

# Enhancing click models with mouse movement information

Zeyang Liu<sup>1</sup> · Jiaxin Mao<sup>2</sup> · Chao Wang<sup>2</sup> · Qingyao Ai<sup>3</sup> ·  
Yiqun Liu<sup>2</sup> · Jian-Yun Nie<sup>4</sup>

Received: 27 March 2016 / Accepted: 21 December 2016  
© Springer Science+Business Media New York 2017

**Abstract** User interactions in Web search, in particular, clicks, provide valuable hints on document relevance; but the signals are very noisy. In order to better understand user click behaviors and to infer the implied relevance, various click models have been proposed, each relying on some hypotheses and involving different hidden events (e.g. examination). In almost all the existing click models, it is assumed that clicks are the only observable evidence and the examinations of documents are deduced from it. However, with an increasing number of embedded heterogeneous components (e.g. verticals) on Search Engine Result Pages, click information is not sufficient to draw a complete picture of process of user examination, especially in federated search scenario. In practice, we can also collect mouse movement information, which has proven to have a strong correlation with examination. In this paper, we propose to incorporate mouse movement information into existing click models to enhance the estimation of examination. The enhanced click models are shown to have a better ability to predict both user clicks and document relevance, than the original models. The collection of mouse movement information has been implemented in a commercial search engine, showing the feasibility of the approach in practice.

---

✉ Zeyang Liu  
liuzeyang0001@gmail.com

Qingyao Ai  
aiqingyao@gmail.com

Yiqun Liu  
yiqunliu@tsinghua.edu.cn

Jian-Yun Nie  
nie@iro.umontreal.ca

<sup>1</sup> Department of Computer Technology and Applications, Qinghai University, Qinghai, China

<sup>2</sup> Department of Computer Science and Technology, Tsinghua University, Beijing, China

<sup>3</sup> College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA, USA

<sup>4</sup> Department of Computer Science and Operations Research, University of Montreal, Montreal, Canada

**Keywords** Mouse movement · Click model · Search engine · Federated search

## 1 Introduction

User interactions with a search engine provides valuable feedback useful for improving document ranking. In particular, user click-through is widely used as implicit relevance feedback, which is, however, very noisy. In order to understand if and how much a user click on a result document implies true relevance, one has to take into account different factors, in addition to document relevance, that may affect user click behaviors. For example, a document ranked at the top position naturally has a higher probability to be clicked, regardless to its relevance. There is thus an acute need to understand how likely a click could mean relevance.

To this end, a number of click models have been proposed (Craswell et al. 2008; Guo et al. 2009a; Chapelle and Zhang 2009), which usually involve additional events (e.g. examination) and different assumptions. For example, Cascade model (Craswell et al. 2008) supposes that users examine results one by one sequentially and leave immediately after clicking a certain document. Dependency Click Model (DCM, Guo et al. 2009a) extends the Cascade Model by allowing users to return to SERPs after result clicks. Dynamic Bayesian Network Model (DBN, Chapelle and Zhang 2009) takes into account not only the position factor, but also the presentation bias. In addition to the assumptions on how users behave on SERPs, a click is generally assumed to subsume examination. According to the commonly used *Examination Hypothesis* (Richardson et al. 2007), one result is clicked if and only if it is both examined and relevant. With this hypothesis, we can infer the relevance of search results and examination behavior based on click behavior on SERPs. Several previous studies proposed to take into account more factors such as user preference (Xing et al. 2013), presentation styles (Wang et al. 2013) or query intents (Hu et al. 2011). However, these models inevitably become more complex and one has to infer more hidden variables (preference, query intent, etc.) from clicks, which is often challenging to do correctly in practice.

The quality of a click model strongly depends on the observed information used for its construction. Most existing click models assume that clicks are the only information available. All the hidden variables and events are induced from click events. If more observed information is used, one can expect to obtain a better click model. In particular, eye-tracking information can reveal the examination behavior of users, and when incorporated into click model construction, can improve the quality of the latter. Unfortunately, in a realistic web search situation, eye-tracking information is unavailable.<sup>1</sup> On the other hand, mouse movement data can be easily collected at large scale. Several recent studies (Buscher et al. 2009; Huang et al. 2011, 2012a; Rodden et al. 2007; Liu et al. 2014) showed that mouse movements strongly correlate with eye fixations and examination. According to strong eye-mind hypothesis (Just and Carpenter 1980) which supposes that what the eyes fixate on is what the mind processes, the eye fixation behavior has close relationship with the result examination behavior. This motivates us to construct a click model by incorporating mouse movements as proxy of eye fixations and examinations, into click models.

<sup>1</sup> Although there exists several relatively inexpensive solutions for eye-tracking devices, they still require a complicated calibration process each time users are engaged and thus not applicable.

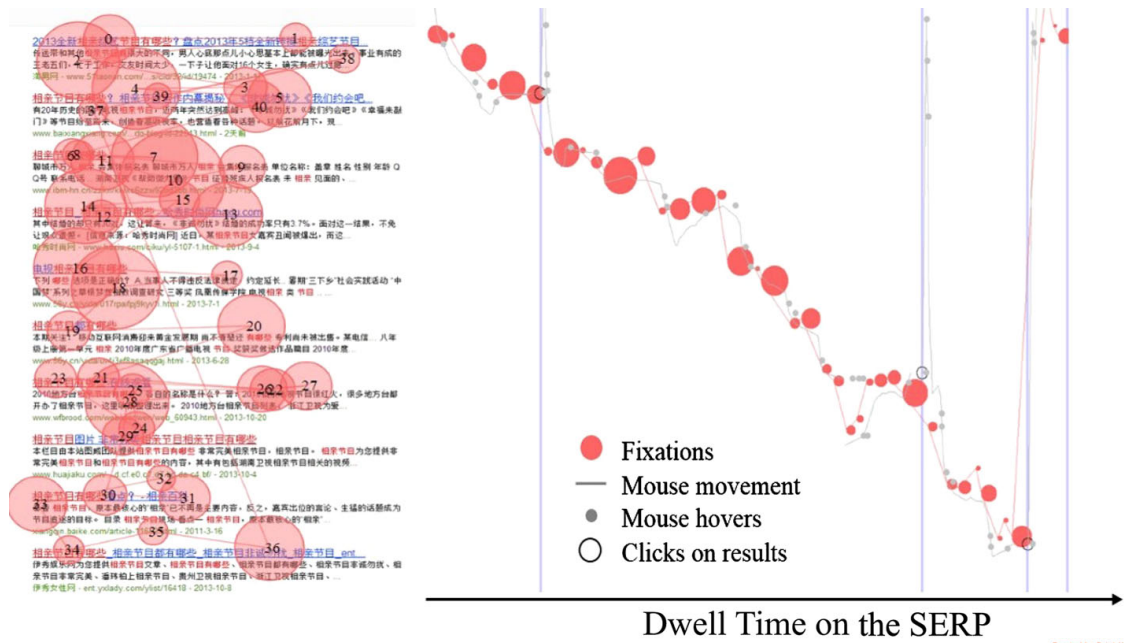
An example of the strong correlation between mouse movements and eye fixations in Web search process can be seen in Fig. 1, which shows a typical example of what we observed on a human subject for the query “TV Dating shows” (see Sect. 3 for more details). The left side shows the sequence of eye fixation on the SERP and the right side provides the corresponding positions of fixations and mouse hovers/clicks. The x-axis represents the dwell time of the session; while the y-axis represents the vertical coordinates of the mouse and the eyes on the screen. As shown in this figure, the subject’s examining process does not always follow the cascade examination hypothesis.<sup>2</sup> However, the vertical coordinates of mouse positions and eye fixations are highly correlated throughout the search session. This example (and our more complete observations in Sect. 3) shows that user examinations could largely be predicted by mouse movements.

Recently, mouse movement information has been used in a number of studies for predicting result relevance (Speicher et al. 2013; Lagun et al. 2014) or search success (Guo et al. 2012). A number of behavior features and mouse movement patterns [a.k.a. motifs (Lagun et al. 2014; Liu et al. 2015a)] are also proposed in these studies. Our work differs from these efforts in that we focus on leveraging mouse movements to predict examinations in click models. Although mouse movements provide us with more information, they may suffer from data sparsity problems because users usually only focus on top search results (Joachims et al. 2005) and there may not be enough interactions with the lower-ranked ones. A combination of user behavior assumptions and eye-tracking evidences in the framework of click model may help us to benefit from both sides and generate better estimation of user behaviors as well as result relevances. Considering the fact that more and more heterogeneous components (e.g. verticals) are incorporated into SERPs and users’ behavior patterns are largely affected by them (Liu et al. 2015b; Wang et al. 2013), it remains unclear how we can make good use of both the existing behavior assumptions and mouse movement features to improve the ranking of search results in a heterogeneous environment. To shed light on this, we conducted both a lab-based eye-tracking study to confirm the correlation between mouse movements and eye fixations, and a large-scale practical search experiment with real search logs to show that our new click model can better account for user interactions and yield better relevance prediction. To our best knowledge, we are among the first to incorporate mouse movement information into click models and validate its effectiveness in practical Web search situation.

Our contributions in this paper are threefold: (1) we propose a model to predict user examinations of results in a heterogeneous search environment using mouse movements; (2) we propose to enhance existing click models by combining the above examination prediction and existing behavior assumptions; (3) a large-scale data set containing both click-through and mouse movement behavior information is collected with the help of a popular commercial search engine; and the effectiveness of the enhanced models are validated on this data set.

The rest of this paper is organized as follows. In Sect. 2, we present some related works on the construction of click models and the analysis of eye/mouse coordination. In Sect. 3, we introduce the data collection process. After that, the examination prediction method based on mouse movement features is presented in Sect. 4. We then describe the details of the proposed click model in Sect. 5 and present the results of large-scale search ranking experiments in Sect. 6. Finally, we present some conclusions and future research directions.

<sup>2</sup> According to cascade assumption (Craswell et al. 2008), users should examine search results from top to bottom one by one sequentially.



**Fig. 1** Mouse/eye tracking results on a SERP for a user after submitting a query “TV Dating shows”

## 2 Related work

Our work is related to three families of prior research: (1) eye/mouse movements on SERPs, (2) user behavior studies in heterogeneous environment and (3) studies on the construction of click models.

### 2.1 Eye/mouse movements on SERPs

The primary goal of most existing studies on eye/mouse movement behaviors on SERPs is to detect the user’s attention allocation mechanism in Web search. Granka et al. (2004) and Joachims et al. (2005) were among the first to investigate the influence of result positions on user’s attention allocation based eye movement information. They discovered a trend of browsing results from top to bottom, leading to the *Cascade Hypothesis* which has been adopted in many follow-up click model construction research. However, recent research finds more complex interaction patterns in users’ search process. For example, Navalpakkam et al. (2013) presented a lab study on the effect of a rich information panel on the right side of the search result column. They found that the widely adopted top-down linear examination order of search results is not applicable for a nonlinear page layout (e.g. one with knowledge graph components or multimedia vertical results). Buscher et al. (2009) also investigated users’ gaze information and built a model that can predict the most salient regions on Web pages.

To avoid the high expense and inconvenience of eye tracking devices,<sup>3</sup> researchers try to find a substitute that is both inexpensive to collect and suitable for large-scale experiments. Chen et al. (2001) recorded subjects’ eye/mouse movements in Web browsing

<sup>3</sup> Although there has been eye-tracking devices that are affordable to most researchers such as EyeTribe (<http://theeyetribe.com/>), it is still not possible for each search user to install one on their PCs or mobile devices. The calibration process before each usage also makes it difficult to persuade users to adopt them during their Web browsing processes.

experiments and found a strong correlation between gaze and mouse positions. To study implicit feedback on users' interest, Claypool et al. (2001) developed a browser to collect clicks, mouse movement, scroll and elapsed time. They found that scroll behavior is highly correlated with users' explicit interest rating. Rodden et al. (2007) went further into the relationship between eye movements and mouse movements, and discovered three patterns of active mouse usage: following the eye vertically with the mouse, following the eye horizontally with the mouse, and using the mouse to mark a promising result. Huang et al. (2012a) examined the question if "gaze is well approximated by cursor", by conducting a search study with 36 subjects and 32 search tasks, to determine when gaze and cursor are aligned. They found that cursor position is closely related to eye gaze on SERPs and can be used to estimate search result relevance and distinguishing bad abandonment, which suggests that users are dissatisfied with the search results, from good abandonment (users can find the answer directly without clicking the result) (Huang et al. 2011).

The high correlation between eye movements and mouse movements motivated many attempts to use mouse movement data for result ranking. Guo and Agichtein (2008) used mouse trajectories on SERPs to infer and disambiguate query intent. They also identified fine-grained features of examination and interaction behavior on both SERPs and landing pages to predict document relevance (Guo and Agichtein 2012; Guo et al. 2012). Speicher et al. (2013) built a cursor tracking system that can collect mouse movement information and estimate result relevance in a large scale. Lagun et al. (2014) try to automatically discover frequent mouse subsequences, or motifs, on landing pages to predict relevance of search results.

We can see that mouse movement information has been used in many existing studies on result examination process of search users. It also proves a useful signal reflecting the user's implicit relevance judgments. However, to our best knowledge, there is no work which tries to adopt mouse movement information into the construction of click models.

## 2.2 User behavior on heterogeneous SERPs

With an increasing number of vertical sources are aggregated into traditional web search results, the user behavior has become more and more various and complex in these heterogeneous search environment. Some prior work have noticed the difference of user behavior in federated search and strive to gain a better understanding of users' search behavior by either conducting user studies or performing large-scale log analyses. For example, Sushmita et al. (2010) showed that different presentation styles may affect users' click behavior. They indicated that users prefer to click more on video results than on image or news verticals. Wang et al. (2013) and Diaz et al. (2013) respectively demonstrated that different appearances may lead to various interaction biases. By performing a crowd-sourcing study with explicit user assessments, Zhou et al. (2013) found a similar behavior trending in federated search. Arguello and Capra (2014) revealed the relationship between users' intents and the senses of queries. By focusing on aggregated search coherence, they found that a so-called spill-over effect generalizes differently across verticals. To take it a step further, Liu et al. (2015b) systematically investigate the differences of user examination behavior in federated search and showed three vertical-aware behavior effects, namely attractiveness effect, examination cut-off effect and the examination spill-over effect.

These existing works have revealed the differences in examination and click behaviors between homogeneous and heterogeneous search environments. As the change of user behavior, especially click behavior on SERPs, may inevitably affect the performance of



click models, it inspires us to further find more general signals or features to improve click models in practical search scenario. Based on our experiment, we showed that mouse movements can provide a way to meet this requirement and help better model users' behaviors.

### 2.3 Click model construction

Click is a major user interaction with a search system and it can be used as implicit relevance feedbacks for ranking (Joachims et al. 2005). However, according to existing studies, user clicks are noisy because they are affected by different *biases*, like *position bias* towards documents higher in the ranking (Joachims et al. 2005; Guan and Cutrell 2007), *attention bias* towards visually salient documents (Wang et al. 2013), and novelty bias towards previously unseen documents (Zhang et al. 2011). In order to take these biases into account, researchers proposed a series statistical generative models to model users' browsing, examining, and clicking behaviors on SERPs. These models are called *click models*. By introducing random variables and dependencies between these random variables, click models can effectively and efficiently extract implicit relevance feedbacks from a large amount of search logs.

Different clicks models make different assumptions and hypotheses on users' browsing process on SERPs. One of the hypotheses shared by most click models is the *Examination Hypothesis*, which assumes that users will click a result if and only if they have examined the result on the SERP and think the result is relevant. Originally formulated by Richardson et al. (2007), the assumption can be described as follows:

$$P(C_i = 1) = P(E_i = 1)P(C_i = 1 | E_i = 1) \quad (1)$$

where  $P(C_i = 1)$  is the probability that the result on position  $i$  is clicked;  $P(E_i = 1)$  is the probability that result  $i$  is examined by users when browsing the SERP; and  $P(C_i = 1 | E_i = 1)$  is the attractiveness or relevance of the document on position  $i$  for the given query  $q$ .

Another commonly used hypothesis on user behavior is the *Cascade Hypothesis*. *Cascade Hypothesis* states that users examine the results on SERPs sequentially top-down. With this hypothesis, Craswell et al. (2008) presented a Cascade Model assuming that users examine results one by one and leave immediately after clicking a certain document. However, the user may look for multiple results and return to the SERP to continue searching even after finding a relevant document. Guo et al. (2009a) extended the Cascade Model by allowing users to return to SERPs after result clicks. They use a group of global parameters to describe users' returning probabilities at different positions. Chapelle and Zhang (2009) further investigated the process of this "returning" behavior. They proposed the Dynamic Bayesian Network Model (DBN) which is the first model to consider presentation bias. 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. If users think a certain result is attractive, they will click it, and if the result can satisfy their needs, they will leave SERPs without returning. Some widely used effectiveness metrics for IR systems, like *Rank-biased Precision* (RBP, Moffat and Zobel 2008) and *Expected Reciprocal Rank* (ERR, Chapelle et al. 2009), are also based on the cascade hypothesis. Zhang et al. (2010) further construct a click model and use it to estimate a suitable parameter for RBP.

Since the assumption in *Cascade Hypothesis* that users examine results sequentially doesn't always fit practical situations, several researchers designed click models based on more complex behavior assumptions. Dupret and Piwowarski (2008) described a User Browsing Model (UBM) by hypothesizing that users' examination behaviors are related to the position of current result and the distance to their last click. Liu et al. (2009) refined the parameter estimation process of UBM and presented a computing friendly model named Bayesian Browsing Model (BBM). Wang et al. (2013) found that such factors as results' appearance also affect users' examination. They therefore constructed a Vertical-aware Click Model that considers biases in addition to the position bias. Some existing studies (Thomas et al. 2013; Wang et al. 2015) used eye-tracking devices to analyze the examination behavior on SERPs and found that users' examination and click behavior is not strictly from top to bottom. Based on these findings, a Partially Sequential Click Model (PSCM, Wang et al. 2015) were proposed to model the non-sequential click behaviors in search logs.

Although many studies showed that mouse movement information is a reasonable substitute for eye tracking, only a few studies tried to incorporate this kind of information into click models. Huang et al. (2012b) tried to analyze users' interactions with search results and extended DBN by adding constraints on its latent variables. However, they left the abundant mouse movement data unexplored and only used mouse hover and scroll information. As we showed in Fig. 1, there exists much richer mouse movement information that is strongly correlated with user's examination behavior, and this information can be used in click models. Speicher et al. (2013) extracted a variety of features from mouse movement data to predict document relevance, but what they built is a regression model that cannot incorporate with existing behavior assumptions.

Most of existing click models, such as UBM (Dupret and Piwowarski 2008), BBM (Liu et al. 2009), Vertical-aware Click Model (Wang et al. 2013), focus on one special search scenario (e.g., heterogeneous or homogeneous search). It means that these models have some limitations in practical search environments, which becomes more and more complex and heterogeneous. In this paper, we proposed a novel framework to construct click models which can be applied in both heterogeneous and homogeneous search scenarios. Therefore, our work extends existing click model research in three aspects: Firstly, we take vertical biases (Wang et al. 2013; Liu et al. 2015b) into consideration when we construct click models and use rich mouse movement information to improve the performance of click models in both heterogeneous and homogeneous search environments. Second, we propose a novel framework to enhance existing click models by combining the examination prediction and behavior assumptions. The last but not the least, our proposed click models are validated on a large-scale data set, which is collected by a popular commercial search engine. That means the proposed click models can be available in real search scenarios.

### 3 Data collection

#### 3.1 Collection procedure

User examination has proven to be a useful event and signal to explain clicks and has been successfully incorporated in most click models. In our study, we do not intend to replace examinations by mouse movements, but rather to use mouse movements to predict

examinations. In such a way, we can still benefit from the existing click models with their assumptions on examinations.

To this end, we collect aligned data on examinations and mouse movements with human subjects. The collection of such data is not intended to merely confirm the strong correlation between them [which has already been shown in previous studies such as Huang et al. (2012a)], but rather to develop a prediction model of examination by mouse movements. To accomplish this goal, we design and implement an experimental search engine system to collect users' behavioral data, including mouse movement, eye movement and click-through information. This system is similar with the one adopted by Liu et al. (2014).

Our experimental study is performed in the following steps. Firstly, to ensure that each participant is familiar with the experimental system and the experiment procedure, they were asked to finish two warm-up tasks. During this step, participants only perceived that their interactions with the search system, including eye activities and mouse movements, will be recorded and were unaware of the real purpose of our study. Then each participant was asked to do calibrations with eye tracking device before starting experiments. After that, they were instructed to finish a series of search tasks with the experimental search engine.

The search tasks adopted in this study were sampled from the query logs of a popular commercial search engine so that they contain the practical users' search intentions. Considering the initial SERP of each task contains one or none vertical result and filtering too much verticals may affect the quality of search results, we select medium frequency queries whose original SERPs have less than three vertical results from the raw search logs. For each sampled query, we created a task description to simulate the information need and to avoid query ambiguity. For example, for the query "TV Dating shows", in the corresponding task description, we asked the participant to "find what TV dating shows are on broadcast recently". In order to make sure that all participants see the same SERPs in each search task, after reading the task description, the participant was presented with the sampled query and its corresponding first result page from the search engine that provides the search logs. The experimental is designed to observe users' examination and click behavior on SERPs. Therefore, participants were allowed to click any result link on the SERP and visit the landing page, but not allowed to reformulate their own queries. There was no time limit for each search task, the participant was instructed to finish a search task when he or she thought the simulated task was completed, or the presented SERP and results are not sufficient for completing the task.

During the search process for each task, participants' eye movements were recorded by the eye-tracker and their mouse movements/click-through behaviors were also logged by the injected JavaScript code on SERPs. We implemented our own version of mouse movement recorder but researchers may also rely on other public solutions such as Mouse Recorder.<sup>4</sup>

In the two steps of the experiment, we recruited 72 participants (37 for the first step and 35 for the second step, with a variety of majors and self-reported Web search expertises). They were all undergraduate students from a university.

Considering head-free eye trackers may make the collected interaction more natural and realistic, we use a Tobii X2-30 eye tracker which has a tracking frequency of 30 Hz and an accuracy of 0.5 of visual angle to capture participants' eye movements and deployed the search system on a 17-in. LCD monitor whose resolution is 1360 \* 768. Internet Explorer

---

<sup>4</sup> <http://www.mouse-recorder.com>.



11 browser was used to display the pages of search system, including the description pages and search result pages. To identify users' examination behaviors, we detect fixations using built-in algorithms from Tobii Studio. In these algorithms, a fixation is generated if recorded gaze locations are close to each other at least 60 ms.

With the above experimental settings, we collected a dataset of 1584 valid sessions after filtering system errors such as browser collapses and users' operation mistakes. A session is the process in which a user tries to complete a search task. Because query reformulation was not allowed in our experiment, each session contained only one query. Therefore, each session could be uniquely denoted as a tuple  $\langle user, query \rangle$ . Because 10 search results were presented to the participant on each SERP (and in each session), 15,840 valid examination instances were collected. Each instance could be uniquely denoted as a tuple  $\langle user, query, result \rangle$ . For each instance, we recorded: (1) the binary event that whether the user examined the result or not when completing the task defined by the query; (2) the binary event that whether the user clicked the result; and (3) the mouse movement behavior related to the result during the session. We use the fixation threshold of 500 ms to judge whether users have examined a result or not. This threshold is consistent with those used in previous studies (Buscher et al. 2009; Wang et al. 2013). We have also repeated our experiments using a fixation threshold of 200 and 1000 ms, which did not lead to any major differences regarding the conclusions of the paper. With the experimental system, we obtained 4989 positive (examined) instances and 10,851 negative (non-examined) instances for the examination prediction process.

### 3.2 SERP generation for search tasks

Considering the impacts of vertical results on user behaviors (Chen et al. 2012; Wang et al. 2013; Arguello and Capra 2014; Liu et al. 2015b), we take two steps to investigate users' mouse/eye behaviors in both homogeneous and heterogeneous environments. In the first step, we removed all advertisements and vertical results from original SERPs so as to focus on the user behavior on homogeneous search pages, which only contain organic results. It is worth noting that the order and content of these results remained consistent for each query.

In the second step, we further investigate the difference of users' mouse movement, click-through and eye activities in federated search. As an increasing number of vertical results are embedded in search results, the aim of this step is to simulate the real-world use of search engine and develop a robust model for practical scenario. Since previous work (Wang et al. 2013; Liu et al. 2015b) reveals that different factors lead to different behavior biases in heterogeneous search environment, we also manipulate three factors namely presentation style, position and quality to simulate possible cases in practical search scenario.

- *Presentation style* According to Liu et al. (2015b), verticals with different presentation styles may have distinct influences on users' behaviors. In this experiment, we focus on five popular verticals that are with different presentation styles, namely textual, encyclopedia, image, application-download and news. See Fig. 2 for examples for these five presentation styles. They are also the same verticals investigated by Liu et al. (2015b).
- *Position* As shown in existing work (Wang et al. 2013; Liu et al. 2015b), vertical position also plays an important role in affecting users' search behavior. Therefore, we

place these verticals at different positions (1, 3 and 5) and collect corresponding user behaviors.

- **Quality** We also manipulate the quality of vertical results to see its impacts on user behaviors and prediction models. Following the previous work (Buscher et al. 2010; Arguello and Capra 2014; Liu et al. 2015b), we use subsets of the terms from the initial task queries to retrieve off-topic verticals. Because these vertical results also contained the same terms from original queries, the appearance of off-topic results appear to be quite similar to the on-topic ones. The aim of this strategy is to make the occurrence of off-topic verticals more reasonable and natural. Note that the order and quality of other organic results is consistent with original SERPs.

For each search task, we generate six different kinds of federated SERPs (2 quality options  $\times$  3 position options) to collect user behavior. Each federated SERP in our experiment consist of one specific vertical and nine organic results. The vertical result was selected from a pool of verticals of different qualities (relevant or irrelevant) and integrated into the SERP at Position 1, 3, or 5 (corresponding to the top, middle or bottom of the first viewport). Therefore, we generate 180 (30 search tasks  $\times$  6 kinds of SERPs) federated SERPs in total. Each participant was asked to complete all 30 search tasks. To make sure that all tasks would be completed with equal opportunities in each SERP condition, we used a Graeco-Latin square design (Buscher et al. 2010; Arguello and Capra 2014; Liu et al. 2015b) to show tasks and conditions to users.

## 4 Examination prediction with mouse movement information

### 4.1 Extraction of mouse movement features

Rodden et al. (2007) identified four different ways in which Web search users act on the mouse: neglecting the mouse while reading a document, using the mouse to help read horizontally or vertically, and using the mouse to mark interesting results. According to this finding, existing studies (Guo and Agichtein 2012; Speicher et al. 2013; Lagun et al. 2014) have proposed a number of mouse movement features to predict examination behavior or relevance feedback information of users. Different from these studies, we focus on incorporating mouse movement information into the construction of click models, which means that the features extracted from mouse movement logs have to be scalable and applicable in practical Web search applications. Another major difference between our work and previous efforts on examination prediction lies in the fact that most of the previous efforts try to predict the whole eye fixation sequence of users, which is not necessary in the construction of click models. The different prediction goals also lead to different feature selection strategies. In our experiment, we extract a variety of mouse movement behavior features, such as mouse scrolling, the distance of mouse movement, time-spending on SERPs and so on. However, not all features are highly beneficial to the examination prediction. After comparing the prediction results of different feature groups, we finally adopt the six features, which have the best performance, as shown in Table 1 in the prediction process.

From Table 1 we can see that the proposed features are all extracted on a per result level. We avoid feature extraction procedures that may incur too much computational cost. Therefore, we do not use features such as mouse movement patterns [(e.g. motif extraction by Lagun et al. (2014), Liu et al. (2015a)]. Instead, we use *MostLeft* and

### Django documentation | Django documentation | Django

<https://docs.djangoproject.com> ▼

How the documentation is organized¶ Django has a lot of documentation. A high-level overview of how it's organized will help you know where to look for certain things:

#### Models

Django does make one adjustment to the Meta class of an abstract base ...

#### Templates

Templates¶ Being a web framework, Django needs a convenient way to ...

#### Django's Official Tutorial

Writing your first Django app, part 1¶ Let's learn by example. Throughout ...

#### Django's Forms

Django's role in forms¶ Handling forms is a complex business. Consider ...

#### Getting Started


Getting started¶ New to Django? Or to Web development in general? Well, ...

#### Django.Contrib.Admin

The Django admin site¶ One of the most powerful parts of Django is the ...

See results only from [docs.djangoproject.com](https://docs.djangoproject.com)

### Badminton - Wikipedia, the free encyclopedia 翻译此页







**Badminton** is a racquet sport played using racquets to hit a shuttlecock across a net. Although it may be played with larger teams, the most common forms of the game ...

<https://en.wikipedia.org/wiki/Badminton> ▼


### Images of flower

<bing.com/images>

See more images of flower

### In the news



#### Apple iPhone 7 Has Two Nasty Surprises

Forbes - 6 hours ago

Apple's iPhone 7 has problems which may make customers regret upgrading...

Some iPhone 7 owners report hissing sounds

Engadget - 7 hours ago

iPhone Next: How iPhone 7 hints at next year's breakthrough

CNET - 18 hours ago

More news for iphone 7

### Skype | Free calls to friends and family

<https://skype.com> ▼

Download Skype and stay in touch with family and friends for free. Get international calling, free online calls and Skype for Business on desktop and mobile.

#### Download Skype

Download Skype for your desktop. Available for ...

#### Skype Sign In

Sign into your Skype account to call and chat with friends ...

#### Windows

Download and launch Skype on any Windows device for ...

#### Skype for Web

Sign into your Skype account to call and chat with friends ...

#### About Skype

Skype and Microsoft have big dreams. Now Skype is part ...

#### Video Chat

Get free video calling on your mobile or desktop. Skype ...

*Textual*


*Encyclopedia*

*Image*

*News*

*Application-download*

Fig. 2 Examples of different kinds of vertical results

 Springer

*HorizontalMoveRight* to capture possible reading behaviors of users. This is because users sometimes use mouse to help read horizontally according to Rodden et al. (2007) and Liu et al. (2015a). The relationships between *MostLeft*, *HorizontalMoveRight* and the corresponding percentage of examined results are shown in Fig. 3.

In this figure the x-axis represents the normalized value of *MostLeft* and *HorizontalMoveRight* and is split into 20 buckets with equal range. The y-axis represents the percentage of examined results in each bucket. As we can see, *MostLeft* is negatively correlated with examinations, except for the left-most positions. On the other hand, *HorizontalMoveRight* is positively correlated with examinations. These observations indicate that when users examine a result, the cursor can stay at or traverse a left-most position (small *MostLeft* value), or move horizontally left to right. The longer the distance traveled horizontally by the cursor, the more likely it indicates an examination.

For the other features adopted in the prediction process, *VerticalDwellTime* is selected together with *DwellTime* because sometimes users just scroll the SERPs without moving the cursor horizontally. Therefore, while user examine certain results, the cursors may not appear in the result zones but they are within the vertical ranges of the corresponding results. *HoverTime* is also selected because some users do not move the cursors while reading contents. The number of actions (*ActionNumber*) shows whether users are interested to interact with the corresponding results and for actions. The actions we count are hovers, clicks, text selection and cursor movements.

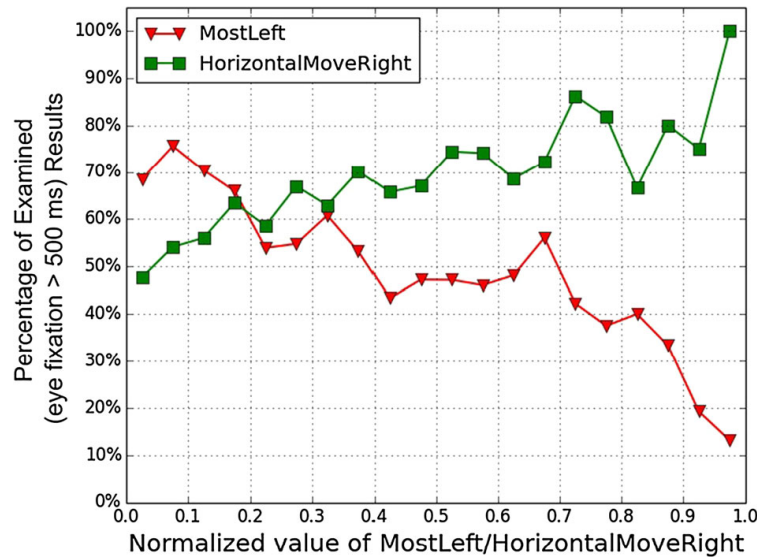
Notice that the mouse movement features are easy to collect in practice. All these features, together with click-through information, have been recorded in the logs of a popular commercial search engine, which we will use in our experiments.

## 4.2 Prediction of examination behaviors

To predict the examination of results with the proposed mouse movement features, we tried five different learning methods that are widely-adopted in related studies. We choose SVM, Logistic Regression, Random Forrest (RF), Decision Tree (J48), and gradient boosting regression tree (GBRT) (Friedman 2001). SVM and Decision Tree are classification models that directly output a binary label (examined or unexamined); while GBRT, Logistic Regression and RF are regression models that predict the probability of examination for each result. We use 0.5 as a threshold to convert regression models' predictions into binary judgment for comparison purposes.

**Table 1** Description of features extracted from mouse movement data

Feature name	Description
<i>MostLeft</i>	The most left position that cursor reaches in the result's display area
<i>HorizontalMoveRight</i>	The total rightward distance of cursor in the result's display area
<i>DwellTime</i>	The total dwell duration that cursor stays in the result's display area
<i>VerticalDwellTime</i>	The total dwell duration that cursor stays within the result's display area horizontally
<i>HoverTime</i>	The total duration that cursor hovers over the result's display area
<i>ActionNumber</i>	The number of cursor actions that happen in the result's display area



**Fig. 3** The relationship between result examination and *MostLeft/HorizontalMoveRight*

To evaluate the performance of different learning methods, we adopt the evaluation metrics of precision, recall, F1-score, accuracy and Matthews correlation coefficient (MCC) score (Matthews 1975). MCC is used in machine learning as a measure of the quality of binary classifications. It is generally regarded as a balanced measure which can be used even if the classes are skewed. MCC is computed as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where  $TP$ ,  $TN$ ,  $FP$ , and  $FN$  are the numbers of true positives, true negatives, false positives, and false negatives.

Considering the difference of user behavior in federated search (Buscher et al. 2010; Arguello and Capra 2014; Liu et al. 2015b), we tested these learning methods with fivefold cross validation both in homogeneous and heterogeneous environment so that we can compare the performance in different search scenario and further select the most robust method to build examination prediction model in our experiment. The comparison results of these five methods are shown in Table 2 (homogeneous environment) and Table 3 (heterogeneous environment), respectively.

In Tables 2 and 3, the best performing result in each column is in bold. Comparing the results of these two tables, we can see that GBRT performs best in terms of MCC and accuracy both in homogeneous and heterogeneous environment, and the performance of its precision, recall and F1 score is also more stable and available than other methods in these two different scenario. For example, in Table 2, we can observe that SVM has a slightly

**Table 2** Examination prediction results of different learning methods in homogeneous environment ( $n = 8900$ , fivefold cross validation)

Model	Prec.	Recall	F1-score	MCC	Accu.
GBRT	0.731	<b>0.733</b>	<b>0.731</b>	<b>0.440</b>	<b>0.733</b>
SVM	<b>0.734</b>	0.711	0.714	0.435	0.711
Logistic	0.730	0.725	0.726	0.437	0.724
RF	0.695	0.698	0.695	0.364	0.698
J48	0.721	0.723	0.719	0.416	0.723



**Table 3** Examination prediction results of different learning methods in heterogeneous environment ( $n = 6940$ , fivefold cross validation)

Model	Prec.	Recall	F1-score	MCC	Accu.
GBRT	0.754	0.650	0.698	<b>0.601</b>	<b>0.849</b>
SVM	0.661	<b>0.773</b>	<b>0.713</b>	0.600	0.833
Logistic	<b>0.759</b>	0.511	0.610	0.520	0.826
RF	0.736	0.654	0.694	0.597	0.846
J48	0.725	0.569	0.637	0.532	0.827

better precision than GBRT in homogeneous search, while it perform worst in terms of precision in heterogeneous environment (shown in Table 3). That means the prediction performance of SVM method might be susceptible to the search environment. Therefore, we chose GBRT in our prediction model for the subsequent steps.

Through the results shown in Tables 2 and 3, we also noticed that the prediction results are not perfect. Although we believe that the mouse movement features contain valuable information about users' examination process, we have to point out that mouse movement data contain considerable noise. In our experiment process, subjects sometimes neglected the mouse while browsing results and left it at the right blank side of SERPs. The correlation between eye/mouse movements also varies depending on users. Those who are more familiar with Web searching, such as subjects in computer science major, move mouse faster than others and usually finish a query with fewer interactions. These factors may affect the construction of our examination prediction model. However, even with this noise in mouse behavior data, we show in the following sections that the click models can benefit from the predicted examinations.

## 5 Click model with mouse movement information

As mentioned in Sect. 2, modern click models focus on removing behavior biases and providing an adjusted estimation of result relevance. For this purpose, these models rely on different hypotheses on users' search interaction process. In this section, we propose modifications on several popular click models by incorporating mouse movements into their construction. To simplify the construction framework, we adopt linear fitting method in our experiment. This method not only helps us analyze the impact of mouse movement information based on the weights of learned parameters, but also provides an extensible framework that is easy to incorporate different hypotheses into the construction of click models. The modified models will be called Click Models with Mouse movement information, or CMwM for short and the corresponding modified model will have the suffix "wM".

### 5.1 DCMwM

DCM (Guo et al. 2009a) extends the *Cascade Model* by adding position dependent variables. These variables represent the probability that users return to examine the result at position  $(i + 1)$  after clicking the result at position  $i$ . If we use  $\lambda_i$  to represent these probabilities, the assumption can be formulated as:

$$\begin{aligned}
 P(e_1 = 1) &= 1 \\
 P(c_i = 1) &= P(e_i = 1)r_{d_i} \\
 P(e_{i+1} = 1) &= \lambda_i P(c_i = 1) + (P(e_i = 1) - P(c_i = 1))
 \end{aligned} \tag{2}$$

where  $P(e_i = 1)$  and  $P(c_i = 1)$  represent respectively the possibility that a user examines and clicks the document at position  $i$ ; and  $r_{d_i}$  is the relevance of document at position  $i$ . The corresponding estimations of  $r_{d_i}$  and  $\lambda_i$  are

$$r_{d_i} = \frac{\# \text{Clicks on the result at position } i}{\# \text{Impressions with last - clicked position } \geq i} \tag{3}$$

$$\lambda_i = 1 - \frac{\# \text{Impressions with last - clicked position} = i}{\# \text{Impressions with the result at position } i \text{ clicked}} \tag{4}$$

Since we can now predict the examination probability of results with mouse movement data, the estimation of examination parameters can be refined. Let us use  $P(m_i = 1)$  to represent the predicted examination probability of the result at position  $i$ . We can modify the parameters in DCMwM using a weighted combination as follows:

$$\lambda'_i = (1 - w)\lambda_i + wP(m_{i+1} = 1) \tag{5}$$

where  $w$  is a weight parameter set empirically (we will show in Sect. 5.3 that this parameter can be learned automatically in position based models such as PUBMwM), and  $\lambda'_i$  is a session-dependent parameter that we use to replace the session-independent parameter  $\lambda_i$  in Eqs. (2) and (4). Note that  $\lambda'_i$  is more informed than the original estimation of  $\lambda_i$  because it also relies on the observation on the examination of the  $(i+1)$ th result rather than inferred from the click behavior only. We use a simple weighted combination here because it can be incorporated to almost any click models that based on the examination hypothesis (see Sect. 2.3), and tuning the weight  $w$  can control the influence brought by the examination prediction models.

The mouse movement information of the result located at position  $(i + 1)$  is used in the estimation of  $\lambda'_i$  because the behavior assumption in DCM supposes that  $\lambda_i$  (and corresponding  $\lambda'_i$ ) represents the probability of returning to the next result at  $(i+1)$ . With a similar procedure of estimating  $\lambda_i$  from (2) as DCM does, we can estimate the hidden variable  $\lambda_i$  as follows:

$$\lambda_i = \frac{|S_i^c| - |S_{i=l}^c| - w \sum_{S_i^c} P(m_{i+1} = 1)}{(1 - w) |S_i^c|} \tag{6}$$

where  $S_i^c$  = Impressions with the result at position  $i$  clicked and  $S_{i=l}^c$  = Impressions with last-clicked position =  $i$ .

With this modification, the relevance prediction formula (3) in DCMwM is the same as the original DCM.

## 5.2 DBNwM

DBN (Chapelle and Zhang 2009) shares a similar hypothesis with DCM in that users can return to SERP and examine more documents after clicking a result. However, instead of using a group of parameters, it describes this process using one global parameter:  $\gamma$ . The relevance prediction of DBN is different from DCM in that it divides the estimation of

document relevance into two steps: estimation of attractiveness and estimation of satisfaction. DBN uses  $\alpha_d$  and  $s_d$  to represent a document's attractiveness and satisfaction, separately. We can formulate its assumptions with the following equations:

$$\begin{aligned} P(e_1 = 1) &= 1 \\ P(c_i = 1) &= P(e_i = 1)\alpha_{d_i} \\ P(e_{i+1} = 1 \mid c_i = 1) &= (1 - s_{d_i})\gamma \\ P(e_{i+1} = 1 \mid e_i = 1, c_i = 0) &= \gamma \\ P(e_{i+1} = 1 \mid e_i = 0) &= 0 \end{aligned} \quad (7)$$

Similar to DCMwM, we can use a weighted combination to incorporate the predicted examination information into DBN, which is:

$$\gamma'_i = (1 - w)\gamma + wP(m_{i+1} = 1) \quad (8)$$

The parameter estimation of DBN uses the EM algorithm, and its original descriptions are stated by Chapelle and Zhang (2009). In order to reduce the influence of good abandonment (Huang et al. 2011) when users visit the search result page, we added a statistical bias, which was inferred from mouse movement information, into the calculation of parameter  $\gamma$  in (8); but we can still use the EM process of DBN for parameter inferences.

Huang et al. (2012b) have extended DBN with mouse information in a different way. They collected users' scroll and hover behaviors on SERPs and added constraints to DBN so that every result that had been scrolled over or hovered over would be treated as examined. To compare the performance of this DBN extension, we also implemented Huang's method and use it as one of our baselines (named HDBN) in Sect. 5.

### 5.3 UBMwM and PUBMwM

UBM (Dupret and Piwowarski 2008) does not follow the framework of Markov process as DCM and DBN do. In UBM, the probability of examining a result is estimated according to the result's distance to the last click position as well as its position on SERP. The behavior assumptions can be formulated as follows:

$$\begin{aligned} P(e = 1 \mid r, d) &= \gamma_{rd} \\ P(c = 1 \mid u, q, r, d) &= \alpha_{uq}\gamma_{rd} \\ P(c = 0 \mid u, q, r, d) &= 1 - \alpha_{uq}\gamma_{rd} \end{aligned} \quad (9)$$

$P(e = 1)$  is the probability of examination,  $P(c = 1)$  is the probability of click;  $r$  is the rank of the result, and  $d$  is the distance between rank  $r$  and the last click position.  $\alpha_{uq}$  represents the relevance between result  $u$  and query  $q$ , and  $\gamma_{rd}$  are a group of session-independent parameters that are used to infer examination probabilities.  $\alpha_{uq}$  and  $\gamma_{rd}$  can be estimated through the EM algorithm.

Similar to DCMwM and DBNwM, we can still extend this model using a weighted combination:

$$P(e = 1 \mid r, d) = (1 - w)\gamma_{rd} + wP(m_r = 1) \quad (10)$$

We can also use EM algorithm to estimate the parameters for UBMwM model. Note that different from DCMwM and DBNwM, the mouse movement information of the result at position  $r$  is adopted here instead of that at position  $(r + 1)$ . This is because the estimation of examination probability in (10) is for the result at position  $r$ .

We can see that the weight parameter  $w$  in (10) is a global one and independent of result positions. However, due to the fact that users pay much more attention to the higher-ranked results on SERPs, it is reasonable to assume that users may interact more frequently with the top results while the lower-ranked ones may receive relatively few interactions. For results with little mouse movement information, it may be difficult for the examination prediction method to judge whether this result is examined or not. Therefore, the prediction based on user behavior assumptions (from the original click model) may be more reliable in these lower positions. If we take this position effect of mouse movement information into consideration, we can replace  $w_r$  with a group of position-dependent parameters as follows:

$$P(e = 1 \mid r, d) = (1 - w_r)\gamma_{r,d} + w_rP(m_r = 1) \quad (11)$$

In Eq. (11), the weight parameters  $w_r$  varies according to different result positions. Since it is not possible for us to preset all  $w_r$  values in advance, we adopt a gradient descent algorithm to learn the parameters so that the difference between a model's predicted click behavior and actual click-through behavior on the training set can be minimized.

As  $P(m_r = 1)$  denotes the prediction of the examination probability with mouse movements, this parameter depends on not only the data set which we collected but also the learning method adopted in the prediction process. In other words, the accuracy of the examination prediction is likely to be susceptible to the potential influence of learning methods, especially in the heterogeneous search environment. To make this model more robust and flexible in practical scenario, we further estimate the examination probabilities as follows:

$$P(e = 1 \mid r, d) = \frac{1}{1 + e^{-(c_0 + w_0\gamma_{r,d} + \sum_{i=1}^k w_i f_i)}} \quad (12)$$

where  $f_i$  is a feature set, which consists of mouse movement features mentioned in Sect. 4.1.  $w_i$  is a weight parameter corresponding to the  $i$ -th feature. We use logistic regression as the basic model framework so that the prediction of examination probability could be normalized to a range of 0 to 1. We can see that this model eliminates the potential influence of training methods and data set and can be adopted to prediction tasks in both homogeneous and heterogeneous environments. Therefore, there is no need to predefine the parameters in this new model. We name this click model as Position-aware UBMwM (PUBMwM) and we will compare its performance with the UBMwM method in experiment part.

## 6 Experimental results

### 6.1 Experiment setups

To evaluate the effectiveness of the proposed models in practical search environment, we collected search logs and mouse movement features (as shown in Table 1) from a popular

commercial search engine.<sup>5</sup> The data set was collected from February 13 to February 28 in 2014. Because existing studies (Wang et al. 2013) show that users' behaviors on vertical results are different from those on ordinary text results, we will discuss the performances of our proposed click models in homogeneous and heterogeneous environment in two separated sections. As is the case for any click model, infrequent queries can be hardly accounted for. Therefore, in this study, we focus on queries having a certain frequency and filtered out those queries that involved in less than 15 sessions in 15 days. For homogeneous environment, we extract a dataset of 79,813 distinct queries and 8,452,125 query sessions (in which no vertical results are on SERPs). We also separated sessions of each query with the ratio of 7:3 as in Guo et al. (2009a) and Dupret and Piwowarski (2008), and obtained a train dataset with 5,916,488 query sessions and a test dataset with 2,535,636 query sessions. For heterogeneous environment, we extract a dataset of 4,348,462 query sessions. It is worth noting that each SERP in these sessions only contains one vertical result. We separate them into a training data set with 3,049,934 query sessions and a testing data set with 1,298,528 query sessions.

In the experiments, we tested the performances of DCMwM, DBNwM, UBMwM and PUBMwM on the collected data set and tried different values of weight parameter  $w$  on the first three models to investigate the influence of mouse movement information. We also labeled the relevance scores of a number of query-result pairs to analyze the performance of the models in relevance estimation.

For the selection of evaluation metrics, we use click perplexity to reflect the models' performance on predicting clicks. The click perplexity for position  $i$  is computed as,

$$p_i = 2^{-\frac{1}{N} \sum_{n=1}^N (C_i^n \log_2 q_i^n + (1 - C_i^n) \log_2 (1 - q_i^n))} \quad (13)$$

where  $N$  is the number of sessions,  $C_i^n$  denotes the binary click event, and  $q_i^n$  is the click probability given by the click model. With the parameters learned from training sessions, click perplexity measures the probability of predicting each click in test sessions. Perplexity has a minimum value of 1, and lower values indicate better performance of the corresponding model.

The click models will also be used to rank search results according to the predicted relevance. To compare models' performance in relevance estimation, we constructed an online labeling system and recruited 30 undergraduate students to label 1,239 query-result pairs (see Sect. 6.4 for more details). For each query-result pair, we showed the query and the screenshot of whole result webpage to students and ask them to label relevance between the query and the result page as "Good", "Fair", "Poor", or "Bad". To guarantee the quality of the labeling process, we use seven queries labeled by the authors to train students in advance. We have each document labeled by three different students, and the Cohen's Kappa coefficient among them is 0.435 (moderate agreement) if we group users annotation into binary judgment ("Good" or "Fair" means relevant while "Poor" or "Bad" means irrelevant). Therefore, we use the majority vote of the binary judgments made by three students in the calculation of relevance evaluation metrics. We choose to use Mean nDCG@N, MAP, Mean ERR [*Expected Reciprocal Rank* (Chapelle et al. 2009)] as in previous studies to evaluate the performance.

All the baseline models are based on the open source implementations provided by Chuklin et al. (2013). Our own implementation of the click models and a sample of the

<sup>5</sup> The search logs were collected from <http://www.sogou.com>, a popular Chinese commercial search engine. Because the dataset is commercial-in-confidence, we are not allowed to release the whole dataset.



mouse movement logs will also be made available to public after the double blind review process.

## 6.2 Click prediction in homogeneous environment

### 6.2.1 Effect of weight parameter $w$

We first compare the click models with mouse movement information incorporated and their original versions. Note that when the weight parameter  $w$  in DCMwM, DBNwM and UBMwM is set to 0, the models retain its original structure; when  $w$  is 1, the model's examination parameters are completely replaced by the predicted values from mouse movement features. Therefore, by varying the value of  $w$ , we will see the impact of mouse movement information.. The comparison results are presented in Table 4.

As shown in Table 4, with a proper setting of  $w$ , the proposed models with mouse movement information outperform their original versions (with smaller perplexity values). Especially, when  $w$  is not too large ( $<0.6$ ), the performance improvement is steady for all three click models. It means that the incorporation of mouse movement information actually helps click models better predict user clicks. A combination of both mouse movement information and the original user behavior assumption produces the best performance (as shown in bold figures) for all three models, which indicates that they both contribute to the prediction of examination.

We also notice that UBMwM obtains the highest and most steady improvement among all the models we tested: it has a +5.20% reduction in perplexity when  $w$  is set to 0.7 and it still performs better than the original model even when  $w$  is set to 1.0 (which means that we only use the predicted examination based on mouse movements). Under the assumptions of UBM, the examination probability of a result is only related to its position and the distance from the last click. Given the position of the last click, each result's examination probability is independent of others. Because we collect mouse movement features on a per-result level, the examination predictions given by our model are also independent of each other. The design of examination prediction model is consistent with UBM's assumptions. That is one possible reason for the phenomenon that UBMwM performs better than DCMwM and DBNwM. It is worth noting that the improvement of UBMwM is not statistically significant. We find that for higher-ranked positions, the performance of UBMwM model is much better than original UBM model while UBMwM model doesn't perform so well at lower-ranked positions. One possible reason is that UBM model hypothesizes that users examination behaviors are related to the position of current result and the distance to their last click, which means abundant mouse movement information at higher-ranked results may affect the performance of UBM model when users examine results at lower-ranked positions. Besides, the lack of mouse movement information at lower-ranked positions also has an important impact on the performance of UBMwM model. That is another possible reason for the phenomenon that the improvement of UBMwM is not statistically significant.

Another interesting observation is that although the HDBN algorithm (Huang et al. 2012b) (which uses part of the mouse movement information—scroll and hover) outperforms the original DBN by 2.94%, it cannot beat the DBNwM algorithm which considers richer mouse interaction data. Meanwhile, the proposed position-aware UBMwM algorithm obtains the best performance among all models. We can see that assigning different  $w$  for different positions actually helps improve the click prediction performance. It

**Table 4** Perplexities of the proposed models with different  $w$  ( $w=0$  corresponds to the original click model without mouse movement information)

Model/ $w$	0.0	0.1	0.2	0.3	0.4	0.5
DCMwM	1.3220	1.3175	1.3150	1.3140	<b>1.3131</b>	1.3136
	–	+1.40%*	+2.17%*	+2.48%	<b>+2.76%</b>	+2.61%
DBNwM	1.2922	1.2846	1.2800	<b>1.2783</b>	1.2791	1.2820
	–	+2.60%**	+4.18%**	<b>+4.76%*</b>	+4.48%*	+3.49%
HDBN	1.2836 (+2.94% compared with original DBN)					
UBMwM	1.2809	1.2763	1.2738	1.2715	1.2695	1.2679
	–	+1.64%	+2.53%	+3.35%	+4.06%	+4.63%
PUBMwM	<b>1.2636 (+6.16% compared with original UBM)</b>					
Model/ $w$	0.6	0.7	0.8	0.9	1.0	
DCMwM	1.3152	1.3217	1.3652	1.4694	1.3185	
	+2.11%	+0.01%	–13.4%	–45.6%*	+1.08%	
DBNwM	1.2871	1.2945	1.3048	1.3192	1.3408	
	+1.75%	–0.79%	–4.31%	–9.24%	–16.6%*	
HDBN	1.2836 (+2.94% compared with original DBN)					
UBMwM	1.2668	<b>1.2663</b>	1.2666	1.2678	1.2703	
	+5.02%	<b>+5.20%</b>	+5.09%	+4.66%	+3.77%	
PUBMwM	<b>1.2636 (+6.16% compared with original UBM)</b>					

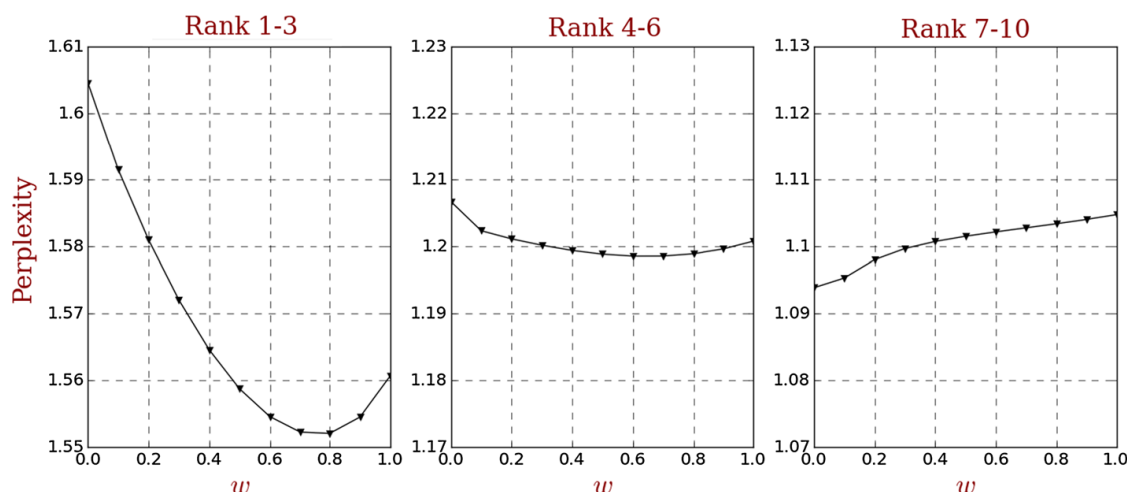
“+” means improvement in performance and decrease in perplexity, and the improvement of perplexity value  $p_1$  over  $p_2$  is computed as  $(p_2 - p_1)/(p_2 - 1) \times 100\%$  (Guo et al. 2009b). Two-tailed t-test is performed for significance while \* and \*\* represent  $p$  value  $<0.05$  and  $0.01$ , respectively

indicates that mouse movement information may play different roles for results in different ranking positions and we will investigate this problem with more detailed analysis.

### 6.2.2 Effect of result position

In our collected user behavior data, each SERP contains exactly 10 results as in most commercial search engines. To show the effect of result position in the performance of click models, we group the ranking positions into three categories: high rank (position 1 to 3), median rank (position 4 to 6) and low rank (position 7 to 10). In Table 4 we already showed that UBMwM performs the best and the most steadily. Therefore, we just show the impact of mouse movements with this model (similar impact is observed on other models).

As shown in Fig. 4, UBMwM’s performance varies significantly in different ranking positions. For high ranked positions, perplexity value decreases with the increase of  $w$  when  $w$  is not too large. Even when  $w$  is set to 1.0, the performance improvement is still noticeable. However, for median-ranked results, the performance increase is not so large but it generally follows the same trends as higher-ranked results. Things are rather different for lower-ranked results, in which the perplexity score show performance loss with the incorporation of mouse movement information. This observation can possibly be explained by the fact that the mouse movement information is not available for a large proportion of lower-ranked results. While interacting with commercial search engines, a large part of



**Fig. 4** Perplexities of UBMwM with different weight parameter  $w$  on different result positions

users may not even scroll the SERP to view results at the bottom positions. According to our statistics, the percentages of query sessions with mouse movement on bottom results are rather low (7.7, 6.2, 5.6% for the results ranked at 8th, 9th and 10th positions). Meanwhile, the corresponding percentages for results in top three results are 78.1, 62.3, 47.6%, respectively.

Indeed, the amount of mouse movement information for different positions plays a determining role in the accuracy and usefulness of the prediction based on mouse movements. In top positions, the proposed examination prediction model can predict a more accurate examination probability according to mouse movement features than the global latent parameters of the original click model. This leads to a better performance of our proposed click models. While in bottom positions, due to the fact that the vector of mouse movement features is often a zero vector, the examination prediction model will predict a much lower examination probability (close to 0) as compared to the global latent parameters of the original click models. The model will thus predict “no click” at these positions. If there is a click event occurring in the testing data set at these positions, the perplexity value will be greatly affected and result in degradation in performance.

Figure 4 shows clearly that user behavior assumptions and mouse movement information play complementary roles in the prediction of examination behavior of search users. Therefore, a combination of them can take advantage of both. The results in Fig. 4 also suggest strongly that different weights should be used for different positions. The weight parameter  $w$  should be high for top positions and decreases gradually. We did not implement a strategy with different weights set manually. However, the PUBMwM uses a similar strategy: With PUBMwM, the weight parameter  $w$  is trained with gradient descent to minimize a model’s perplexity on the train set. Therefore, it could choose suitable  $w$ s for different positions and gain better performance. The results in Table 4 validates our assumption because PUBMwM gains the best performance.

The weight parameters learned from training set by PUBMwM (see Table 5) can also provide us with some insight of the different roles of mouse movement information plays in different positions—the larger the weight, the more important the mouse movement information.

As shown in Table 5, PUBMwM’s  $w$  gradually decreases from position 2–10 as we expected, which shows that mouse movement information is more important at higher positions for which more mouse movement information is available.

### 6.2.3 Effect of query frequency

We tested the effectiveness of the proposed model for queries with different frequencies. As UBMwM performs best according to Table 4, we only show its performance in this experiment. We group the queries in our data set into three bins: 15–100, 100–1000 and over 1000. Note that queries with less than 15 occurrences are removed to avoid data sparsity problem as described in Sect. 5.1. The comparison of results is shown in Fig. 5.

From this figure we can see that UBMwMs' performance on high-frequency queries is higher than that of low frequencies, which is expected. This observation is consistent with the findings in existing studies (e.g. Guo et al. 2009a, b) that click models perform better on queries with high frequencies. Meanwhile, their perplexities' share the same trend with the increase of  $w$ . These results indicate that mouse movement data can improve models' performance irrespective of the queries' frequency.

### 6.2.4 Effect of query intent

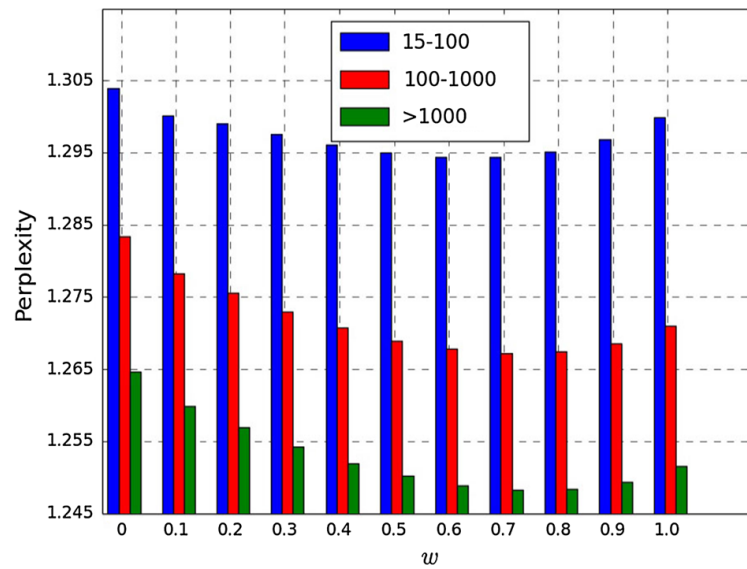
Guo and Agichtein (2008) observed that users' mouse movement behaviors vary depending on their search intent. Therefore, we also investigated the proposed models' performance on queries with different intents. We use the taxonomy introduced by Broder (2002), which classifies queries into three intents: navigational, transactional, and informational. We randomly sampled 1,110 queries in total (about 1.4% in the data set) and had their intents annotated by professional assessors. We got 270 navigational, 418 transactional, and 422 informational queries and then separately calculated the models' perplexities for each query set as shown in Fig. 6.

We can see from Fig. 6 that with proper parameter settings, click perplexity decreases from 1.188 to 1.186 (with +1.06% improvement in prediction performance) for navigational queries, from 1.32 to 1.30 (+6.25%) for transactional queries, and from 1.24 to 1.22 (+8.33%) for informational queries. It means that after incorporating mouse movement information, transactional and informational queries obtain more performance improvement than navigational queries.

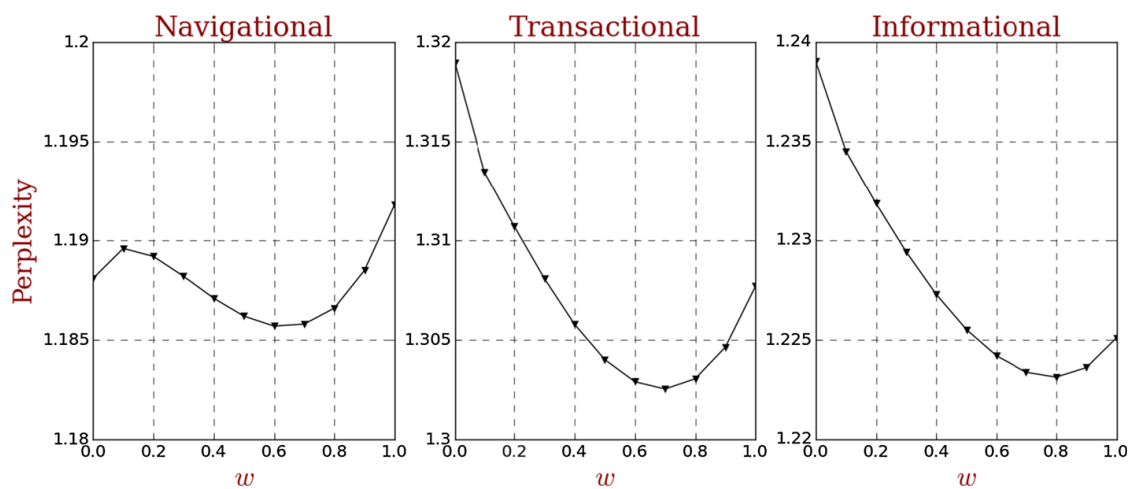
This phenomenon may be caused by the fact that users have different behavior patterns for different search tasks Liu et al. (2006). Transactional and informational search tasks are more complex than navigational ones, users need to read more information before making click decisions on transactional and informational queries. Therefore, the mouse interactions on SERPs are more abundant and the examination prediction based on mouse features are more accurate for transactional and informational queries. Meanwhile, navigational queries are relatively simple in general and receive fewer interactions in search sessions. Therefore, the information provided by mouse movements may be limited for navigational queries.

**Table 5** Parameter  $w$  automatically learned by PUBMwM on different result positions

Position	1	2	3	4	5
$w$	0.7495	0.8510	0.8101	0.7653	0.6668
Position	6	7	8	9	10
$w$	0.5648	0.3948	0.1812	0.0907	0.0621



**Fig. 5** Perplexities of UBMwM with different weight parameter  $w$  on queries with different frequencies



**Fig. 6** Perplexities of UBMwM with different  $w$  on different query intents ( $w = 0$  means original UBM)

### 6.3 Click prediction in heterogeneous environment

With the systematical analysis of the different performance of DCMwM, DBNwM, UBMwM and PUBMwM in Sect. 6.2, we are also curious about the performance of these models in heterogeneous search environment. Considering multiple multimedia verticals may have more complex influence on users' behavior, we focus the situation on the one-vertical SERPs in our experiment, which may have covered most practical scenario, and click model with multiple vertical results will be investigate in the future. For examination prediction models, we adopted the data set which we collected in the second-step experiment (see Sect. 3.1) and use click perplexity to evaluate click models' performance. Since PUBMwM obtains much better prediction performance than other models in homogeneous environment, this model was selected and tested in the heterogeneous environment.

Moreover, according to Wang et al. (2013) and Liu et al. (2015b), the factors of vertical results also affect user examination behavior. In this section, we further incorporated these



potential impacts into our model and investigated the performance of the modification. Table 6 shows the extra vertical factors which were taken into account in our experiment. It is worth noting that all of these vertical features can also be easily extracted from the raw search log of real-world search engines so that we are able to deploy the extended model into the practical scenario. Additionally, considering the influence of learning methods in the examination prediction, we also compare the performances of Eqs. (11) and (12) which are both applicable for the proposed position-aware UMBwM Model.

Table 7 presents perplexity scores of both UBM and Position-aware UBMwM. To analyze the influence of vertical factors, we investigate the performance of PUBMwM with and without vertical features, respectively. As shown in Table 7, we can observe the position-aware UMBwM model still performs better than UBM model in heterogeneous environment, whose perplexity is 1.27% less than that of UBM. That means the proposed position-aware UBMwM model is applicable and robust both in homogeneous and heterogeneous environment. We can also find the model which contains vertical features outperforms the original position-aware UBMwM and obtain the best performance among all models. Although the improvement is relatively limited in our experiment, it is significant considering the large size in test set. It also demonstrates the incorporation of vertical impacts may improve the performance of click models to some extent.

#### 6.4 Relevance prediction

To verify our models' performances in relevance estimation, we sampled a number of queries from the test dataset and labeled their results' relevance scores. Due to limited human resources, we focus on the queries in which original click models and revised ones generate much different rankings. Similar strategies have also been used in previous studies (e.g. Awadallah and Zitouni 2014; Li et al. 2015). By this means, we want to show whether the proposed model with mouse movement improves original one. However, we have to point out that this sampling strategy may lead to larger differences in model performances than random sampling because in many cases, the result rankings generated by different models are not so different.

To locate the query cases that show actual differences in result rankings, we choose the rankings generated by UBMwM because we only have limited annotation resources and UBMwM gain best performance among all three revised models according to Table 4. We calculated the Jaccard distances between the top 3 results of these two ranking lists for all queries in the test set. After that, we selected 252 queries with the largest distances and their corresponding query-result pairs' relevance scores were annotated. Notice that on these sampled queries, UBMwM may perform better than UBM, or the reverse. In total 1,239 pairs were annotated because there were generally 5 or 6 documents for each query.

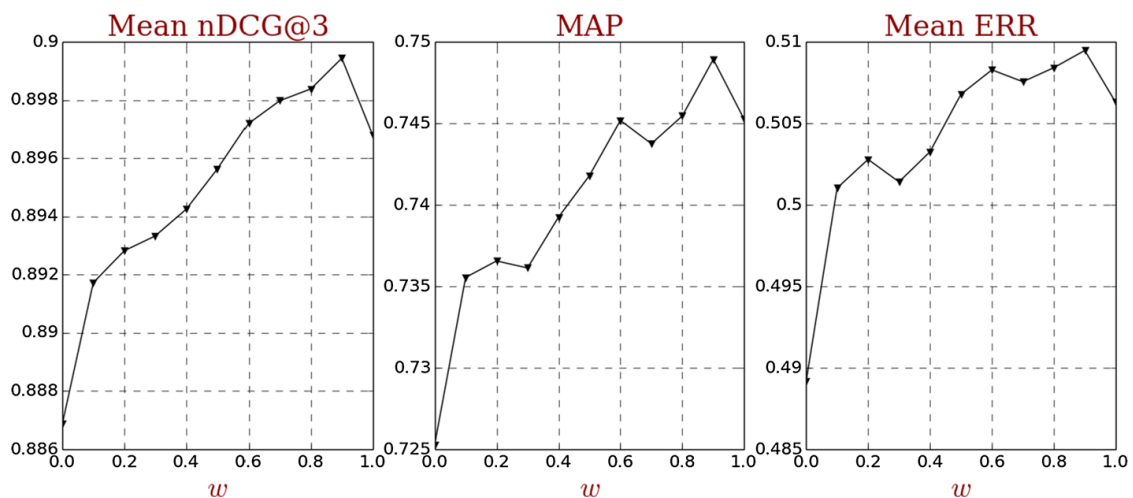
**Table 6** Vertical-aware features adopted in heterogeneous search environment

Feature name	Description
ResultType	A vector which represents the presentation style of vertical
PositionVertical	The relative position of the current result against the vertical on the SERP
Distance	The absolute distance from the current result to the vertical result
NumberOfImages	The number of images in this current results. (e.g. for an organic result, the value is 0)

**Table 7** Perplexity comparison of UBM, position-aware UBMwM click model with different features

Click model	Perplexity
UBM	1.2324
PUBMwM (with Mouse-only features)	1.2170 (+6.63%*)
PUBMwM (Mouse and Vertical features)	<b>1.2145 (+7.70%*)</b>

“+” presents improvement in performance and decrease in perplexity when comparing with original UBM, and the improvement of perplexity value  $p_1$  over  $p_2$  is computed as  $(p_2 - p_1)/(p_2 - 1) \times 100\%$  (Guo et al. 2009b). Two-tailed  $t$  test is performed for significance while \* and \*\* represent  $p$  value  $<0.05$  and  $0.01$ , respectively

**Fig. 7** Relevance prediction results of UBMwM with different  $w$  ( $w = 0$  means original UBM)

We choose Mean nDCG@3, MAP and Mean ERR to evaluate the performance of UBMwM and UBM (UBMwM turns to UBM when  $w = 0$ ) with the binary relevance judgments and the results are shown in Fig. 7.

The experimental results in Fig. 7 show that UBMwM outperforms the original UBM in relevance estimation with different kinds of metrics. The performance of UBMwM gradually improves along with the increase of  $w$  and reach its best when  $w$  is 0.9. The improvement over the original UBM is all significant with paired  $T$  test ( $p < 0.01$ ). This result shows that the modified click model has an improved estimation of document relevance once mouse movements are incorporated.

## 7 Conclusions and future work

Click models rely on user behavior assumptions to predict the examination behavior of search users and to infer relevance of results. However, behavior assumptions sometimes do not fit the practical Web search situation and a better estimation requires richer sources of users' interaction information. Mouse movement data is regarded as an important substitute for eye fixation behavior and therefore is adopted in this work to predict examination behavior of users. We incorporate the prediction results into a number of popular click models so that examination is predicted by both mouse movement

information and behavior assumptions. Lab studies show that the prediction algorithm gain promising results and experimental results on large scale practical search behavior data show effectiveness of the revised click models in both homogeneous and heterogeneous environments. The improvement compared to the original model is steady across different models, queries with different frequencies and queries with different search intents. Additionally, we investigated mouse movement's effect on different result positions and find that the predicted examinations seem to be more accurate on high ranks than that on low ranks.

We also acknowledge some limitations in our work, which may be addressed in future studies. The weighted combination method we used to incorporate mouse movement information is not suitable for all click models such as those in which examination probability is not explicitly defined. We also plan to involve more behavior features, especially personalized features into the construction of prediction models in the future. In addition to using the predicted examination probability into click models, one can also incorporate mouse movement events directly into click models, to replace examination. To do this, new assumptions in relation with mouse movements should be defined. We will also investigate this possibility in the future.

**Acknowledgements** This work is supported by Tsinghua University Initiative Scientific Research Program (2014Z21032), National Key Basic Research Program (2015CB358700) and Natural Science Foundation (61472206, 61073071) of China. Part of the work has been done at the Tsinghua-NUS NEXt Search Centre, which is supported by the Singapore National Search Foundation & Interactive Digital Media R&D Program Office, MDA under research Grant (WBS:R-252-300-001-490).

## References

- Arguello, J., & Capra, R. (2014). The effects of vertical rank and border on aggregated search coherence and search behavior. In *CIKM'14* (pp. 539–548). ACM.
- Awadallah, A. H., & Zitouni, I. (2014). Machine-assisted evaluation for search preference judgments. In *CIKM'14* (pp. 51–60). ACM.
- Broder, A. (2002). A taxonomy of web search. In *ACM SIGIR forum 2002* (Vol. 36, pp. 3–10). ACM.
- Buscher, G., Cutrell, E., & Morris, M. R. M. (2009). What do you see when you're surfing?: Using eye tracking to predict salient regions of web pages. In *CHI'09* (pp. 21–30). ACM.
- Buscher, G., Dumais, S. T., & Cutrell, E. (2010). The good, the bad, and the random: An eye-tracking study of ad quality in web search. In *SIGIR'10* (pp. 42–49). ACM.
- Chapelle, O., & Zhang, Y. (2009). A dynamic Bayesian network click model for web search ranking. In *WWW'09* (pp. 1–10). ACM.
- Chapelle, O., Metlzer, D., Zhang, Y., & Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In: *CIKM'09* (pp. 621–630). ACM.
- Chen, M. M. C., Anderson, J. R. J., & Sohn, M. H. M. (2001). What can a mouse cursor tell us more?: Correlation of eye/mouse movements on web browsing. In *CHI'01 Extended Abstracts on Human Factors in Computing Systems* (pp. 281–282).
- Chen, D., Chen, W., Wang, H., Chen, Z., & Yang, Q. (2012). Beyond ten blue links: Enabling user click modeling in federated web search. In *WSDM'12* (pp. 463–472). ACM.
- Chuklin, A., Serdyukov, P., & de Rijke, M. (2013). Using intent information to model user behavior in diversified search. In: *ECIR'13* (pp. 1–13). Springer.
- Claypool, M., Le, P., Wased, M., & Brown, D. (2001). Implicit interest indicators. In *IUI'01* (pp. 33–40). ACM.
- Craswell, N., Zoeter, O., Taylor, M., & Ramsey, B. (2008). An experimental comparison of click position-bias models. In *WSDM'08* (pp. 87–94). ACM.
- Diaz, F., White, R. W., Buscher, G., & Liebling, D. (2013). Robust models of mouse movement on dynamic web search results pages. In *CIKM'13* (pp. 1451–1460). ACM.
- Dupret, G. E., & Piwowarski, B. (2008). A user browsing model to predict search engine click data from past observations. In *SIGIR'08* (pp. 331–338). ACM.

- Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.
- Granka, L. A., Joachims, T., & Gay, G. (2004). Eye-tracking analysis of user behavior in WWW search. In *SIGIR'04* (pp. 478–479). ACM.
- Guan, Z., & Cutrell, E. (2007). An eye tracking study of the effect of target rank on web search. In *CHI'07* (pp. 417–420). ACM.
- Guo, Q., & Agichtein, E. (2008). Exploring mouse movements for inferring query intent. In *SIGIR'08* (pp. 707–708). ACM.
- Guo, Q., & Agichtein, E. (2012). Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behavior. In *WWW'12* (pp. 569–578). ACM.
- Guo, Q., Lagun, D., & Agichtein, E. (2012). Predicting web search success with fine-grained interaction data. In *CIKM'12* (pp. 2050–2054). ACM.
- Guo, F., Liu, C., & Wang, Y. M. Y. (2009a). Efficient multiple-click models in web search. In *WSDM'09* (pp. 124–131). ACM.
- Guo, F., Liu, C., Kannan, A., Minka, T., Taylor, M., Wang, Y. M., & Faloutsos, C. (2009b). Click chain model in web search. In *WWW'09* (pp. 11–20).
- Hu, B., Zhang, Y., Chen, W., Wang, G., & Yang, Q. (2011). Characterizing search intent diversity into click models. In *WWW'11* (pp. 17–26). ACM.
- Huang, J., White, R., & Buscher, G. (2012a). User see, user point: Gaze and cursor alignment in web search. In *CHI'12* (pp. 1341–1350). ACM.
- Huang, J., White, R. R. W., & Dumais, S. (2011). No clicks, no problem: using cursor movements to understand and improve search. In *CHI'11* (pp. 1225–1234). ACM.
- Huang, J., White, R. W., Buscher, G., & Wang, K. (2012b). Improving searcher models using mouse. In *SIGIR'12* (pp. 195–204). ACM.
- Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting click through data as implicit feedback. In *SIGIR'05* (pp. 154–161). ACM.
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87, 329–354.
- Lagun, D., Ageev, M., Guo, Q., & Agichtein, E. (2014). Discovering common motifs in cursor movement data for improving web search. In *WSDM'14* (pp. 183–192). ACM.
- Li, L., Kim, J. Y., & Zitouni, I. (2015). Toward predicting the outcome of an A/B experiment for search relevance. In *WSDM'15* (pp. 37–46). ACM.
- Liu, Y., Chen, Y., Tang, J., Sun, J., Zhang, M., Ma, S., & Zhu, X. (2015a). Different users, different opinions: Predicting search satisfaction with mouse movement information. In *SIGIR'15* (pp. 493–502). ACM.
- Liu, C., Guo, F., & Faloutsos, C. (2009). BBM: Bayesian browsing model from petabyte-scale data. In *KDD'09* (pp. 537–546). ACM.
- Liu, Z., Liu, Y., Zhou, K., Zhang, M., & Ma, S. (2015b). Influence of vertical result in web search examination. In *SIGIR'15* (pp. 193–202). ACM.
- Liu, Y., Wang, C., Zhou, K., Nie, J., Zhang, M., & Shaoping, M. (2014). From skimming to reading: A two-stage examination model for web search. In *CIKM'14* (pp. 849–858). ACM.
- Liu, Y., Zhang, M., Ru, L., & Ma, S. (2006). Automatic query type identification based on click through information. In *AIRS'06* (pp. 593–600). Springer.
- Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein. Structure*, 405(2), 442.
- Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1), 2.
- Navalpakkam, V., Jentzsch, L.L., Sayres, R., Ravi, S., Ahmed, A., & Smola, A. (2013). Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *WWW'13* (pp. 953–964).
- Richardson, M., Dominowska, E., & Ragno, R. (2007). Predicting clicks: Estimating the click-through rate for new ads. In *WWW'07* (pp. 521–530). ACM.
- Rodden, K., Ruthven, I., & White, R. (2007). Exploring how mouse movements relate to eye movements on web search results pages. In *ACM SIGIR Forum*.
- Speicher, M., Both, A., & Gaedke, M. (2013). TellMyRelevance!: Predicting the relevance of web search results from cursor interactions. In *CIKM'13* (pp. 1281–1290). ACM.
- Sushmita, S., Joho, H., Lalmas, M., & Villa, R. (2010). Factors affecting click-through behavior in aggregated search interfaces. In *CIKM'10* (pp. 519–528). ACM.
- Thomas, P., Scholer, F., & Moffat, A. (2013). *What users do: The eyes have it* (pp. 416–427). Berlin: Springer.

- Wang, C., Liu, Y., Wang, M., Zhou, K., Nie, J. Y., & Ma, S. (2015). Incorporating non-sequential behavior into click models. In *SIGIR'15* (pp. 283–292). ACM.
- Wang, C., Liu, Y., Zhang, M., Ma, S., Zheng, M., Qian, J., & Zhang, K. (2013). Incorporating vertical results into search click models. In *SIGIR'13* (pp. 503–512). ACM.
- Xing, Q., Liu, Y., Nie, J.Y., Zhang, M., Ma, S., & Zhang, K. (2013). Incorporating user preferences into click models. In *CIKM'13* (pp. 1301–1310). ACM.
- Zhang, Y., Chen, W., Wang, D., & Yang, Q. (2011). User-click modeling for understanding and predicting search-behavior. In *KDD'11* (pp. 1388–1396). ACM.
- Zhou, K., Cummins, R., Lalmas, M., & Jose, J. M. (2013). Which vertical search engines are relevant? In *WWW'2013* (pp. 1557–1568). ACM.
- Zhang, Y., Park, L. A. F., & Moffat, A. (2010). Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval*, 13(1), 46–69.