Incorporating Vertical Results into Search Click Models

Chao Wang[1], Yiqun Liu[1], Min Zhang[1], Shaoping Ma[1], Meihong Zheng[2], Jing Qian[2], Kuo Zhang[1]
[1] State Key Laboratory of Intelligent Technology and Systems,
[1] Tsinghua National Laboratory for Information Science and Technology,
[1] CS&T Department, Tsinghua University, Beijing, 100084, China P.R.
[2] Department of Psychology, Tsinghua University, Beijing, 100084, China P.R.
chaowang0707@gmail.com

ABSTRACT
In modern search engines, an increasing number of search result pages (SERPs) are federated from multiple specialized search engines (called *verticals*, such as Image or Video). As an effective approach to interpret users' click-through behavior as feedback information, most click models were designed to reduce the position bias and improve ranking performance of ordinary search results, which have homogeneous appearances. However, when vertical results are combined with ordinary ones, significant differences in presentation may lead to user behavior biases and thus failure of state-of-the-art click models. With the help of a popular commercial search engine in China, we collected a large scale log data set which contains behavior information on both vertical and ordinary results. We also performed eye-tracking analysis to study user's real-world examining behavior. According these analysis, we found that different result appearances may cause different behavior biases both for vertical results (local effect) and for the whole result lists (global effect). These biases include: examine bias for vertical results (especially those with multimedia components), trust bias for result lists with vertical results, and a higher probability of result revisitation for vertical results. Based on these findings, a novel click model considering these biases besides position bias was constructed to describe interaction with SERPs containing verticals. Experimental results show that the new Vertical-aware Click Model (VCM) is better at interpreting user click behavior on federated searches in terms of both log-likelihood and perplexity than existing models.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms
Algorithms, Measurement, Experimentation

Keywords
Federated search, Click model, User behavior analysis

1. INTRODUCTION
Millions of users submit queries to search engines every day. As the web search click logs reflect users' preferences regarding search result documents, these logs are considered to be invaluable sources of information for improving search performance. The information stored in click-through behavior data can be used in many research areas such as click-through rate (CTR) prediction, web search ranking, query recommendation, and so on. With the help of these applications, search engine can better help users to satisfy their information needs.

While analyzing click-through data, key concerns include how to construct a click model to interpret users' examination and clicking preferences and how to obtain unbiased document relevance estimations. Much effort has been made on this research topic. State-of-art click models such as user browsing model [3], click chain model [5] and dynamic Bayesian network click model [4] have shown their power in fitting the real-world data and predicting future clicks. Although these existing click models have gained much success in modeling ordinary search results, they were not designed for result lists with non-Web-page results (or *verticals*), which were provided by multiple heterogeneous vertical search engines and incorporated into a large number of SERPs. According to their appearances, we classify vertical results into three categories:

- Text vertical: The text vertical is made up of a few blue links and textual snippets, such as news search results or wiki information shown in SERPs.
- Multimedia vertical: The multimedia vertical is made up of a group of multimedia components, such as video or photo search results.
- Application vertical: The application vertical contains a button or a form embedded into SERPs to help users finish certain tasks, such as a download button or an exchange rate calculator.

Figure 1 shows examples of these verticals. We can see that the vertical results have different layout and presentation forms compared to ordinary search results. It is reasonable to suppose that they may lead to different user examination behavior and click preference. Therefore, most of previous click models, which assume all results are homogeneous, may not describe user behavior on these SERPs correctly. Chen et al. [7] made the first step to model user behavior in vertical results. They found that users were more likely to examine the vertical and the ordinary web documents nearby. They also indicated that users are more likely to be satisfied with vertical results and end the whole search session. However, although they also divided verticals into several kinds, they didn't take into account that different vertical types influence users' behavior differently.

With the help of a popular commercial search engine in China, we collected a large number of log data which contain behavior information on both vertical and ordinary results (see Section 3). By analyzing the logs we found that more than 80% SERPs of this Chinese commercial search engines contain verticals, and different result appearances caused different behavior biases both for the vertical results (local effect) and for the whole result lists (global effect). So when analyzing user behavior for modern search engines, taking verticals into account is very important. We also performed eye-tracking experiments to look into users' actual

Permission to make digital or hard copies of all or part 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 bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR'13, July 28-August 1, 2013, Dublin, Ireland.

Copyright 2013 ACM 978-1-4503-2034-4/13/07...$15.00.

Figure 1. Different kinds of vertical results federated into SERPs (with Bing.com as an example)

examining behaviors on vertical and ordinary results. Based on these findings, a novel click model named Vertical-aware Click Model (VCM) was constructed to take these biases into consideration. The major contributions of this work include:

1. Vertical results are grouped into three categories according to their appearances on SERPs. How users interact with these vertical results and other results on a same SERP are analyzed in terms of both large scale click-through log and laboratory eye-tracking analysis.
2. User behavior bias on different kinds of vertical results are concluded which affect users' examination or click behavior either globally (for the whole SERP) or locally (for a particular vertical result).
3. A novel click model named Vertical-aware Click Model (VCM) is constructed which takes these user behavior biases into consideration. The new model is able to incorporate these biases into a calculable framework and improve click prediction performances.

The rest of the paper is organized as follows: we first provide an overview of related work in Section 2. The user behaviors in click logs are investigated in Section 3. In Section 4, an eye-tracking experiment is designed to see users' examining behavior on different vertical categories. Based on click log analysis and eye-tracking experiment, we incorporate our observations into four bias types and build Vertical-aware Click Model to model users' search behavior in Section 5. Section 6 is devoted to experimental studies. The paper is concluded in Section 7.

2. RELATED WORK

2.1 Click Model

Granka et al. [1] were among the first to carry out eye-tracking experiments to analyze users' decision process as they scan SERPs. Joachims et al. [10, 11] compared the implicit click feedbacks against the explicit relevance judgments and defined *position bias* in users' decision processes, namely, that documents appearing in higher positions always attract more clicks, even when they are less relevant than documents in lower positions. Richardson et al. [2] further proposed *examination hypothesis* to define such kind of position bias in Web search user behavior. Their hypothesis stated that a Web document must be examined before clicked, and user-perceived document relevance was defined conditional probability of being clicked. Craswell et al. [8] further proposed *cascade model* for describing where the first click happened when users linearly looked through search results. More recent works have tried to improve click models by incorporating practical behavior hypotheses. The user browsing model (UBM) [3] states that the probability of examining a document is not a constant but affected by clicks before this document. The dynamic Bayesian networks model (DBN) [4] states that users choose to examine the next document if they are unsatisfied with the clicked document. The click chain model (CCM) [5] assumes that the probability to examine the next document after a click depends on the relevance of the clicked document and user-behavior parameters. Chen et al. [6] proposed a task-centric click model (TCM) to model user click behavior in task level. They indicated that users tend to express their information needs incrementally in a task and thus tend to click fresh documents that are not included in the results of previous queries. They also found that click behavior in industrial search engines is often noisy and proposed a Noise-aware Click Model (NCM) [23] to characterize the noise degree of a click. To a certain extent, these models succeed in interpreting search users' click-through behavior and they also help improve the performance of relevance estimation based on implicit feedback information. More mathematic details of click model related to our work will be presented in section 5.

2.2 Federated Search

Federated search integrates vertical results into web searches. Heterogeneous search results are promising for promoting users' search experiences.

Most prior work focused on predicting which verticals are relevant to a query (*vertical selection*). Diaz et al. [14] first carried out a system to collect news dynamically and aggregated them into web search results. Arguello et al. [15, 16] showed that in vertical selection, query-logs are useful. They proposed a vertical ranking method by the query likelihood given the vertical's query-log language model. They also attempted to reuse training data from a set of existing verticals to obtain a predictive model for a new vertical. K Zhou et al. [19] presented an approach that considers both reward and risk within the task of vertical selection.

Some work focused on merging documents retrieved from multiple ranked lists of selected information sources into a single list (*result merging*). Arguello et al. [17, 18] proposed and three learning based approaches and concluded that the best approaches are those that allow the learning algorithm to learn a vertical-specific relationship between features and relevance. Dzung Hong et al. [20] studied on existing result merging methods and showed that learning a set of combination weights for multiple centralized retrieval algorithms is not flexible enough to deal with heterogeneous information sources. They proposed a mixture probabilistic model to learn more appropriate combination weights with respect to different types of information sources.

2.3 Web Search Biases

When using search engine on the web, a lot of potential biases will affect user's searching behavior. The most famous bias is *position bias* [10, 11] which means documents appearing in higher positions always attracted more clicks. Many click models were proposed to eliminate this influence as shown in Section 2.1. Judit et al. [21] studied user preferences for different orderings of search results and confirmed that users tend to choose one of the first results on the results page. They also observed a *site reputation bias* that pages from well-known sites are considered favorably by the subjects. Beyond the *position bias*, Y Yue et al. [22] quantified the effect of bolded keyword matches in the title and abstracts to measure the attractiveness of search results.

Most of these prior work adopted a lot of information in SERP organization and users' search behavior. However, to the best of our knowledge, few previous studies have considered federated search and user click behavior. Chen et al. [7] made the first step

to construct click models for federated results. They tried to automatically infer heterogeneous documents relevance based on user click behavior. However, they didn't look into different types of vertical results as well as the examination behavior on federated results. In this paper, we make a further step to analyze the influence on users' click preference and examining sequence when adding different types of result presentation forms into SERP and build a click model that more accurately reflects users' real search behaviors.

3. CLICK-THROUGH DATA ANALYSIS

To construct an effective model for federated searches, we look into real-world search user behavior log data and compare ordinary Web results with different kinds of verticals using a number of click-through behavior features.

The adopted search log data come from one of the most popular commercial search engines in China. The data set contains 53,080,107 query sessions with 15,149,469 distinct queries during the time period from April 1st to April 3rd, 2012. At the session level, it contains 19.22% sessions with no vertical result and 80.78% with one or more vertical results. At the query level, 18.43% queries contain no vertical result while 81.57% contain one or more vertical results. We can see that a majority of SERPs contain vertical results, which makes the construction of a vertical-aware click model quite necessary. For the purpose of avoiding the combinatorial effects of multiple vertical results, in this paper we only consider the sessions with one vertical result and leave multi-vertical session processing as future work.

After manually labeling the category of most common tags which this Chinese commercial search engine uses to identify verticals, we found that most verticals fall into one of these three categories: text vertical, multimedia vertical and application vertical. Therefore, we focus our analysis on these kinds of verticals, which are also most popular ones for other search engines as far as we know. In this Chinese commercial search engine, each result page contains at most 10 results (including verticals) by default, unlike some other search engines, such as Bing, which inserts verticals into ordinary results and thus increases the number of results per page. But we don't think this will greatly affect the reliability of results in this paper.

3.1 Global Statistics

Firstly, we study the global influence of different kinds of verticals. Figure 2 shows the amount of sessions in which vertical result is placed from rank 1 to rank 10. We can see that text verticals are almost uniformly distributed to each position except that there are a bit more in position 1 and 9. The amount of sessions for multimedia and application verticals decreases when it comes to a lower ranking position, especially for application verticals. We can see that application vertical is barely placed at rank 10 (only less than 100 cases in our log data set which contains over 53 million sessions). To keep the statistical meaning of our analysis, statistics which base count is less than 1,000 are abandoned in the following part of the paper.

To confirm whether there is a global influence of vertical result to other ordinary web results on a same page, we also look into the average CTR of the first page when vertical result appears. In Figure 3, average CTR means the average click-through rate of all results on the SERP; the dotted line shows the average CTR of SERPs without verticals. From this figure, we observe that the existence of vertical results leads to differences in average CTR. When there is a multimedia vertical in a SERP, the average CTR is higher than that of SERP without verticals. On the contrary, average CTR becomes lower in the existence of application verticals. It indicates that users are more likely to click on some results when there is a multimedia vertical in SERP. The average CTR scores are even higher when multimedia results are placed at positions 6-9, which are perhaps caused by the fact that ordinary results on these relatively lower positions will not be clicked while multimedia ones attract user clicks even in these positions. The drop in average CTR of application verticals may be related with the fact that they directly help most users to finish their tasks without having to look into or clicking on other results.

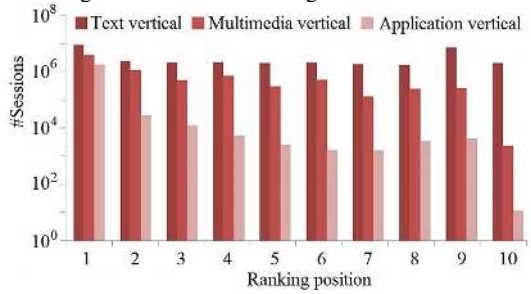


Figure 2. Amount of sessions in which different kinds of vertical results appear from rank 1 to rank 10

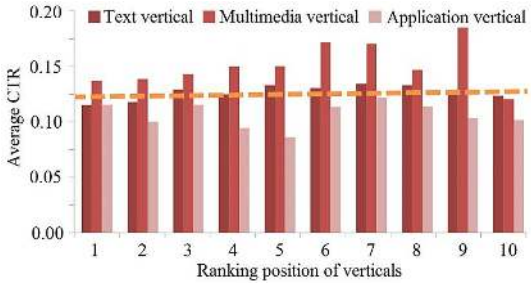


Figure 3. Average CTR of the first page when different kinds of vertical results appear from rank 1 to rank 10

3.2 Click Position Statistics

To further analyze the impact of verticals on SERPs, we compare the click distribution (CD) [9] of both vertical and ordinary results to see whether users are more willing to interact with verticals. Click distribution is defined as the degree of click concentration compared with other results on a same page at query level and it is commonly adopted in click through behavior analysis. We take the average click distribution of each result in SERPs without verticals as the baseline. For each vertical class, we calculate the average click distribution for each position and compare it with the baseline score in the same ranking position.

Heat maps in Figure 4 are used to show the differences between vertical results and baselines using brighter/darker color as a sign for larger/smaller differences, respectively. In this figure, the category axis shows where the vertical result is placed in SERP; while the value axis represents result positions in SERP. So the grid with coordinate (i, j) in this figure shows the j-th result document's CD score when a certain kind of vertical result is placed at the i-th position. Brighter color means a larger difference compared with the CD on SERP without verticals. From the Figure 4(a), we can see that the click distribution of SERP with text vertical is almost the same as SERP with no verticals. Only the CD values for text vertical results themselves (see the diagonal line from upper left to lower right) are slightly higher. Figure 4(b) shows that multimedia vertical result (also see the diagonal line) are with higher CD scores than ordinary results in the same position, which also shows a sign of user preference for this kind of vertical results. Figure 4c shows that the top-ranked application vertical results also get higher click

concentration while application verticals placed at lower positions are with almost the same CD values as ordinary results.

From the above observations in Figure 4 we can see that the CD values of multimedia vertical as well as top-ranked application vertical are higher than those of ordinary results in same positions. When we look into users' click through behavior more detailed, we found that other click behaviors on vertical results are also different from ordinary ones. In Figure 5, first click distribution means how many percentages of users' first clicks are on a certain position. From Figure 5 we can see that when text or application vertical appears at the first position, it attracts a lot more first clicks than ordinary results do. For multimedia vertical, it attracts more clicks not only when it is placed at the top of the result list, but also when it is ranked among the first 5 or 6 results, which usually means in the first screen of results without scrolling.

According to the comparison of CD and first click distribution within different kinds, we can see that user behaves differently on click preference with different kinds of verticals. For multimedia class, users may pay more attention to multimedia vertical result. They are more likely to see the vertical result directly and click it. While for text class and application class, users will pay more attention to vertical result only if the vertical is placed at the top-ranking positions in SERP. This can be regarded a sign for larger examination probabilities for multimedia verticals and top-ranked text/application verticals, which we will show explicitly via eye-tracking statistics in Section 4.

3.3 Click Sequence Statistics

We have shown in Section 3.2 that multimedia verticals as well as top-ranked text/application verticals are higher in the values of either CD or first click contribution. We also want to find out what happens after users click on these vertical results. How they interact with other results on SERPs also plays an important part in the construction of a vertical aware click model.

Compared with the sequential clicking behavior which is already well-defined by existing click models, we focus on the possible revisiting behavior after users click on vertical results. From Figure 4 and 5 we can see that there are only slight chances that text/application verticals are firstly clicked when they are not placed at the 1st position. Therefore, most result revisits after clicking vertical results first should come from queries with multimedia results.

In Figure 6, revisit means that a user clicks on another higher ranked result after having clicked on a lower result. We can see from this figure that top documents before the vertical result cause higher revisit proportion. Figure 7 further shows that the second click after clicking a multimedia vertical tends to be on the 1st ranking position. It means that when users first click a multimedia vertical, there is a large chance that they continue their search session by revisiting the results that they previous skipped.

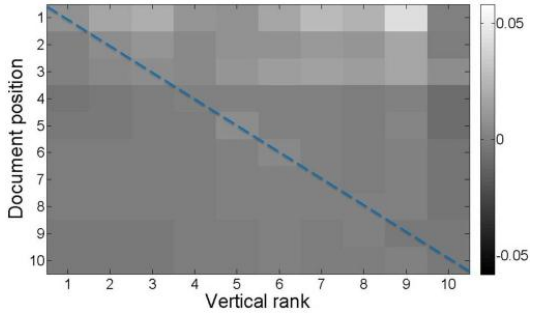
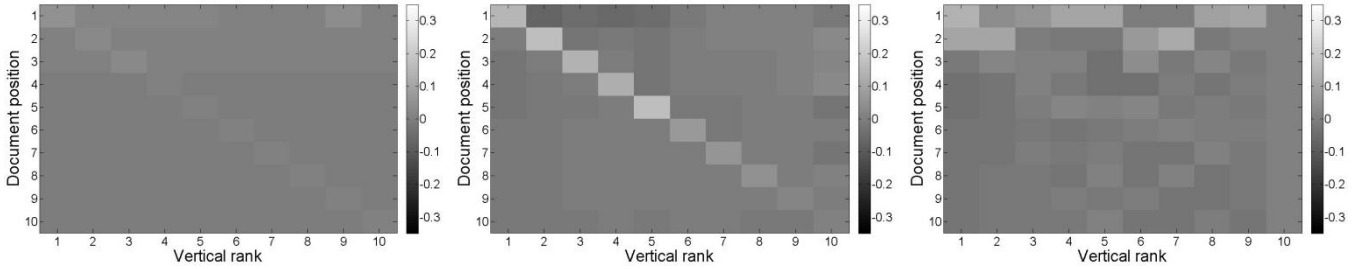


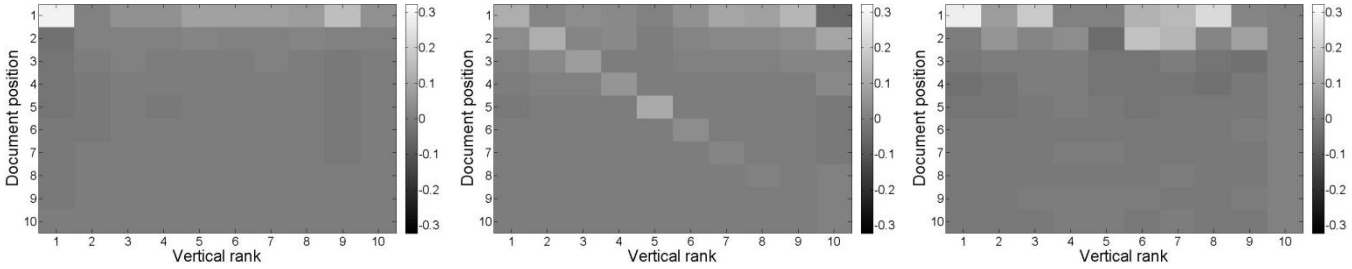
Figure 6. Revisit click distribution when vertical result appears from rank 1 to rank 10 compared with SERPs with no verticals. (Brighter color means a higher revisit probability compared with SERPs with no verticals)

According to the comparison of revisit distribution and second click position distribution to those of the ordinary class, we get a relatively clear picture on how users interact with SERPs with multimedia verticals. We may conclude that for result lists with multimedia vertical results, users have a probability to directly examine and click on the vertical result and then scan back to the top of SERP. It looks like that although multimedia results interrupt the normal search interaction process because they are attractive with image or video contents, people will resume the interrupted session by starting from the top of the result list again.



(a) (b) (c)

Figure 4. Average click distribution of SERPs with (a) text (b) multimedia (c) application vertical when vertical result appears from rank 1 to rank 10 compared with SERPs with no verticals. (Brighter color means a higher CD value)



(a) (b) (c)

Figure 5. Average first click distribution of SERPs with (a) text (b) multimedia (c) application vertical when vertical result appears from rank 1 to rank 10 compared with SERPs with no verticals. (Brighter color means a higher first click probability)

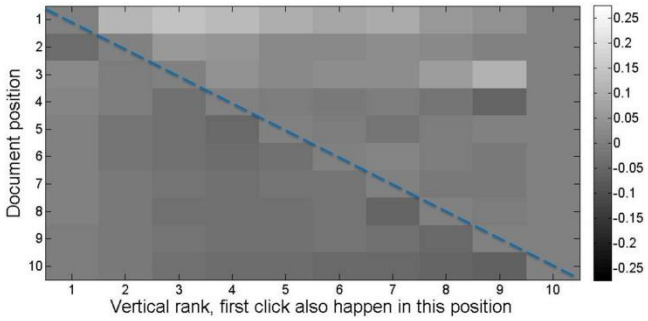


Figure 7. Average second click distribution of SERPs after first clicking a multimedia vertical. (Brighter color means a higher second click probability compared with SERPs with no verticals)

3.4 Log Data Analysis Findings

In summary, we can conclude the influence of vertical results on users' click-through behavior into four aspects:

CT 1. Different verticals have different global influence on users' clicking preference. Multimedia vertical increases global CTR while application vertical decreases global CTR.

CT 2. Click distribution of multimedia vertical results as well as top-ranked application vertical results are higher than ordinary results at same ranking positions.

CT 3. Multimedia vertical which are at the first screen and text/application vertical results which are ranked 1st attract more first clicks than ordinary results.

CT 4. After clicking a multimedia vertical result first, there is probability that users revisit results ranked higher than the vertical starting from the very beginning of the ranking list.

4. EYE-TRACKING ANALYSIS

After analyzing click-through behavior from search engine click logs, we conclude four main influences when different vertical results are federated into SERPs. However, click preference of users can be extracted directly from the click logs, while the examining behavior remains unknown from the click logs. Therefore, we designed an eye-tracking experiment to investigate how users examine SERPs with vertical results.

4.1 Eye-tracking Data Collecting

The following environment is designed to resemble the typical utilization of a WWW search engine. Subjects are asked to search 20 given queries one by one using the same search engine that helped us to collect log data as described in Section 3. Queries are selected from the same click log data set. Among these 20 queries, 5 of them contain no vertical result in their SERPs (ordinary class), the other 15 queries are assigned exactly 1 vertical result in their SERPs. Query No. 6-10 are with text verticals; 11-15 are with multimedia verticals while 16-20 are with application verticals, as shown in Table 1. Each of these queries is also assigned a description text to avoid possible ambiguities. There are no restrictions on subjects' click actions. Subjects are told that the experiment looks into how people search on the Web, but are not told that we were specifically interested in their examining behavior on vertical results of the search engine.

We recruit 23 subjects with cash compensation for their participation. All of these subjects are undergraduate students from a university and indicate at least a general familiarity with search engine interface. Due to the inability of precisely calibrated eye tracking for one subject, comprehensive eye movement data is recorded for 22 of them. The gender distribution is split between 18 males and 5 females (typical for most departments in the university).

Table 1. Search queries adopted in the eye-tracking experiment (N, I and T represent Navigational, Informational and Transactional queries, respectively)

Query	Description	Intent	Type
4399 弹弹堂	An online game site	N	Ordinary
王立军最新消息	Recent information about Lijun Wang	I	
武道至尊免费下载	Free download of a novel	T	
大连实德总裁亲戚名单	Family information of a famous entrepreneur	I	
吞噬星空最新章节列表	Recent update of a science fiction novel	T	
重生之鸿蒙无极神诀	The name of a novel	T	Text Vertical
日本预测 9 级地震	Earthquake prediction for Japan	I	
优酷网看古装电视剧婚姻向右浪漫向左	A TV series at youku.com	T	
北京出事谣言	Political rumors in Beijing	I	
传奇私服发布网	An online game	N	
果宝特攻第二部全集	Resource of a Comic TV series	T	Multi-media vertical
马布里老婆的照片+图	Picture of a famous CBA basketball player's wife	I	
qq 头像闪图	Multimedia plugin for an online chatting software	T	
葫芦娃全集	A Chinese Comic TV series	T	
wwe 美国职业摔角 sd	Video of WWE	T	
快播播放器下载	Software download	T	Appli-cation Vertical
百度影音	Software download	T	
火车票网上订票官网	Official website for train tickets reservation	N	
vagaa	Software download	T	
飞信下载 2012 正式版官方下载	Software download	T	

The manipulations to the result pages were performed by a transfer server which shows exactly the same SERPs as the original search engine except the domain name. The server automatically eliminated all advertising content, so that the SERPs for all subjects would look as uniform as possible, with approximately the same amount of results appearing within the first scroll screen. When a subject searches for a query, if the SERP contains a vertical result, the vertical result position will be randomly placed by the server at:

1. Position 1 (top of the first screen).
2. Position 3 (middle of the first screen).
3. Position 5 (bottom of the first screen).
4. Position 10 (out of the first screen and bottom of the first SERP which can be seen only by mouse scrolling down).

None of the changes were detectable by the subjects. While being asked after their query sessions, none of the subjects suspect any manipulation. This server is also used to log all click-through behavior and all SERPs subjects visit.

All subjects' eye movements are recorded using a SMI RED250 eye-tracker, which utilizes infrared to reconstruct the subjects' eye position. BeGaze Experimental Center is used for the simultaneous acquisition and analysis of the subjects' eye movements. With this tracking device, the following indicators of ocular behaviors are recorded: fixations, saccades, pupil dilation, and scan paths [12]. Among these behaviors, we focus on eye fixation, which is the most relevant metric for evaluating information processing in online search. In this paper, eye fixation is defined as a spatially stable gaze lasting for approximately 200-300 milliseconds, during which visual attention is directed to a specific area of the visual display.

4.2 Do Users Examine Verticals First?

To analyze which result the user first pays attention to, we collect subjects' first two seconds eye fixations on the screen. Figure 8 shows two samples from eye-tracking data which shows users' watching area on SERP with different kinds of verticals or no vertical results. From this figure we can see that users pay most attention to the first result when there is no vertical in SERP (which should be regarded as sign for position bias). However, when there is a multimedia vertical result at the third position, it attracts a lot of users' direct attentions.

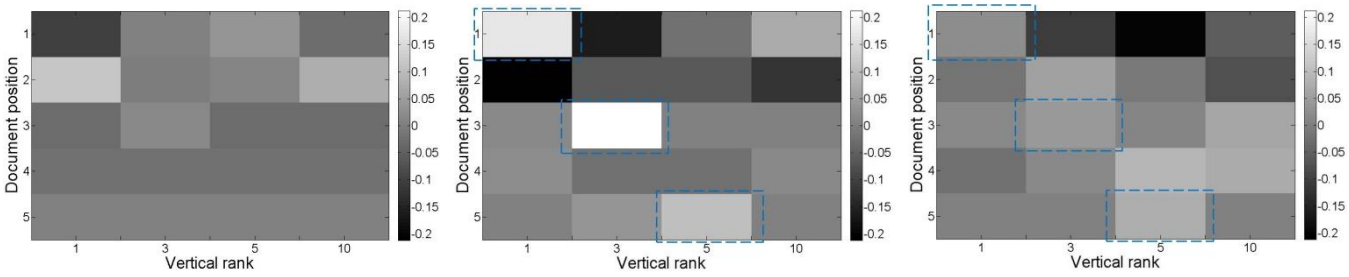
We set 250 milliseconds as the threshold of fixation action and labeled each document's boundary manually. Then we can record each subject's eye examining sequence on document results for each SERP. This statistical result shows users' examining sequential behaviors. We compared users' first examining behavior for each vertical class using the same form with users' first click distribution we analyzed in previous section. Figure 9 shows the subjects' first examining distribution at each document position for each vertical class compared with no vertical situation. From Figure 9(b) we can see that when multimedia vertical is placed on the first screen (rank 1, 3 and 5); it actually attracts more attention than ordinary results. This conclusion is consistent with CT 2 in section 3.5. For application verticals, Figure 9(c) shows that application vertical attracts slightly more attention than ordinary result. Meanwhile Figure 9(a) shows that text vertical doesn't attract much user attention.

4.3 Behavior after Examining Verticals First

To validate the findings in Section 3 about how users behave after clicking a vertical results first. We look into the examining sequences of subjects when they examine vertical results firstly. Two typical examining sequential patterns are extracted from eye-tracking data and shown in Figure 10. We can see from this figure that users examine results above verticals either sequentially (top-down) or from the ones next to verticals first (bottom-up). Further statistics in Table 2 show that most of the subjects examine back to top results after examining a vertical first. This finding accords with our assumption in Section 3.3 that users will resume a top-down search session after being interrupted by a vertical result.

(a) **(b)**

Figure 8. Heat map of the subjects' eye fixation areas in first 2 seconds on (a) SERP with no vertical (b) SERP with multimedia vertical placed at the 3rd position



(a) **(b)** **(c)**

Figure 9. First examining distribution of SERPs with (a) text (b) multimedia (c) application vertical when vertical result is placed at rank 1, 3, 5 and 10 compared with SERPs with no verticals. (Brighter color means a higher first examining rate on document position. We don't show document positions from 6 to 10 here because almost no subjects examine results located on them first)

(a)

(b)

Figure 10. Typical eye-tracking cases of examining previous document (a) bottom up (b) top down after first examining verticals. The examining sequence is from green circles, yellow circles (dashed border) to red circles (solid border)

Table 2 shows the proportion of each examining patterns when user first examines verticals. We can see that after examining vertical result first, most subjects (89% users who examine 3rd result first and 100% users who examine 5th result first) scan back to the previous results. This shows that users may be attracted by the vertical's presentation and change their examination sequence; meanwhile results on top of the ranking list are always valued and not omitted by users.

Table 2. Proportion of different examining behaviors after user first examines a vertical result

Vertical Rank	#Subject[1]	Next	Previous	Back to Top
3	9	0.11	0.22	0.67
5	2	0.00	0.00	1.00

4.4 Eye-tracking Analysis Findings

In summary, we can conclude the main influence of vertical results on users' examining behavior into two aspects:

ET 1. Multimedia and application vertical results are examined more frequently compared with ordinary web results.

ET 2. After examining a vertical result first, most users will scan back to examine the previous results before the vertical either bottom-up or top-down.

[1] 11 subjects out of all 22 subjects examine vertical result first when vertical is placed at the 3rd or 5th position. We don't consider sessions in which verticals are placed at the 1st position because there would be no "previous" or "back to top" patterns.

5. VERTICAL-AWARE CLICK MODEL

We first state some definitions and notations that will be used in the following part. A search session within the same query is called a *query session*. A web search user initializes a *query session s* by submitting a *query q* to the search engine. The SERP can be represented as $D = (d_1, \dots, d_M)$ sequentially (M = 10 if we only consider the first search result page), where d_i is document at position i from the top of the page.

Examination, click and document relevance are treated as probabilistic events. In particular, for a given query session, we use binary random variables E_i, C_i and A_i to represent the examination, click and document attractiveness events of the document at position i. The corresponding, examination and click probabilities for position i are denoted by $(E_i = 1)$, $P(C_i = 1)$ and $P(A_i = 1)$, respectively.

5.1 Preliminaries

We first introduce two important hypotheses: *examination hypothesis* and *cascade hypothesis*, which are the foundations of most existing click models.

The *examination hypothesis* [13] can be summarized as follows:

$$P(C_i = 1 | E_i = 0) = 0 \quad (1)$$

$$P(C_i = 1 | E_i = 1) = r_{d_i} \quad (2)$$

where $i = 1, \dots, M$. r_{d_i} is defined as the document relevance, which is the conditional probability of a click event after examination. Given E_i, $C_i$is conditionally independent on previous examine/click events.

The *cascade hypothesis* in [8] assumes that users always begin the examination at the first document. The examination is strictly linear from top to bottom of the search result page, so a document is examined only if all previous documents are examined:

$$P(E_1 = 1) = 1 \quad (3)$$

$$P(E_{i+1} = 1 | E_i = 0) = 0 \quad (4)$$

Given E_i, E_{i+1} is conditionally independent of all examine/click events above i, but may depend on the click C_i.

The user browsing model (UBM) [3] is based on the examination hypothesis, but it doesn't follow the cascade hypothesis. Instead, it assumes that the examination probability E_i depends on its own position and the previous clicked position l_i:

$$P(E_i = 1 | C_{1:i-1}) = \gamma_{i,i-l_i} \quad (5)$$

Given click $C_{1:i-1}$, E_i is conditionally independent of all previous examination events$E_{1:i-1}$. If there is no click before i, l_i is set to 0. The probability of a query session under UBM is:

$$P(C_{i:M}) = \prod_{i=1}^{M} (r_{d_i} \gamma_{i,i-l_i})^{C_i} (1 - r_{d_i} \gamma_{i,i-l_i})^{1-C_i} \quad (6)$$

5.2 Modeling Biases

We can see from Section 3 and 4 that users treat verticals differently from ordinary results. Now, we want to develop an effective click model for federated search containing both vertical and ordinary results. Notice that we only consider the situation that only one vertical appears in the SERP. Therefore, if there are two or more verticals in SERP, we only keep the first vertical and simply regard others as ordinary results.

5.2.1 Attraction Bias

According to the conclusions of ET 1 and CT 3 described in previous sections, certain vertical result (e.g. multimedia vertical) will attract users' attention directly and cause users to examine and thus click it first. So we formalize the assumption as:

Assumption 1 (Attraction Bias): If there is a vertical placed in the SERP, there is probability that users examine it first.

We use the binary random variable F to represent the event of examining the vertical result first. Thus, the bias can be summarized as

$$P(F = 1) = \phi_{t_v,l_v} \quad (7)$$

where t_v is the class of the vertical result and l_v is the position of vertical result. $\{\phi\}$ is a group of global parameters that should be estimated according to different kinds and positions of verticals.

5.2.2 Global Bias

From the conclusions of CT 1 and CT2 on user's clicking behavior described in section 3, we formalize the assumption about global preference for different vertical results as follows:

Assumption 2 (Global Bias): If there is a vertical placed in the SERP and user examines it first, the user will have a global impression on the whole page, which will affect user's examining and click probability of all results in the SERP.

This bias can be summarized as:

$$P(E_i = 1|F = 1) = P(E_i = 1|F = 0) + \theta_{q,i} \quad (8)$$

$$P(A_i = 1|E_i = 1, F = 1) = P(A_i = 1|E_i = 1, F = 0) + \beta_{q,i} \quad (9)$$

$\{\beta\}$ and $\{\theta\}$ are a group of parameters that represent the additional global impression of each document when users examine the vertical result first.

5.2.3 First Place Bias

From the conclusion of CT 3 in Section 3, we found that when text or application verticals are placed at the first place, users will click on these verticals much more than ordinary results. Meanwhile, because users are likely to be satisfied by the verticals, they may not click on other results any more. Therefore, we can conclude the following bias:

Assumption 3 (First Place Bias): If there is a vertical placed in the SERP and the vertical is placed at the first position, there is probability that users click more on these verticals and less on other results.

We may use another group of parameters to describe the additional influence caused by this bias. However, this group of parameters will simply occur in the same place as $\{\beta\}$ and $\{\theta\}$. Thus, $\{\beta\}$ and $\{\theta\}$ is sufficient to describe the *global bias* and the *first place bias* simultaneously.

5.2.4 Sequence Bias

In section 3 we found that users may revisit (CT 4) after clicking a vertical first. According to the eye-tracking analysis (ET 2), we also found that most users will scan back to examine the previous results either bottom-up or top-down after examining the vertical first. So we summarize the points above and make a non-sequential examining assumption as follow:

Assumption 4 (Sequence Bias): If there is a vertical placed in the SERP and user examines it first, after examining the vertical result users will scan back to the previous documents in either bottom up or top down sequence.

We use binary random variable B to represent the event of examining the previous document in bottom up sequence. Suppose that the SERP's document list $D = (d_1, \dots, d_M)$. After examining the vertical first, if the user decides to scan in top down sequence, the following examining sequence is $d_1, \dots, d_M$; if the user decides to scan in bottom up sequence, the following examining sequence of this SERP will change to $d_{l_v}, \dots, d_1, d_{l_v+1}, \dots, d_M$. So the bias can be summarized as:

$$P(B = 1|F = 0) = 0 \quad (10)$$

$$P(B = 1|F = 1) = \sigma_{t_v,l_v} \quad (11)$$

$\{\sigma\}$ is a group of global parameters and can vary according to different kinds and positions of verticals.

We use the four biases above to represent conclusions we made in click-through log analysis and eye-tracking analysis. Then we propose a click model named Vertical-aware Click Model to take these biases into account.

5.3 Vertical-aware Click Model

The biases above can embrace the assumption of most existing click models depending on the examination hypothesis. Therefore, our Vertical-aware Click Model (VCM) can be constructed based on many existing click models (e.g., the UBM and DBN)

We choose the UBM as a basis in the following experiment. When $F = 0$, the examining and click probability is the same as UBM. The global impression parameters have been introduced into the formulas for $F = 1$. Thus, the VCM can now be summarized as:

$$P(C_i = 1|E_i = 0) = 0 \quad (12)$$

$$P(C_i = 1|E_i = 1) = P(A_i = 1|E_i = 1) \quad (13)$$

$$P(F = 1) = \phi_{t_v,l_v} \quad (14)$$

$$P(E_i = 1|F = 0, C_{1:i-1}) = \gamma_{i,i-l_i} \quad (15)$$

$$P(E_i = 1|F = 1, C_{1:i-1}) = \gamma_{i,i-l_i} + \theta_{q,i} \quad (16)$$

$$P(A_i = 1|E_i = 1, F = 0) = \alpha_{q,i} \quad (17)$$

$$P(A_i = 1|E_i = 1, F = 1) = \alpha_{q,i} + \beta_{q,i} \quad (18)$$

$$P(B = 1|F = 0) = 0 \quad (19)$$

$$P(B = 1|F = 1) = \sigma_{t_v,l_v} \quad (20)$$

Figure 11 shows the decision-making process of VCM. When user begins with a query session, the user will have the opportunity to examine the vertical first if there is a vertical result in SERP. After examining the vertical first, the user will decide to scan back to the previous document in bottom up sequence or top down sequence.

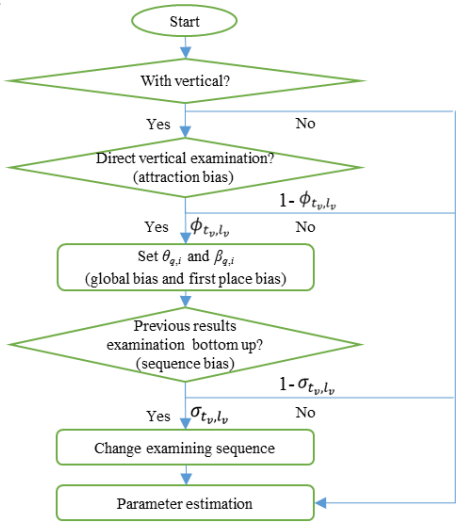


Figure 11. Decision-making process of VCM

5.4 Model Inference

We use the expectation-maximization (EM) algorithm to complete the inference step. The EM algorithm is used to find the maximum likelihood estimates of parameters, including the attraction bias parameters ϕ_{t_v,l_v}, the global bias and first place bias parameters $\beta_{q,i}$ and $\theta_{q,i}$, the sequence bias parameters σ_{t_v,l_v}, the examining probability parameters $\gamma_{i,i-l_i}$ and the document

relevance parameters $\alpha_{q,i}$. The EM iteration alternates between performing an E-step, which creates a function for the expectation of the Log-Likelihood evaluated using the current estimate for the parameters, and M-step, which computes parameters maximizing the expected *Log-Likelihood* found on the E-step.

The traditional EM algorithm leads to the following updating formula, here we show how to update $\alpha_{q,i}$; the corresponding formulas for the other parameters can be derived analogously.

Supposing that there are N sessions and C_i^j denotes the click action of i-th document associated with the j-th session. q^j is the search query associated with the j-th session. A t superscript indicates the estimate at iteration t:

$$I_i^j = I(C_i^j = 1) \tag{21}$$

$$P_q^j = \prod_{i=1}^{M} \left(\alpha_{q,i}^t \gamma_{i,i-l_i}^t\right)^{I_i^j} \left(1 - \alpha_{q,i}^t \gamma_{i,i-l_i}^t\right)^{1-I_i^j} \tag{22}$$

$$A_{q,i} = \sum_{j=1}^{N} I(q^j = q)\, P_q^j \left[I_i^j + (1 - I_i^j) \times \frac{\alpha_{q,i}^t \left(1-\gamma_{i,i-l_i}^t\right)}{1-\alpha_{q,i}^t \gamma_{i,i-l_i}^t}\right] \tag{23}$$

$$B_{q,i} = \sum_{j=1}^{N} I(q^j = q)\; P_q^j \left[(1 - I_i^j) \times \frac{(1-\alpha_{q,i}^t)}{1-\alpha_{q,i}^t \gamma_{i,i-l_i}^t}\right] \tag{24}$$

where $I(\cdot)$ is the indicator function, and $\alpha_{q,i}^{t+1} = \frac{A_{q,i}}{A_{q,i}+B_{q,i}}$.

6. EXPERIMENTS AND DISCUSSIONS

In this section, we compare the VCM with UBM model with click perplexity and log-likelihood as metrics to measure the effectiveness of these two click models. UBM is selected as our baseline because we want to confirm whether the new model VCM in which four biases is added can better interpret user click behavior than the original model or not.

6.1 Experiment Setups

The click logs used for training and testing click models are sampled from a popular Chinese commercial search engine during a week in April 2012. To prevent the evaluations from being biased by extremely high-frequency queries, we allow each query at most 10^4 sessions. For each query, we sort its sessions by timestamp information and split sessions into the training and testing sets at a ratio of 4 : 1. Altogether 306,750 queries and 11,558,016 sessions were collected and their query frequency distributions are shown in Table 3.

Table 3. Query frequency distribution of experiment data set

Query Frequency	# Queries	# Sessions
1-10	228,290	688,129
10-$10^{1.5}$	43,280	777,642
$10^{1.5}$-10^2	21,060	1,157,448
10^2-$10^{2.5}$	9,103	1,573,706
$10^{2.5}$-10^3	3,341	1,802,170
10^3-$10^{3.5}$	1,140	1,980,876
$10^{3.5}$-10^4	536	3,578,045

For training the baseline model UBM, we use the inference algorithms introduced in the original paper [3]. For VCM, we use the inference method introduced in Section 5. If there are two or more verticals in SERP, we only keep the first vertical and simply regard others as ordinary results.

As for evaluation metrics, perplexity and log-likelihood were adopted by a number of previous works (e.g. [3, 5, 7]). In our experiment we also use these two metrics to show effectiveness of VCM compared with the original UBM.

Perplexity measures the accuracy for each position instead of the whole session. It is computed for binary click events at each position in a query session independently. The perplexity of the entire dataset is the average of p_i over all positions. A smaller value indicates better prediction accuracy, and perfect click prediction will have a perplexity of 1.0000. The improvement of perplexity value p_1 over p_2 is given by $(p_2 - p_1)/(p_2 - 1)$ .Log-likelihood (LL) is also widely used to measure model fitness. Given the document impression for each query session in the test data, LL is computed as the average log probability of observed click events under the trained model. A larger LL indicates better performance, and the optimal value is 0. The improvement of LL value l_1 over l_2 is computed as $(exp(l_1 - l_2) - 1)$.

6.2 Results and Discussions

Figure 12 presents perplexity scores for both UBM and VCM over different positions. Average click perplexity over all positions for VCM is 1.2792, which is 14.06% better than that of UBM (1.3249). The improvement is significant and almost adequate in all ranking positions, which indicates that biases introduced in our vertical-aware click model can learn a better accuracy at all positions compared with UBM. Table 4 also shows comparison results for different vertical types, from which we can see that the improvement of VCM in multimedia and application vertical is larger. This phenomenon is reasonable because multimedia and application vertical results have significant different appearances compared with ordinary results and user click/examination behavior on these two types of verticals are also different from ordinary ones according to Section 3 and 4. It indicates that VCM better models user behavior on these vertical results by incorporating more biases besides position bias.

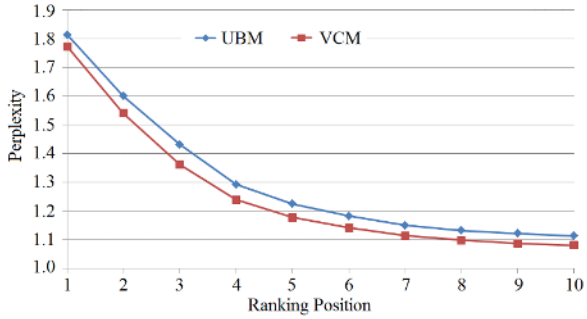


Figure 12. Perplexity comparison of UBM and VCM for results in different ranking positions

Table 4. Perplexity comparison of UBM and VCM for queries with different vertical types

	UBM	VCM	VCM Improvement
Text vertical	1.2266	1.2139	5.58%
Multimedia vertical	1.3735	1.3071	17.78%
Application vertical	1.1908	1.1601	16.09%
Without vertical	1.2388	1.2285	4.33%

Figure 13 and Table 5 present log-likelihood comparison results for VCM and UBM for queries with different frequencies. The overall LL result for VCM is -3.1128, which is 19.46% better than that of UBM (-3.3005). Similar with the perplexity results, it also shows that VCM outperforms UBM for almost all kinds of queries. Especially, Figure 13 shows that the improvement is even larger for low-frequency queries (or tailed queries). The average LL score of VCM for queries with less than 100 appearances improves UBM by 57.89%. Queries with multimedia results benefit the most among all vertical types according to Table 5,

which is similar with the results shown in Table 4.

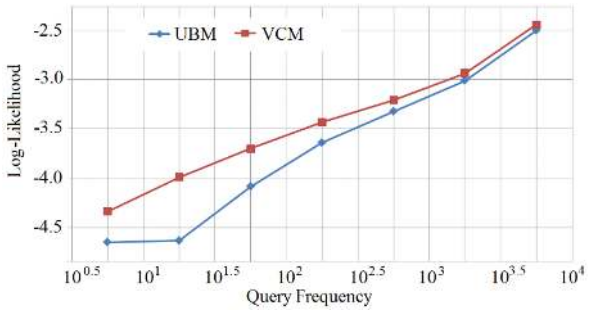


Figure 13. Log-likelihood comparison of UBM and VCM for queries with different frequencies

Table 5. Log-likelihood comparison of UBM and VCM for queries with different vertical types

	UBM	VCM	VCM Improvement
Text vertical	-2.9093	-2.7968	11.90%
Multimedia vertical	-4.1142	-3.8638	28.44%
Application vertical	-2.2671	-2.1427	13.24%
Without vertical	-3.0256	-2.9646	6.29%

The experimental results show that VCM which take four biases into consideration can better interpret user click behavior than the original UBM in terms of both perplexity and Log-likelihood. As the introduced parameters $\{\beta\}$ and $\{\theta\}$ can better interpret the additional influence brought by vertical, the original examine probability $\{\gamma\}$ is more close to the situation without vertical in SERP. Therefore, VCM can also interpret slightly better than original UBM even for session without verticals.

7. CONCLUSIONS AND FUTURE WORK

Nowadays vertical results appear in over 80% SERPs of commercial search engines. In order to solve the problem of incorporating vertical results into search click models, we look into both large scale click-through log collected from a popular search engine and laboratory based eye-tracking data of 22 participants. We found that click-through and result examining behaviors between SERPs with and without verticals are different from each other. Such behaviors are even different for verticals with different presentation forms such as text vertical, multimedia vertical and application vertical. A number of behavior differences are concluded into four biases: attraction bias, global bias, first place bias and sequence bias. A click model named Vertical-aware Click Model (VCM) is constructed and its effectiveness is evaluated in term of perplexity and log-likelihood.

In the future, we would like to extend the VCM model to cover SERPs with multiple vertical results. We will also work on incorporating VCM into a ranking model to improve ranking performance of search result lists with vertical results.

8. ACKNOWLEDGMENTS

This work was supported by Natural Science Foundation (60903107, 61073071), National High Technology Research and Development (863) Program (2011AA01A205) of China and Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

9. REFERENCES

[1] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in WWW search. In SIGIR'04, pages 478-479, 2004.

[2] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In WWW '07, pages 521–530, 2007.

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

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

[5] F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y. Wang, and C. Faloutsos. Click chain model in web search. In WWW'09, pages 11–20, 2009.

[6] Y. Zhang, W. Chen, D. Wang, and Q. Yang. User-click modeling for understanding and predicting search-behavior. In KDD 2011, pages 1388-1396, 2011.

[7] D. Chen, W. Chen and H. Wang. Beyond ten blue links: enabling user click modeling in federated web search. In WSDM 2012, pages 463-472, 2012.

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

[9] U. Lee, Z. Liu and J. Cho, Automatic Identification of User Goals in Web Search, in the WWW'05, pages 391-400, 2005

[10] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In SIGIR '05, pages 154–161, 2005.

[11] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. ACM Trans. Inf. Syst., 25(2):7, 2007.

[12] K. Rayner. Eye movements in reading and information processing. Psychological Bulletin, 124:372–252, 1998.

[13] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In WWW '07, pages 521–530, 2007.

[14] F. Diaz. Integration of news content into web results. In WSDM 2009, pages 182-191, 2009.

[15] J. Arguello, F. Diaz and J. Callan. Sources of evidence for vertical selection. In SIGIR'09, pages 315-322, 2009.

[16] J. Arguello, F. Diaz. Vertical selection in the presence of unlabeled verticals. In SIGIR 2010, pages 691-698, 2010.

[17] J. Arguello, F. Diaz. Learning to aggregate vertical results into web search results. In CIKM 2011, pages 201–210, 2011.

[18] J. Arguello, F. Diaz and J. Callan. A methodology for evaluating aggregated search results. In Proceedings of ECIR 2011. Springer-Verlag, pages 141–152.

[19] K Zhou, R Cummins. Evaluating Reward and Risk for Vertical Selection. In CIKM 2012, pages 2631-2634, 2012.

[20] Dzung Hong, Luo Si. Mixture model with multiple centralized retrieval algorithms for result merging in federated search. In SIGIR 2012, pages 821-830, 2012.

[21] Judit Bar-Ilan, Kevin Keenoy and Mark Levene. Presentation bias is significant in determining user preference for search results—A user study. In Journal of the American Society for Information Science and Technology, pages 135–149, 2009.

[22] Y Yue, R Patel, H Roehrig. Beyond position bias: examining result attractiveness as a source of presentation bias in clickthrough data. In WWW 2010, pages 1011-1018, 2010.

[23] W. Chen, D. Wang and Y. Zhang. A Noise-aware Click Model for Web Search. In WSDM 2012, pages 313-322, 2012.