha_{uq}^{v+1} &= \alpha_{uq}^{initial} - \eta \sum_i \nabla(w_i \times f(\alpha_{uq})).
\end{aligned} \tag{35}$$

The update formulas obtained for the TACM are exactly the same as for the PSCM, except for the  $\alpha_{uq}$  parameter. This is because we only make the assumption that different click dwell times only show influence on result relevance estimation. Therefore, the examination sequence is the same as the PSCM. However, according to Equation (33), the  $\gamma_{imn}$  will be different from the PSCM because the  $\alpha_{uq}$  parameters are different in these two click models.

## 5. EXPERIMENTS

To test the effectiveness of the proposed TACM, we compared its performance to several existing click models for click prediction and relevance estimation. In addition to our basic PSCM [Wang et al. 2015], we also chose some popular position-based click models

(the UBM [Dupret and Piwowarski 2008] and DBN [Chapelle and Zhang 2009]) and sequence-based click models (the POM [Wang et al. 2010], THCM [Xu et al. 2012], and TCM [Xu et al. 2010]) as our baselines.

We performed two types of experiments to validate our model. We evaluated the click model in terms of predicting click probabilities (click perplexity) from search logs and used the predicted relevance as a signal for document ranking, and we evaluated each click model's ranking performance with traditional IR metrics (in this article, we use the NDCG metric [Järvelin and Kekäläinen 2002]).

### 5.1. Experimental Setup

As described earlier, we applied the same method here to address the limitations of the TCM and POM to adapt them for performance comparison. As for other baseline models, we refer to the implementations from Chuklin et al. [2013]. Our own implementation of the PSCM and TACM can be found at <http://www.thuir.cn/group/yqliu>, as well as a sample of the experimental dataset.

#### 5.1.1. Baseline Model Adaptation.

*Temporal click model.* As mentioned in Section 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 (five possible situations for two results [Xu et al. 2010]), it faces an exponential explosion problem when the number of results becomes large. Therefore, we cannot expand this model to  $M$  results in one SERP ( $M$  equals 10 in our dataset). To compare this model to other existing click models that can handle an arbitrary number of results in a SERP, we made a trivial expansion of the TCM: we separated these results into  $M/2$  pairs  $((1, 2), (3, 4), \dots, (M-1, M))$  and implemented the TCM for each pair separately. Then, from each pair, we can deduce the two results' relevance and click probabilities. We therefore combine  $M/2$  pairs together to generate click prediction and relevance prediction for the whole results list.

*Partially observable Markov model.* Although the POM can model nonsequential 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 Section 2. It is unfair to compare the POM with other models. To make the POM more suitable for click and relevance prediction tasks, we modified the original POM by setting a relevance score for each specific document-query pair.

According to search logs, clicks can be reorganized as a temporal sequence of behaviors by recorded time stamps,  $E = \langle E_1, E_2, \dots, E_t, \dots, E_T \rangle$ , where  $t$  represents the events' relative order and  $E_t$  represents the corresponding ranking of the result being examined at time  $t$ , and  $C = \langle C_1, C_2, \dots, C_t, \dots, C_T \rangle$ , where  $C_t$  represents whether the corresponding result is clicked or not. From search logs, we can only observe which results were clicked by users. Based on the assumption that a user must examine a result before clicking on it (the examination hypothesis [Craswell et al. 2008]), we can infer that the clicked results must have been examined. Therefore, a user may examine some results in his or her browsing process but not click them given a click sequence observation  $O = \{(E_1 = e_1, C_1 = 1), \dots, (E_T = e_T, C_T = 1)\}$ . Thus, an arbitrary  $O' = \{(E'_1, C'_1), \dots, (E'_k, C'_k), \dots, (E'_K, C'_K)\}$  can be generated based on the original observation  $O$ , where  $O \subseteq O'$ . The POM assumes that the probability of original observation is the summation of the probabilities of all compatible examination sequences. Furthermore, the POM makes the first-order assumption that the currently examined result only depends on previous examinations. Therefore, the POM can be represented as follows:

$$P(O) = \sum_{O'} P(O') = \sum_{O'} \prod_{i=1}^K P(C_i | E_i) P(E_i | E_{i-1}), \quad (36)$$



$$P(C_i = 1|E_i = m) = c_m, \quad (37)$$

$$P(E_i = n|E_{i-1} = m) = e_{mn}, \quad (38)$$

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

However, in the original POM, given the examination of a result, the click probability is only dependent on the result position (Equation (37)). 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, Equation (37) is revised as such:

$$P(C_i = 1|E_i = m) = \alpha_{uq}, \quad (39)$$

where  $\alpha_{uq}$  is the relevance of a query-document pair. Thus, the click probability no longer depends on the rank position but depends on the search query. Once we obtain  $\alpha_{uq}$ , we can compare the POM to other click models in terms of click perplexity and NDCG. The parameter estimation formulation is made similar to the original model [Wang et al. 2010] by using the EM algorithm.

The estimation formula for the iteration process ( $v + 1$ ) is as follows:

$$e_{mm}^{(v+1)} = \frac{\sum_q \sum_{qs} \sum_{O'} P(O'|\Lambda^{(v)}) \sum_i I(E_{i+1} = n, E_i = m)}{\sum_q \sum_{qs} \sum_{O'} P(O'|\Lambda^{(v)}) \sum_i I(E_i = m)}, \quad (40)$$

$$\alpha_{qu}^{(v+1)} = \frac{\sum_{qs} \sum_{O'} P(O'|\Lambda^{(v)}) \sum_i I(E_i = m, C_i = 1, d_i = u)}{\sum_{qs} \sum_{O'} P(O'|\Lambda^{(v)}) \sum_i I(E_i = m, d_i = u)}, \quad (41)$$

where  $\Lambda^{(v)}$  represents the parameter for iteration process ( $v$ ),  $qs$  is the list of corresponding sessions of query  $q$ , and  $I(\cdot)$  is the indicator function.

**5.1.2. Datasets.** To show the effectiveness of the proposed click models, we utilize two real-world large-scale datasets collected by Sogou from China and Yandex<sup>3</sup> from Russia. The detailed statistics for the two datasets can be found in Table I. To better examine the value of dwell time information, we filtered the data sessions without clicks or those that contained only a single click. Please note that the datasets used here are not the same as those in Wang et al. [2015], as the dataset in the PSCM paper [Wang et al. 2015] does not contain click dwell time information. According to the following experimental results, the major findings are the same as those obtained based on the previous datasets.

## 5.2. Evaluation of Click Prediction

As in the experiments for the TACM, we used two search logs (see Table I) to compute the click perplexity of each model. For each dataset, we split all query sessions into training and testing sets in a ratio of 70% to 30%, as did many previous studies (e.g., Chen et al. [2012] and Wang et al. [2013]).

<sup>3</sup>The Yandex dataset is publicly available at <https://www.kaggle.com/c/yandex-personalized-Web-search-challenge/data>.

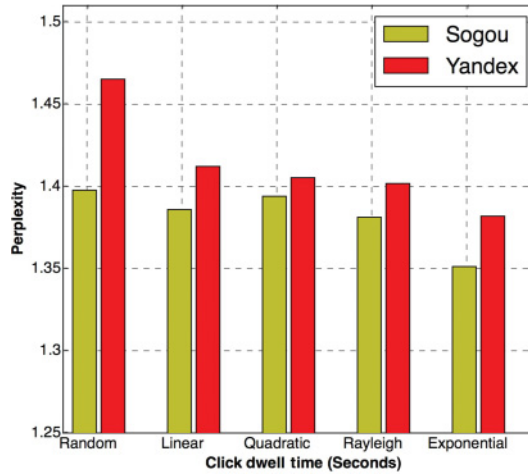


Fig. 4. Click perplexity of different dwell time mapping functions on Data-C and Data-Y.

*5.2.1. The TACM With Different Mapping Functions.* We first want to investigate which dwell time mapping function is more suitable for the TACM. Therefore, we implemented the linear mapping function, quadratic mapping function, Rayleigh mapping function, and exponential mapping function described in the previous section. To test whether the dwell time information was actually useful or not, we also implement a random mapping function that randomly generates an information gain value no matter what the dwell time is.

The results are shown in Figure 4. We can see that the linear mapping function, quadratic mapping function, Rayleigh mapping function, and exponential mapping function are better than the random mapping function. Therefore, adding dwell time information as a positive correlation with user satisfaction can actually improve the model's performance. We can also find that the exponential mapping function performs best among all mapping functions. According to our statistics in Section 3, the click dwell time in different situations varies significantly. Therefore, using a fixed threshold (e.g., 30 seconds) in the mapping function (as in linear and quadratic functions) may not be a good idea. It accords with the conclusions in Kim et al. [2014] that in different cases, users need different amounts of time to be satisfied with result clicks. Therefore, we choose the exponential mapping function in the following experiments.

*5.2.2. Overall Comparison.* After choosing the proper dwell time mapping function, we compared performances of the TACM with other existing models in the Sogou and Yandex datasets.

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

Table I. Two Large-Scale Commercial Search Logs (Different Languages) Used to Evaluate the Proposed Click Models

Data	Data-C	Data-Y
Description	Sogou's logs	Yandex's logs
#Distinct Queries	149,947	2,643,339
#Sessions	3,431,378	5,999,999
Experiments	Click perplexity NDCG	Click perplexity

Note: The number sign (#) stands for "Number of."

Table II. Overall Click Perplexity of Each Model on Data-C and Data-Y

Model	Data-C	TACM Improvement	Data-Y	TACM Improvement
TACM	1.346	—	1.382	—
PSCM	1.477	+27.5%	1.428	+10.7%
UBM	1.562	+38.4%	1.611	+37.8%
DBN	1.593	+41.7%	1.670	+43.0%
POM	2.174	+70.5%	1.876	+56.4%
THCM	2.040	+66.7%	2.121	+65.9%
TCM	3.156	+84.0%	3.833	+86.5%

Note: All improvements are statistically significant according to the  $t$ -test with a  $p$ -value  $< 10^{-5}$ .

Table III. Click Perplexity of Each Model for Different Query Frequencies (Data-C)

Model	(0, 10]	TACM Improvement	(10, 100]	TACM Improvement	(100, inf)	TACM Improvement
TACM	1.361	—	1.332	—	1.391	—
PSCM	1.396	+8.8%	1.466	+28.8%	1.545	+28.3%
UBM	1.752	+52.0%	1.595	+44.2%	1.560	+30.2%
DBN	2.024	+64.7%	1.634	+47.6%	1.585	+33.2%
POM	2.870	+80.7%	2.281	+74.1%	2.268	+69.2%
THCM	2.813	+80.1%	2.266	+73.8%	2.089	+64.1%
TCM	5.216	+91.4%	3.596	+87.2%	3.520	+84.5%

better coped with in our model. We can also see that by adding dwell time information, the TACM can better predict click action than the PSCM.

**5.2.3. Comparison for Different Query Frequencies.** In addition to the the overall comparison, we also compared different models for different query frequencies. We separated queries in our dataset into three groups according to the query appearance count: a low-frequency group, middle-frequency group, and high-frequency group. Table III and Table IV shows the performance of different models. We can see that popular click models such as the UBM perform better when query frequency increases, whereas the PSCM performs better for low-frequency queries. Although the PSCM's performance decreases for high-frequency queries, it is still better than other popular models. In contrast to these models, the TACM performs best for middle-frequency queries. The standard deviation of the TACM at different query frequencies was only 0.03, whereas it was 0.08 for the PSCM and 0.20 for the UBM. Therefore, the TACM's performance is much more stable than other click models. This can be explained as follows. For low-frequency queries, since the data amount is very small, the amount of click dwell time data points may not be sufficient to reveal the actual user preference on different results. As the query frequency becomes higher and higher, the amount becomes statistically significant for user preference estimation. Therefore, the TACM performs well for middle-frequency queries. However, for high-frequency queries, since the amount of data is so sufficient that click information is enough for estimating user preference,

Table IV. Click Perplexity of Each Model for Different Query Frequencies (Data-Y)

Model	(0, 10]	TACM Improvement	(10, 100]	TACM Improvement	(100, inf)	TACM Improvement
TACM	1.415	—	1.379	—	1.455	—
PSCM	1.400	-3.8%	1.466	+18.6%	1.562	+19.0%
UBM	1.931	+55.4%	1.590	+35.7%	1.545	+16.6%
DBN	2.432	+71.0%	1.628	+39.6%	1.570	+20.2%
POM	3.567	+83.8%	1.860	+55.9%	1.787	+42.2%
THCM	3.335	+82.2%	2.269	+70.1%	2.069	+57.4%
TCM	7.450	+93.6%	4.181	+88.1%	4.020	+84.9%

Table V. Click Perplexity of Each Model for Different Query Lengths (Data-C)

Model	(0, 2]	TACM Improvement	(2, 4]	TACM Improvement	(4, 6]	TACM Improvement	(6, inf)	TACM Improvement
TACM	1.350	—	1.359	—	1.358	—	1.349	—
PSCM	1.397	+11.8%	1.409	+12.2%	1.414	+13.5%	1.410	+14.9%
UBM	1.671	+47.8%	1.725	+50.5%	1.753	+52.5%	1.728	+52.1%
DBN	1.890	+60.7%	1.955	+62.4%	1.990	+63.8%	1.975	+64.2%
POM	2.776	+80.3%	2.767	+79.7%	2.762	+79.7%	2.784	+80.4%
THCM	2.658	+78.9%	2.706	+79.0%	2.757	+79.6%	2.741	+80.0%
TCM	5.013	+91.3%	4.927	+90.6%	4.901	+90.8%	4.943	+91.1%

the improvement of the TACM is lower than the result in the middle-frequency query situation.

*5.2.4. Comparison for Different Query Lengths.* We also compared different models for different query lengths. In the Yandex dataset, they used query ID to identify a specific query rather than the query itself. Therefore, we conducted this experiment only in the Sogou dataset. We separated queries in our dataset into four groups according to the query lengths: (0,2], (2,4], (4,6] , (6, inf). Table V illustrates the performance of different models. We can observe that the TACM’s performance is much better than other click models at all query lengths. The standard deviation of the TACM at all query lengths was only 0.004, whereas it was 0.007 for the PSCM and 0.035 for the UBM. This result also proved that the TACM’s performance is much more stable. We also find that the performance of our proposed model increases with query lengths. This may be explained as follows. Users’ search intent becomes more complex with increasing query length, and therefore the dwell time information becomes more important in revealing the result preference than click information.

### 5.3. Evaluation of Relevance Estimation

As a click model also provides a prediction of the relevance of a document for a query,  $\alpha_{uq}$ , we can rank documents according to this value. The ranking results can be measured using NDCG [Järvelin and Kekäläinen 2002]. This evaluation was performed only on Data-C, for which human evaluators could be recruited to judge document relevance. The same evaluation cannot be done on Data-Y, as the data has been encoded as unreadable code and no relevance information is available.

For a random sample of 600 queries in Data-C, several professional assessors (from Sogou.com, without knowing any information about this work) annotated several results’ relevance scores for each query. The annotation was performed with five grades (“Perfect,” “Excellent,” “Good,” “Fair,” and “Bad”) as in most existing studies, such as Yang et al. [2010]. Majority voting was adopted to decide on the relevance score if there were conflicts (at least three assessors were involved in each query-result pair annotation). Due to limited human resources, the top five results for 345 queries were annotated, whereas only the top three results were annotated for the other 540 queries.

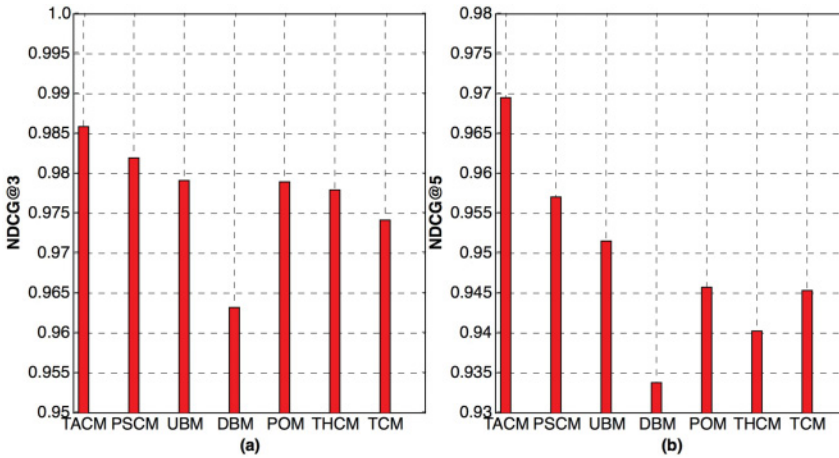


Fig. 5. Relevance estimation performance in terms of NDCG@3 and NDCG@5 for Data-C (all differences are statistically significant ( $p$ -value  $< 0.05$ ) according to paired  $t$ -tests).

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.

From Figure 5, we can see that the PSCM achieves better performance than the UBM and DBN. This result is consistent with our previous experimental results in Wang et al. [2015]. We can also see that the TACM achieves even better performance than the PSCM. This result shows that by properly incorporating click dwell time information, we can generate more accurate relevance estimation. Furthermore, we can see that the performance difference is more obvious in NDCG@5. This shows that our model predicts much more accurate relevance than other models in lower positions. This may be because the amount of user click action in lower positions is much lower than in top positions, and in this situation, the click dwell time may be a more reliable signal of relevance compared to the user click count.

## 6. CONCLUSIONS AND FUTURE WORK

In this article, we address the problem of properly incorporating temporal 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 the non-first-order examination assumption. We also made an analysis of user click dwell time in different search logs. Based on our findings, we proposed a new click model named *TACM*, which incorporates both click dwell time information and nonsequential click behaviors into click models while following the two assumptions on the 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 a test on query-result relevance estimation. The experimental results also showed that the *TACM* outperforms existing models in relevance evaluation tasks.

This study shows the importance for a click model to correctly cope with users' interactions. Compared to previous models, the assumptions made in our model are more realistic and correspond better to observations from practice. Our experimental results show that different click dwell times among click actions indicate different kinds of feedback information. A longer dwell time represents that users are willing to spend more time on the search result and may further indicate that the search result contains more useful information. We also find that the law of diminishing marginal

utility exists in a search environment, as the negative exponential mapping function performs best among several different mapping functions.

The proposed model can be further improved in several aspects. As different search users may follow different behavior patterns, we plan to add factors that can tell the difference between different users and thus make our model more personalized. Meanwhile, we will try to improve the dwell time mapping function to make it more adaptable to scenarios with different search intents. Furthermore, with more and more multimodal content incorporated into search interfaces, SERPs become more and more heterogeneous. We plan to extend the TACM to model user behaviors in a heterogeneous search environment. In addition, we believe that the proposed model can be extended to model user interaction behaviors other than clicks in Web search scenarios (hover, scroll, etc.)

## REFERENCES

- Eugene Agichtein, Eric Brill, and Susan Dumais. 2006a. Improving Web search ranking by incorporating user behavior information. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*. ACM, New York, NY, 19–26.
- Eugene Agichtein, Eric Brill, Susan T. Dumais, and Robert Ragno. 2006b. Learning user interaction models for predicting Web search result preferences. In *Proceedings of the 29th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*. ACM, New York, NY, 3–10.
- Alexey Borisov, Ilya Markov, Maarten de Rijke, and Pavel Serdyukov. 2016a. A context-aware time model for Web search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16)*. ACM, New York, NY, 1–10.
- Alexey Borisov, Ilya Markov, Maarten de Rijke, and Pavel Serdyukov. 2016b. A neural click model for Web search. In *Proceedings of the 25th International Conference on World Wide Web (WWW'16)*. 531–541.
- Andrei Broder. 2002. A taxonomy of Web search. *ACM SIGIR Forum* 36, 2, 3–10.
- Georg Buscher, Ludger van Elst, and Andreas Dengel. 2009. Segment-level display time as implicit feedback: A comparison to eye tracking. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*. ACM, New York, NY, 67–74.
- Georg Buscher, Ryan W. White, Susan Dumais, and Jeff Huang. 2012. Large-scale analysis of individual and task differences in search result page examination strategies. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM'12)*. ACM, New York, NY, 373–382.
- Olivier Chapelle and Ya Zhang. 2009. A dynamic Bayesian network click model for Web search ranking. In *Proceedings of the 18th International Conference on World Wide Web (WWW'09)*. 1–10.
- Danqi Chen, Weizhu Chen, Haixun Wang, Zheng Chen, and Qiang Yang. 2012. Beyond ten blue links: Enabling user click modeling in federated Web search. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM'12)*. 463–472.
- Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. Click models for zeeb search. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 7, 3, 1–115.
- Aleksandr Chuklin, Pavel Serdyukov, and Maarten De Rijke. 2013. Click model-based information retrieval metrics. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*. ACM, New York, NY, 493–502.
- Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM'08)*. ACM, New York, NY, 87–94.
- Edward Cutrell and Zhiwei Guan. 2007. What are you looking for? An eye-tracking study of information usage in Web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'07)*. ACM, New York, NY, 407–416.
- Georges Dupret and Ciya Liao. 2010. A model to estimate intrinsic document relevance from the clickthrough logs of a Web search engine. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10)*. ACM, New York, NY, 181–190.
- Georges Dupret and Benjamin Piwowarski. 2008. A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*. ACM, New York, NY, 331–338.

- Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems* 23, 2, 147–168.
- Laura A. Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'04)*. ACM, New York, NY, 478–479.
- Fan Guo, Chao Liu, Anitha Kannan, Tom Minka, Michael J. Taylor, Yi Min Wang, and Christos Faloutsos. 2009b. Click chain model in Web search. In *Proceedings of the 18th International Conference on World Wide Web (WWW'09)*. 11–20.
- Fan Guo, Chao Liu, and Yi Min Wang. 2009a. Efficient multiple-click models in Web search. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM'09)*. 124–131.
- Qi Guo, Dmitry Lagun, Denis Savenkov, and Qiaoling Liu. 2012. Improving relevance prediction by addressing biases and sparsity in Web search click data. In *Proceedings of the 2012 Workshop on Web Search Click Data (WSCD'12)*. 71–75.
- Maya R. Gupta and Yihua Chen. 2011. *Theory and Use of the EM Algorithm*. Now Publishers, Delft, Netherlands.
- Jeff Huang, Ryen W. White, and Susan Dumais. 2011. No clicks, no problem: Using cursor movements to understand and improve search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*. ACM, New York, NY, 1225–1234.
- Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 4, 422–446.
- Thorsten Joachims, Laura A. Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting click through data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)*. ACM, New York, NY, 154–161.
- Youngho Kim, Ahmed Hassan, Ryen W. White, and Imed Zitouni. 2014. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM'14)*. ACM, New York, NY, 193–202.
- Chao Liu, Ryen W. White, and Susan Dumais. 2010. Understanding Web browsing behaviors through Weibull analysis of dwell time. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. ACM, New York, NY, 379–386.
- Yiqun Liu, Chao Wang, Ke Zhou, Jianyun Nie, Min Zhang, and Shaoping Ma. 2014. From skimming to reading: A two-stage examination model for Web search. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM'14)*. ACM, New York, NY, 849–858.
- Lori Lorigo, Bing Pan, Helene Hembrooke, Thorsten Joachims, Laura A. Granka, and Geri Gay. 2006. The influence of task and gender on search and evaluation behavior using Google. *Information Processing and Management* 42, 4, 1123–1131.
- Rory Navalpakkam, Vidhya, Sujith Ravi, Amr Ahmed, and Alex Smola. 2013. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. 953–964.
- Keith Rayner. 2009. Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology* 62, 8, 1457–1506.
- Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research and Applications*. ACM, New York, NY, 71–78.
- Mark D. Smucker and Charles L. A. Clarke. 2012. Time-based calibration of effectiveness measures. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*. ACM, New York, NY, 95–104.
- Chao Wang, Yiqun Liu, Meng Wang, Ke Zhou, Jian-Yun Nie, and Shaoping Ma. 2015. Incorporating non-sequential behavior into click models. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'15)*. ACM, New York, NY, 283–292.
- Chao Wang, Yiqun Liu, Min Zhang, Shaoping Ma, Meihong Zheng, Jing Qian, and Kuo Zhang. 2013. Incorporating vertical results into search click models. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*. ACM, New York, NY, 503–412.
- Kuansan Wang, Nikolas Gloy, and Xionglong Li. 2010. Inferring search behaviors using partially observable Markov (POM) model. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10)*. 211–220.

- Ryen W. White and Diane Kelly. 2006. A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. ACM, New York, NY, 297–306.
- Danqing Xu, Yiqun Liu, Min Zhang, Shaoping Ma, and Liyun Ru. 2012. Incorporating revisiting behaviors into click models. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM'12)*. ACM, New York, NY, 303–312.
- Wanhong Xu, Eren Manavoglu, and Erick Cantu-Paz. 2010. Temporal click model for sponsored search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. ACM, New York, NY, 106–113.
- Hui Yang, Anton Mityagin, Krysta M. Svore, and Sergey Markov. 2010. Collecting high quality overlapping labels at low cost. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. ACM, New York, NY, 459–466.
- Yisong Yue, Rajan Patel, and Hein Roehrig. 2010. Beyond position bias: Examining result attractiveness as a source of presentation bias in click through data. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. 1011–1018.
- Yuyu Zhang, Hanjun Dai, Chang Xu, Jun Feng, Taifeng Wang, Jiang Bian, Bin Wang, and Tie-Yan Liu. 2014. Sequential click prediction for sponsored search with recurrent neural networks. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*.

Received February 2016; revised August 2016; accepted August 2016