Characterizing Search Intent Diversity into Click Models

Botao Hu[1,2]*, Yuchen Zhang[1,2]*, Weizhu Chen[2,3], Gang Wang[2], Qiang Yang[3]
Institute for Interdisciplinary Information Sciences, Tsinghua University, China[1]
Microsoft Research Asia, Beijing, China[2]
Hong Kong University of Science and Technology, Hong Kong[3]
{botao.a.hu, zhangyuc}@gmail.com, {wzchen,gawa}@microsoft.com,
{wzchen,qyang}@cse.ust.hk

ABSTRACT

Modeling a user's click-through behavior in click logs is a challenging task due to the well-known position bias problem. Recent advances in click models have adopted the examination hypothesis which distinguishes document relevance from position bias. In this paper, we revisit the examination hypothesis and observe that user clicks cannot be completely explained by relevance and position bias. Specifically, users with different search intents may submit the same query to the search engine but expect different search results. Thus, there might be a bias between user search intent and the query formulated by the user, which can lead to the diversity in user clicks. This bias has not been considered in previous works such as UBM, DBN and CCM. In this paper, we propose a new *intent hypothesis* as a complement to the examination hypothesis. This hypothesis is used to characterize the bias between the user search intent and the query in each search session. This hypothesis is very general and can be applied to most of the existing click models to improve their capacities in learning unbiased relevance. Experimental results demonstrate that after adopting the intent hypothesis, click models can better interpret user clicks and achieve a significant NDCG improvement.

Categories and Subject Descriptors

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

General Terms

Algorithm

Keywords

Click Model, Intent Bias, Intent Diversity, Search Engine, User Behavior

1. INTRODUCTION

Click-through logs record user activities on search pages and encode user preferences of search results. Click-through logs can be collected at a very low cost, and the analysis of them can help to understand the user's latest preference tendencies. Naturally, many studies have attempted to discover user preferences from click-through logs to improve the relevance of search results [12, 11, 1].

It is well known that clicks are "informative but biased" [4], and it is a challenging task to estimate unbiased relevance from click-through logs. One typical bias affecting user clicks is the so-called position bias: a document appearing in a higher position is more likely to attract user clicks even though it is not as relevant as other documents in lower positions. Thus, the click-through rate is not a proper measure of relevance. This bias was first noticed by Granka et al. [7] in their eye-tracking experiment and some follow-up investigations have been made to alleviate this bias so that the unbiased relevance can be inferred from the clicks. Richardson et al. [16] proposed to increase the relevance of the documents in lower positions by using a multiplicative factor. This idea was later formalized as the examination hypothesis [4], which assumes that the user will click a search result only after examining its search snippet. In other words, given an examined document, only its relevance determines the user click. The examination hypothesis decouples document relevance from position bias where the position bias is formulated as the probability that a document is examined by a user. Recently, many interesting studies have been made to refine click models using the examination hypothesis. UBM[6], DBN[3], CCM[8], BBM[13], GCM[19] are typical models which can extend the capabilities of the examination hypothesis.

The examination hypothesis assumes that, if a document has been examined, the click-through rate of the document for a given query is a constant number whose value is determined by the relevance between the query and the document. We argue that users with different search intents, however, may submit the same query to the search engine. In other words, a single query may not truly reflect user search intent. Take the query "iPad" as an example. A user submits this query because she wants to browse general information about iPad, and the results from `apple.com` or `wikipedia.com` are attractive to her. In contrast, another user who submits the same query may look for information such as user reviews or feedback on iPad. In this situation, search results like technical reviews and discussion forums are more likely to be clicked. This example indicates that the attractiveness of a search result is not only influenced by its relevance but also determined by the user's intrinsic search intent behind the query.

*This work was done when the authors were interns at Microsoft Research Asia.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0632-4/11/03.

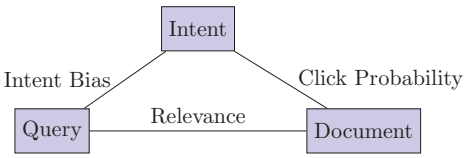


Figure 1: The triangular relationship among intent, query and document. The edge connecting two entities measures the degree of match between two entities.

We design an experiment to validate that the relevance between a query and a document is not a constant number. In the experiment, we collect search sessions, and partition them into two groups according to different search intents. We note that after eliminating the position bias effect, most queries (96.6%) have significantly different click-through rates on two intent groups. (Please refer to Section 3 for the experimental details.) In other words, the click-through rate of an examined document varies greatly across different search sessions due to the diversity in search intent.

Figure 1 describes the triangular relationship among intent, query and document, where the edge connecting the two entities measures the degree of match between them. Each user presumably has an intrinsic search intent before submitting a query. When a user comes to a search engine, she formulates a query according to her search intent and submits it to the search engine. The *intent bias* measures how well the query matches the intent, i.e., the degree of match between the intent and the query. The search engine receives the query and returns a list of ranked documents, while the *relevance* measures the degree of match between a query and a document. The user examines each document and, if a document better satisfies her information need, she is more likely to click this document.

The triangular relationship suggests that the user click is determined by both intent bias and relevance. If a user does not clearly express her information need in the input query, there is a large intent bias. Thus, the user is unlikely to click the document that does not meet her search intent, even if the document is very relevant to the query. The examination hypothesis can be considered as a simplified case, that it regards the search intent and the input query as equivalent and ignores the intent bias. Thus, the relevance between a query and a document may be mistakenly estimated when only the examination hypothesis is adopted.

In this paper, we incorporate the concepts of intent and intent bias to propose a novel hypothesis, the *intent hypothesis*, to explain how user clicks are affected by intent bias, relevance and position bias. The intent hypothesis can enhance the analytical power of the examination hypothesis, characterize search intent diversity and interpret user clicks better. Click models that adopt the intent hypothesis can estimate more accurate and unbiased relevance.

This paper's contributions are four-fold. First, we empirically demonstrate the limitations of the examination hypothesis and suggest that the position bias is not the only bias affecting click behavior. Second, we propose the novel intent hypothesis to enhance the capability of modeling user search behavior. Third, because the intent hypothesis is general, we apply it to two typical click models, UBM and DBN, and adopt a Bayesian inference method to model the intent hypothesis. This inference method is capable of learning on very large scale click-though logs. Finally, the experiment has been conducted on 3.6 million queries and one billion search sessions, and the results illustrate the advantages of adopting the intent hypothesis.

This paper is organized as follows: In Section 2, we briefly review the previous research on click models including their specifications and hypotheses. In section 3, we empirically validate that the examination hypothesis can not well interpret real click-through data. In Section 4, we propose our intent hypothesis and its inference method. In Section 5, the experiment on real datasets shows the advantages of adopting the proposed hypothesis. In Section 6, we analyze the intent bias that we estimated from the experiment and discover some insightful results.

2. BACKGROUND

We start by introducing definitions and notations which will be used throughout the paper. A user submits a *query* q and the search engine returns a *search result page* containing M (usually $M = 10$) documents, denoted by $\{d_{\pi_i}\}_{i=1}^{M}$, where π_i is the index of the document at the i-th position. The user *examines* the summary of each search result and *clicks* some or none of them. Here the *summary* includes the search title, snippets and URL. A search session within the same query is called a *search session*, denoted by s. Clicks on sponsored ads and other web elements are not considered in one search session. We regard subsequent query re-submission or re-formulation as a new session. The terms *url*, *document* and *result* have the same meaning, and we use them indiscriminately in the context.

We define three binary random variables, C_i, E_i and R_i to model user click, user examination and document relevance events at the i-th position:

- C_i: whether the user clicks on the result;
- E_i: whether the user examines the result;
- R_i: whether the document is relevant

where the first event is observable from search sessions and the last two events are hidden. $\Pr(C_i = 1)$ is the click-through rate of the i-th document, $\Pr(E_i = 1)$ is the probability of examining the i-th document, and $\Pr(R_i = 1)$ is the relevance of the i-th document. We use the parameter r_{π_i} to represent the *document relevance* as

$$\Pr(R_i = 1) = r_{\pi_i} \tag{1}$$

Next, we introduce the examination hypothesis mentioned in Section 1. The *examination hypothesis* was originally proposed by Richardson et al. [16] and later formalized by Craswell et al. [4]:

HYPOTHESIS 1 (EXAMINATION HYPOTHESIS). *A document is clicked if and only if it is both examined and relevant, which can be formulated as*

$$E_i = 1, R_i = 1 \Leftrightarrow C_i = 1 \tag{2}$$

where R_i and E_i are independent of each other.

Equivalently, Formula (2) can be reformulated in a probabilistic way:

$$\Pr(C_i = 1|E_i = 1, R_i = 1) = 1 \tag{3}$$

$$\Pr(C_i = 1|E_i = 0) = 0 \tag{4}$$

$$\Pr(C_i = 1|R_i = 0) = 0 \tag{5}$$

After summation over R_i, this hypothesis can be simplified as

$$\Pr(C_i = 1|E_i = 1) = r_{\pi_i} \quad (6)$$
$$\Pr(C_i = 1|E_i = 0) = 0 \quad (7)$$

Thus, as [16] explains, the document click-through rate is represented by

$$\begin{aligned}\Pr(C_i = 1) &= \sum_{e\in\{0,1\}} \Pr(E_i = e)\Pr(C_i = 1|E_i = e) \\ &= \underbrace{\Pr(E_i = 1)}_{\text{position bias}}\underbrace{\Pr(C_i = 1|E_i = 1)}_{\text{document relevance}}\end{aligned}$$

where the position bias and the document relevance are decomposed. This hypothesis has been used in most of the state-of-the-art click models to alleviate the position bias problem. Next, we will briefly review recent research on click models in which DBN and UBM are used to implement the intent hypothesis in the paper.

2.1 Models Under Cascade Hypothesis

The *cascade hypothesis* was originally proposed by Craswell et al. [4] to simulate the user search habit.

HYPOTHESIS 2 (CASCADE HYPOTHESIS). *A user examines search results from top to bottom without skips, and the first document is always examined:*

$$\Pr(E_1 = 1) = 1 \quad (8)$$
$$\Pr(E_{i+1} = 1|E_i = 0) = 0 \quad (9)$$

The *cascade model* [4] combines the examination hypothesis and the cascade hypothesis, and it further assumes that the user stops the examination after reaching the first click and abandons the search session:

$$\Pr(E_{i+1} = 1|E_i = 1, C_i) = 1 - C_i \quad (10)$$

This model is too restrictive and can only deal with the search sessions with at most one click.

The *dependent click model* (DCM) [9] generalizes the cascade model to sessions with multiple clicks and introduces a set of position-dependent parameters, i.e.,

$$\Pr(E_{i+1} = 1|E_i = 1, C_i = 1) = \lambda_i \quad (11)$$
$$\Pr(E_{i+1} = 1|E_i = 1, C_i = 0) = 1 \quad (12)$$

where λ_i represents the probability of examining the next document after a click. These parameters are globally shared across all search sessions. In this model, a user is simply assumed to examine all the subsequent documents below the position of the last click. In fact, if a user is satisfied with the last clicked document, she usually does not continue examining the following results.

The *dynamic Bayesian network model* (DBN) [3] assumes that document attractiveness determines the user click, and the user satisfaction determines whether the user examines the next document. Formally speaking,

$$\Pr(E_{i+1} = 1|E_i = 1, C_i = 1) = \gamma(1 - s_{\pi_i}) \quad (13)$$
$$\Pr(E_{i+1} = 1|E_i = 1, C_i = 0) = \gamma, \quad (14)$$

where the parameter γ is the probability the user examines the next document without clicks, and the parameter s_{π_i} is the user satisfaction. Experimental comparisons show that DBN outperforms other click models based on the cascade hypothesis. DBN employs the *expectation maximization* algorithm to estimate parameters, which may require a great number of iterations for convergence. Zhu et al. [19] introduced a Bayesian inference method, *expectation propagation* [14], for DBN.

2.2 User Browsing Model

The *user browsing model* (UBM) [6] is based on the examination hypothesis but does not follow the cascade hypothesis. Instead, it assumes that the examination probability E_i depends on the previous clicked position $l_i = \max\{j \in \{1, \cdots, i-1\} \mid C_j = 1\}$ as well as the distance between the i-th position and the l_i-th position:

$$\Pr(E_i = 1|C_{1:i-1}) = \beta_{l_i, i-l_i} \quad (15)$$

If there are no clicks before the position i, l_i is set to 0. The likelihood of a search session under UBM can be stated in a quite simple form:

$$\Pr(C_{1:M}) = \prod_{i=1}^{M} \left(r_{\pi_i}\beta_{l_i,i-l_i}\right)^{C_i} \left(1 - r_{\pi_i}\beta_{l_i,i-l_i}\right)^{1-C_i} \quad (16)$$

where there are $M(M+1)/2$ $\{\beta_{i,j}\}$ parameters shared across all search sessions. The *Bayesian browsing model* (BBM) [13] follows the same assumptions of UBM but adopts a Bayesian inference algorithm.

3. REVISITING EXAMINATION HYPOTHESIS

As we mentioned above, the examination hypothesis is the basis of most existing click models. The hypothesis is mainly aimed at modeling the position bias in the click log data. In particular, it assumes that the probability of a click's occurrence is uniquely determined by the query and the document after the document is examined by the user.

In this section, we use a controlled experiment to demonstrate that the assumptions built into by the examination hypothesis cannot completely interpret the click-through log. We show that, given a query and an examined document, there is still diversity among click-through rates on this document. This phenomenon clearly suggests that the position bias is not the only bias that affects click behaviors.

In order to perform the experiment, we collected click-through logs for one month in which each session contains the top ten returned documents in the search result page. We selected the search sessions that have at least one click on the documents at positions 2 to 10. Since it is widely believed that the user browses search results from top to bottom, thus, from the fact that at least one of the documents at the last nine positions is clicked, we can assume that the document at the first position is always examined. The search sessions are further divided into two groups with respect to the number of clicks at the last nine positions: one group includes sessions which have exactly one click at the last nine positions, while another group includes sessions which have at least two clicks at these positions. For each search query, the click-through rate is calculated on the same document and this document is at the first position. We randomly chose five queries and reported the click-through rate values on two groups of sessions in Figure 2.

According to the examination hypothesis, the relevance between a query and a document is a constant number, if

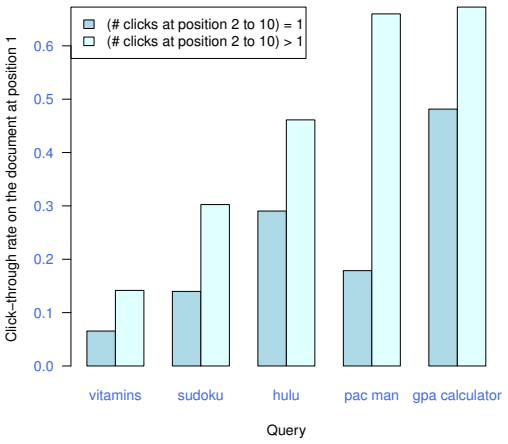


Figure 2: The document click-through rate values on two groups of search sessions over five randomly picked queries. One group includes sessions with exactly one click at positions 2 to 10, and another group includes sessions with at least two clicks at positions 2 to 10. For each query, the click-through rate is calculated on the same document and this document is always at the first position.

the document has been examined. It implies that the click-through rate in the two groups should be equivalent to each other, since the document at the top position is considered always to be examined. As shown in Figure 2, however, none of the queries presents the same click-through rate value on the two groups. Instead, it is observed that the click-through rate in the second group is significantly higher than that in the first group since the P-values of t-test on these two groups is much less than 1%.

In order to investigate the generalization of this analysis, we subtract the click-through rate in the first group from that in the second group and plot the distribution of this difference over all search queries. Figure 3 illustrates the difference of the click-through rate values on two groups for all queries. The resulting distribution matches a Gaussian distribution whose center is at a positive point about 0.2. Specifically, we found that the number of queries whose corresponding difference is located within $[-0.01, 0.01]$ occupies only 3.34% of all queries, which indicates that the examination hypothesis cannot precisely charaterize the click behaviors for most of the queries.

Since we believe that the users have not read the last nine documents when they are browsing the first document, whether the first document has a click is an independent event to the click on the last nine documents. Thus, the only possible explanation for the observed phenomenon is that there is an intrinsic search intent behind the query and that this intent leads to the click diversity in the two groups. In Section 4, we will characterize this diversity by the concept of search intent and propose the intent hypothesis for click models.

4. MODELING INTENT DIVERSITY

4.1 Intent Hypothesis

We propose a new hypothesis called the *intent hypothesis*. The intent hypothesis preserves the concept of *examination* proposed by the examination hypothesis. Moreover, our hypothesis assumes that a document is clicked only after it meets the user's search intent, i.e. it is needed by the user. Since the query partially reflects the user's search intent, it is reasonable to assume that a document is never needed if it is irrelevant to the query. On the other hand, whether a relevant document is needed is uniquely influenced by the gap between the user's intent and the query. From this definition, if we are sure that the user always submits the query which exactly reflects her search intent, then the intent hypothesis will be reduced to the examination hypothesis.

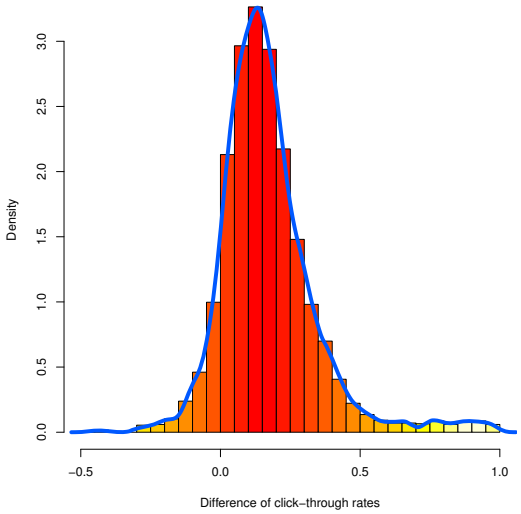
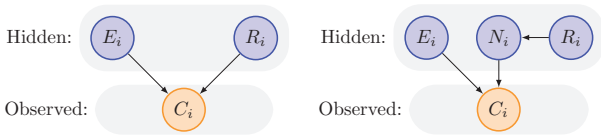


Figure 3: The distribution of the CTR difference on two group search sessions.

Formally, the intent hypothesis includes the following three statements:

1. The user will click a document if and only if it is examined and needed by the user.
2. If a document is irrelevant, the user will not need it.
3. If a document is relevant, whether it is needed is only influenced by the gap between the user's intent and the query.

Figure 4 compares the graphical models of the examination hypothesis and the intent hypothesis. We can see in the intent hypothesis a latent event N_i is inserted between R_i and C_i, in order to distinguish the occurrence of being relevant and being clicked.



(a) Examination Hypothesis (b) Intent Hypothesis

Figure 4: The graphical models of the examination hypothesis and the intent hypothesis

It order to represent the intent hypothesis in a probabilistic way, we first introduce some symbols. Suppose that there are m documents in the session s. The i-th document is denoted by d_{π_i} and whether it is clicked is denoted by C_i. C_i is a binary variable. $C_i = 1$ represents that the document is clicked and $C_i = 0$ represents that it is not clicked. Similarly, whether the document d_{π_i} is examined, whether

it is relevant and whether it is needed are respectively represented by the binary variables E_i, R_i and N_i. Under this definition, the intent hypothesis can be formulated as:

$$E_i = 1, N_i = 1 \Leftrightarrow C_i = 1 \tag{17}$$

$$\Pr(R_i = 1) = r_{\pi_i} \tag{18}$$

$$\Pr(N_i = 1|R_i = 0) = 0 \tag{19}$$

$$\Pr(N_i = 1|R_i = 1) = \mu_s \tag{20}$$

Here, r_{π_i} is the relevance of the document d_{π_i}, and μ_s is defined as the intent bias. Since the intent hypothesis assumes that μ_s should only be influenced by the intent and the query, μ_s is shared across all documents in the same session, which means that it is a global latent variable in session s. However, in different sessions, the intent bias is supposed to be different.

After we combine (17), (18), (19) and (20), it is not difficult to derive that:

$$\Pr(C_i = 1|E_i = 1) = \mu_s r_{\pi_i} \tag{21}$$

$$\Pr(C_i = 1|E_i = 0) = 0 \tag{22}$$

Compared to Equation (6) derived from the examination hypothesis, Equation (21) adds a coefficient μ_s to the original relevance r_{π_i}. Intuitively, it can be seen that we take a μ_s discount off the relevance. Especially, if the value of μ_s is fixed to 1, it means that there will be no intent bias and that our hypothesis will degenerate into the examination hypothesis.

4.2 Click Model Implementations

For all previous click models based on the examination hypothesis, the switch from the examination hypothesis to the intent hypothesis is quite simple. Actually, we only need to replace formula (6) with formula (21) without changing any other specifications. Here, the latent intent bias μ_s is local for each session s. Every session maintains its own intent bias, and the intent biases for different sessions are mutually independent.

When the intent hypothesis is adopted to construct or reconstruct a click model $\mathcal{M}$, the resulting click model is referred to as Unbiased-$\mathcal{M}$. In this paper, we choose two state-of-the-art models, DBN and UBM, to illustrate the impact of the intent hypothesis. The new models based on DBN and UBM are called Unbiased-DBN and Unbiased-UBM respectively.

4.3 Inference

4.3.1 Parameter Estimation

As specified above, when an unbiased model is constructed, we estimate the value of μ_s for each session. After all of the μ_s are known, then other parameters of the click model (such as relevance) can be learned. However, since the estimation of μ_s relies on learning the results of other parameters, the entire inference process has deadlocks. To avoid this problem, we adopt an iterative inference as shown in Algorithm 1.

Every iteration consists of two phases. In Phase A, we learn the click model parameter Θ based on the estimated values of μ_s of the last iteration. In Phase B, we estimate the value of μ_s for each session based on the parameters Θ learned in Phase A. Here, the likelihood function that we want to maximize is the conditional probability that the actual click events of this session occur under the specification of the click model, with μ_s being treated as the condition. Phase A and Phase B should be executed alternatively and iteratively until all parameters converge.

Algorithm 1 Iterative inference of unbiased model

Require: Input a set S of sessions as training data and an original click model $\mathcal{M}$ (Its own parameter set is denoted by Θ.)

1: Initialize the intent bias $\mu_s \leftarrow 1$ for each session s in S.
2: **repeat**
3: Phase A: We learn every parameter in Θ using the original inference method of $\mathcal{M}$ while we fix the values of μ_s according to the latest estimated values of μ_s.
4: Phase B: We estimate the value of μ_s for each session, using maximum-likelihood estimation, under the learning result of parameters Θ generated in phase A.
5: **until** all parameters converge

This general inference framework can be modified to be more efficient if the parameters except μ_s could be learned through online Bayesian inference. In this case, the inference is still online even after the estimations of μ_s are included. Specifically, when a session is loaded, we use the posterior distributions learned from the previous sessions to give an estimation for μ_s. We then use the estimated value of μ_s to update the distribution of other parameters. Since the distribution of every parameter changes little before and after the update, we do not need to reestimate the value of μ_s anymore, so that no iterative steps are needed. Thus, after all the parameters are updated, we just load in the next session and continue the learning process.

As described in Section 2, both UBM and DBN can employ the Bayesian Paradigm to infer the parameters. According to the method mentioned above, as a new session is loaded for training, there are three steps to execute:

1. We integrate over all the parameters except μ_s to derive the likelihood function $\Pr(C_{1:m}|\mu_s)$.
2. We maximize this likelihood function to estimate the value of μ_s.
3. Fixing the value of μ_s, we update other parameters by the Bayesian inference method.

Such online Bayesian inference facilitates the single-pass and incremental computation, which is appealing for very large-scale data processing.

4.3.2 Click Prediction

Given a test session, the joint probability distribution of click events in this session can be calculated by the following formula.

$$\Pr(C_{1:m}) = \int_0^1 \Pr(C_{1:m}|\mu_s)p(\mu_s)d(\mu_s) \tag{23}$$

In order to determine $p(\mu_s)$, we investigate the distribution of the estimated μ_s in the training process and draw a density histogram of μ_s for each query. Then we use the density histogram as an approximation to $p(\mu_s)$. In our implementation, we evenly divide the range $[0, 1]$ into 100 segments and count the density of μ_s that fall into each of the segments, and then we treat this density distribution as $p(\mu_s)$.

It is worth noting that our method is not able to predict the exact value of the intent bias for future sessions. This is because the intent bias can only be estimated when the actual user clicks are available, but in the testing data, the user click is hidden and should be unknown to the click model. Thus, we average the prediction result of future clicks over all intent bias according to the distribution of the intent bias counting from the training set. This averaging step might lose the advantage of the intent hypothesis. In an extreme case that a query never occurs in the training data, our model will set the intent bias to be 1, where the intent hypothesis degenerates to the examination hypothesis and gives the same prediction result as the original model.

4.4 An Example: User Browsing Model

We take the User Browsing Model (UBM) as an example to demonstrate how to apply the intent hypothesis to a click model. A Bayesian inference procedure to estimate the parameters is also introduced.

4.4.1 Model

Given a search session s, UBM takes the document relevances and transition probabilities as its parameters. As we mentioned in Section 2, the parameters in a single session can be denoted by $\Theta = \{r_{\pi_i}\}_{i=1}^{M} \cup \{\beta_{l_i,i-l_i}\}_{i=1}^{M}$. In addition, if we want to apply the intent hypothesis to UBM, then a new parameter should be maintained. This parameter is the intent bias for session s, which we denote by μ_s. Under the intent hypothesis, the revised version of the UBM model is formulated by (21), (22) and (15).

According to the above model specification, we derive the likelihood $\Pr(s|\Theta, \mu_s)$ for session s as :

$$\begin{aligned}
&\Pr(s|\Theta, \mu_s) \triangleq \Pr(C_{1:M}|\Theta, \mu_s) \\
=&\prod_{i=1}^{M}\sum_{k=0}^{1}\big[\Pr(C_i \mid E_i = k, \mu_s, r_{\pi_i})\cdot \qquad (24)\\
&\qquad \Pr(E_i = k \mid C_{1:i-1}, \beta_{l_i,i-l_i})\big] \\
=&\prod_{i=1}^{M}\left(\mu_s r_{\pi_i}\beta_{l_i,i-l_i}\right)^{C_i}\left(1-\mu_s r_{\pi_i}\beta_{l_i,i-l_i}\right)^{1-C_i} \qquad (25)
\end{aligned}$$

Here, C_i represents whether the document at position i is clicked. The overall likelihood for the entire dataset is the product of the likelihood for every single session.

4.4.2 Parameter Estimation

We adopt the Bayesian Paradigm to infer the parameters. The learning process is incremental: we load and process search sessions one by one, and the data for each session is discarded after it has been processed in the Bayesian inference. Given a new incoming session s, we update the distribution of each parameter $\theta \in \Theta$ based on the session data and the click model. Before the update, each θ has a prior distribution $p(\theta)$. We compute the likelihood function $P(s|\theta)$, multiply it to the prior distribution $p(\theta)$, and derive the posterior distribution $p(\theta|s)$. Finally, the distribution of θ is updated with respect to its posterior distribution.

Let's examine this updating procedure in more detail. First, we integrate the likelihood function (25) over Θ to derive a marginal likelihood function only conditioned on the intent bias:

$$\Pr(s|\mu_s) = \int_{\mathbb{R}^{|\Theta|}} p(\Theta)\Pr(s|\Theta, \mu_s)d\Theta$$

Since $\Pr(s|\mu_s)$ is a unimodal function, we can maximize it by the ternary searching procedure on the parameter μ_s, which is in the range of $[0, 1]$. The optimal value for μ_s is then denoted by μ_s^*.

With μ_s optimized, we derive the posterior distributions of each parameter $\theta \in \Theta$ via the Bayes' Rule:

$$p(\theta|s, \mu_s = \mu_s^*) \propto p(\theta)\int_{\mathbb{R}^{|\Theta'|}} \Pr(s|\Theta, \mu_s = \mu_s^*)p(\Theta')d\Theta'$$

where $\Theta' = \Theta\backslash\{\theta\}$ for short notation.

The final step is to update $p(\theta)$ according to $p(\theta|s, \mu_s = \mu_s^*)$. To make the whole inference process tractable, it is usually necessary to restrict the mathematical form of $p(\theta)$ to a specific distribution family. Here, we adopt the Probit Bayesian Inference (PBI) proposed by Zhang et al. [18] to implement the final update. PBI connects each θ with an auxiliary variable x through the probit link $\theta = \Phi(x)$, and restricts $p(x)$ always to the Gaussian family. Thus, in order to update $p(\theta)$, it is sufficient to derive $p(x|\mu_s = \mu_s^*)$ from $p(\theta|\mu_s = \mu_s^*)$ and approximate it by a Gaussian density. Then we use the approximation to update $p(x)$ and further update $p(\theta)$. For more details, please refer to [18].

Since the learning is incremental, the update procedure is executed once for each session.

5. EXPERIMENTS

In this section, we test the intent hypothesis with two state-of-the-art click models, DBN and UBM, and the original examination hypothesis in DBN and UBM are replaced by the intent hypothesis. We denote the new click models by Unbiased-UBM and Unbiased-DBN respectively. In the experiment, we firstly use the estimated relevance from click models to rank the documents and then evaluate the ranking using the human labeled relevance with respect to the normalized discounted cumulative gains (NDCG) [10]. Secondly, we use log-likelihood to evaluate how accurately the Unbiased-UBM and Unbiased-DBN predict user future clicks over UBM and DBN.

5.1 Experimental Setting

Training and testing datasets: The search sessions used to train and evaluate click models were collected from a commercial search engine in the U.S. market in the English language in January 2010. A session consists of a input query, a list of returned documents on the search result page and a list of clicked positions. We collected the session subject to the following constraints: (1) the search session is on the first result page returned by the search engine; (2) all clicks in the session are on the search result but neither on sponsored ads nor on other web elements. In order to prevent the whole dataset from becoming dominated by the extremely frequent queries, we allow each query at most 10^6 sessions. We also filter the search sessions to remove queries with low frequency less than $10^{1.5}$. For each query, we sort its search sessions according to the time stamp when the query is sent to the search engine and split them into the training and the testing sets at a ratio of 3:1. In total, we collect approximately one billion sessions over 3.6 million distinct queries. The detailed information about the dataset is summarized in Table 1.

Human judgment relevance: The manually labeled data is used as the ground truth for evaluating the relevance estimated from click models. The human relevance system

Query Frequency	# Query	# Document	# Session	#HRS Query	# HRS Ratings
$10^{1.5}$ to 10^{2}	2,503,666	56,985,022	133,499,657	2,402	428,236
10^{2} to $10^{2.5}$	782,494	24,846,850	131,894,026	2,410	443,646
$10^{2.5}$ to 10^{3}	241,528	11,411,317	128,167,508	2,132	383,308
10^{3} to $10^{3.5}$	740,53	5,184,211	124,342,019	1,869	371,900
$10^{3.5}$ to 10^{4}	21,871	2,225,700	115,626,785	1,317	285,318
10^{4} to $10^{4.5}$	6,111	873,366	101,222,539	893	223,911
$10^{4.5}$ to 10^{5}	1,688	356,813	88,668,305	471	133,469
$> 10^{5}$	616	196,746	139,407,426	265	88,236
Total	3,632,027	102,080,025	962,828,265	11,759	2,358,024

Table 1: The summary of the data set collected from one month of click logs.

Model		NDCG@1	NDCG@3	NDCG@5	NDCG@7	NDCG@10
UBM	UBM	0.578	0.577	0.587	0.606	0.596
	Unbiased-UBM	0.660	0.628	0.632	0.648	0.633
	Improvement	14.14%	8.90%	7.71%	6.94%	6.25%
DBN	DBN	0.562	0.569	0.584	0.604	0.596
	Unbiased-DBN	0.621	0.613	0.620	0.636	0.623
	Improvement	10.47%	7.74%	6.19%	5.22%	4.55%

Table 2: The experimental results on NDCG

(HRS) randomly picks a set of queries and requires editors to label the relevance between these queries and their corresponding search documents. For each query-document pair, editors give five ratings ranging from 0 to 4, corresponding to five scales: *bad*, *fair*, *good*, *excellent*, and *perfect*. HRS rating for a query-document pair is derived by averaging the ratings of this pair from several editors. On average, 200.53 documents for a query have HRS ratings. In total, HRS generated 2 million ratings for our data set. The last two columns in Table 1 show the summarized HRS rating information.

5.2 Accuracy in Relevance Estimation

One important ability of the click model is to estimate the document relevance. The trained click model is able to provide the estimated relevance for each query-document pair. We can rank all documents under a query according to the estimated relevance and compare this predicted ranking with the human judgment ranking. We expect the accuracy of the relevance estimation can be improved after eliminating the effect of the intent bias.

The normalized discounted cumulative gain (NDCG) [10] is a well-known metric for measuring the divergence between the predicted ranking and human judgments. NDCG is calculated cumulatively from the top of the result list to the bottom with the gain of each result discounted at lower ranks. Higher NDCG values correspond to a better ranking result. We report the arithmetic mean of NDCG over multiple queries. Precisely, given a ranking, the integer sequence $\{g_i\}$ denotes the editorial relevances of the documents ordered by the ranking. The NDCG at a particular rank threshold K is defined as:

$$\text{NDCG@}K = \frac{1}{Z@K} \sum_{i=1}^{K} \frac{2^{g_i} - 1}{\log(1+i)}$$

where $Z@K$ is the normalization to make the ideal ranking (i.e. the ranking obtained by ordering the documents according to their editorial relevance) to have NDCG value of 1. We report NDCG over multiple queries using the arithmetic mean. We use the relative NDCG improvement to evaluate the model with the intent hypothesis and with the examination hypothesis.

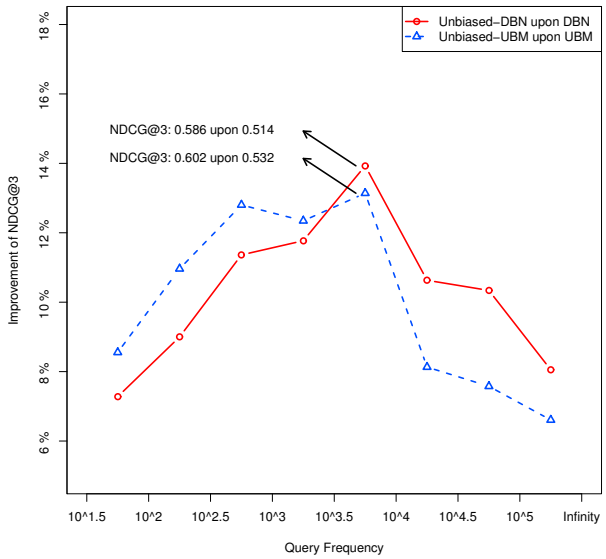


Figure 5: The relative NDCG improvement over query frequencies

We list NDCG evaluation results for the whole dataset in Table 2. It shows that NDCG@1 has been improved by 14.14% from 0.578 to 0.660 for UBM and 10.47% from 0.562 to 0.621 for DBN. NDCG@10 has been improved 6.25% for UBM and 4.55% for DBN. We clearly see that the new click models with the intent hypothesis outperform previous models with the examination hypothesis. The lower rank threshold K is, the higher NDCG@K improvement rate achieved. We perform the significance test for the NDCG improvements at all ten rank threshold K's, and find that the P-values of t-test are all less than 0.01%. Therefore, we conclude that the NDCG improvement after adopting the intent hypothesis is statistically significant.

Furthermore, we investigate at which query frequency the intent hypothesis contributes the greatest improvement. We

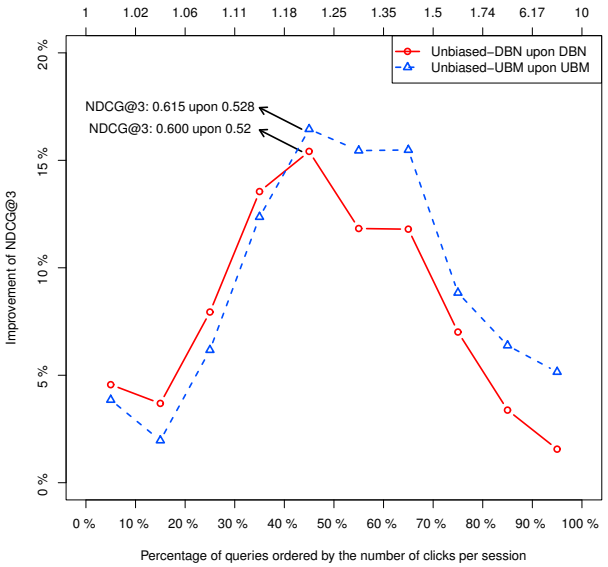


Figure 6: The relative NDCG improvement over average number of clicks

plot the relative NDCG@3 improvement across different query frequencies for UBM and DBN in Figure 5. We can see the NDCG improvement is very consistent across all query frequencies. Each of the curves approximately forms an interesting unimodal pattern. In the beginning, as the query frequency increases, the increase on NDCG improvement is mainly because we can learn the intent bias more accurately with the increase in data. After reaching the NDCG improvement peak at about 10^4, the relevance improvement becomes less, since the relevance estimation from the baseline model has become more accurate as more search sessions are used for training. As we can see, the intent hypothesis helps the baseline model with the examination hypothesis improve the relevance estimation for most queries, especially queries whose frequencies are between 10^3 to 10^4.

From the analysis in Section 3, we can see that the number of clicks within a search session is related to the intent bias, and it is interesting to see the relevance improvement on queries with different numbers of clicks. Thus, for each query, we calculate the average number of clicks over all sessions. We split the query set into 10 equal subsets according to the increasing order of the average number of clicks. We plot the curves of the improvement rates of NDCG@3 for these 10 subsets in Figure 6. The ten corresponding quantiles of the average number of clicks are aligned along the x-axis above the box. From the figure, we clearly see that the curves also have a unimodal form. As we know, a query with lower average number of clicks tends to be navigational, while a query with higher average number of clicks tends to be informational. In the beginning, along with the increase of the average number of clicks, the intent hypothesis leads to more significant improvement. Thus, the intent hypothesis makes the click model to more accurately characterize informational queries, which usually have the diversity of search intents. If the average number of clicks become large enough, the user's intrinsic intents will become too ambiguous to be characterized by the intent bias. Thus, after a peak point where there is already quite large average number of clicks, the improvement will start to drop with the increase in the average number of clicks.

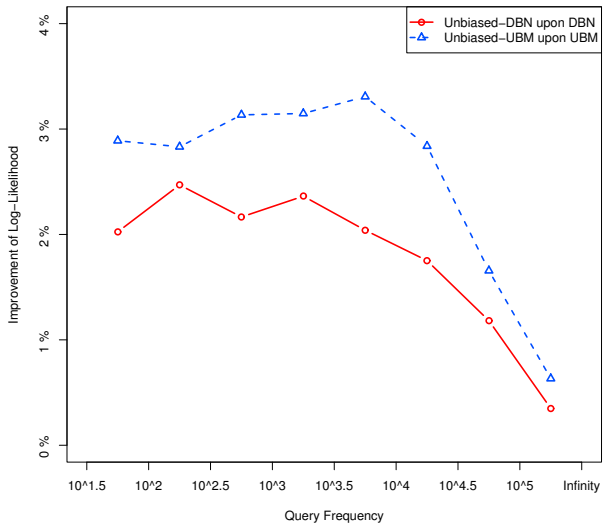


Figure 7: The Log-likelihood improvement over query frequencies

5.3 Accuracy in Click Prediction

After training the model, the parameters in click model have been estimated and we can use it to predict the joint probabilities $\Pr(C_{1:m})$ for all click configurations. We evaluate the click prediction by log-likelihood (LL), which has been widely used to measure the fitness in click models such as UBM [6] and CCM [8]. Its value indicates the logarithm of the joint probability of user click events in testing datasets predicted by the trained click model. A larger LL indicates better prediction accuracy, and the optimal value is 0. The improvement of LL value ℓ_1 over ℓ_2 is computed as $(\exp(\ell_1 - \ell_2) - 1) \times 100\%$. We report average LL over multiple sessions using arithmetic mean.

Figure 7 demonstrates the relative log-likelihood improvements on Unbiased-UBM and Unbiased-DBN over UBM and DBN on different query frequencies. For frequent queries, the baseline model can also accurately predict clicks. So the improvement curves drop along query frequencies. The overall improvements in log-likelihood are 2.10% for DBN and 2.96% for UBM. We can see that the performance of our new models on log-likelihood is close to the baseline model. As introduced in Section 4.3.2, the prediction results are the average of cases over all intent biases according to the distribution of intent bias computed from the training set.

6. DISCUSSION

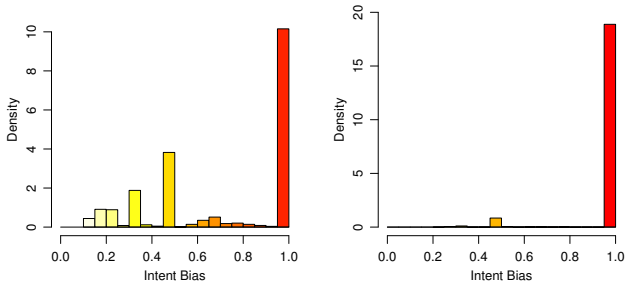
In this section, we report several interesting findings associated with the concept of intent bias. Our study suggests that the calculation of intent bias is not only helpful in estimating the unbiased document relevance but also allows us to investigate more deeply some other web search mechanisms.

6.1 Intent Bias Distribution

We discovered that the sessions of informational queries and navigational queries have significantly different intent bias distributions. This property allows us to design an automatic method for classifying queries into two classes — informational and navigational.

Let us start with two examples. In Figure 8(a), we report the density distribution of intent bias for the query

"photosynthesis" in a density histogram. In this histogram, the density distribution is multi-modal, which means that the search pattern of all users can be clustered into several groups, each with an intent bias. This observation coincides with our intuition that in an informational query such as "photosynthesis", it is hard to reflect the exact search intent of every user. However, since a density near 1 is larger than the density at other values, we can conclude that the majority of users tends to click on the documents which are relevant. In Figure 8(b), we report the density distribution of intent bias for the query "paypal", which is a typical navigational query. Different from the informational query, in this case the sessions with an intent bias near 1 dominate all the sessions. This result is also intuitive, because we believe that a navigational query is much more likely to precisely reflect the user's intent than an informational query.



(a) query "photosynthesis" (b) query "paypal"

Figure 8: Density histogram of intent bias on the two example queries

In order to numerically characterize the difference between the distributions of Figure 8(a) and Figure 8(b), we calculate the *entropy* of the intent bias distribution for each query. The entropy measures the degree of the diversity of the intent biases behind a single query: a higher entropy value suggests that the distribution of intent biases is more diversified, while a lower entropy value suggests that the distribution is more concentrated. Obviously, the entropy for the query "photosynthesis" (2.4611) is higher than that for "paypal" (0.1452).

To give a more convincing conclusion, we manually chose 200 informational queries and 200 navigational queries, and plotted the distribution of entropies on these two classes of queries in Figure 9. It is observed that the entropy for navigational queries is mostly located near zero, while the entropy for informational queries are distributed around a positive value far away from zero. This observation exactly coincides with our intuition that there is a larger degree of diversity in intent biases for informational queries.

The above analysis suggests that search engines can maintain a query classifier based on the entropy of the intent bias. With such a classifier, the search engine can treat queries with a single intent and with multiple intents differently so as to satisfy the user's information need behind the query.

6.2 Relationship between Intent Bias and Other Click Patterns

According to the intent hypothesis, the intent bias of a search session is deterministically derived from the click events. Thus, there should be some specific connections between the intent bias and other click patterns, such as the click positions and the number of clicks. In Figure 10, we report the

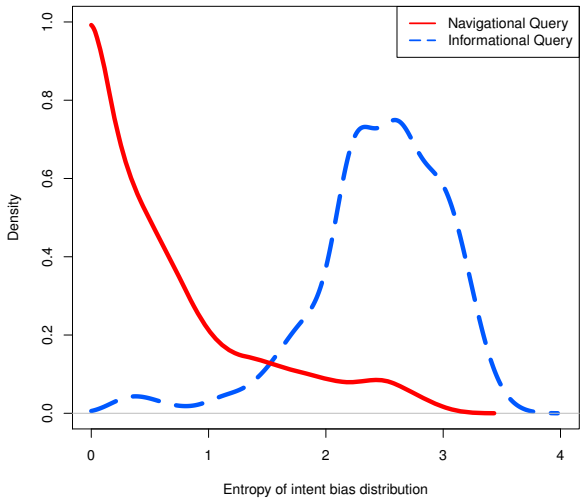


Figure 9: The distribution of the entropy of the intent bias over the two groups of queries, i.e., the informational queries and the navigational queries.

relation between intent bias and three statistical quantities, including the first-click position, the last-click position and the number of clicks in the session. These quantities are averaged among all the search sessions based on three months of data. In order to avoid the misapprehension, it is necessary to note that the higher value of μ, which is listed along the x-axis, means the lower intent bias. For example, the rightmost endpoint of x-axis corresponds the case that there is no intent bias.

As illustrated in Figure 10, the first-click position decreases as the value of μ increases. This phenomenon is natural, since the lower the intent bias becomes, the more similarity there is between the user's search intent and the query she issues. Since the search engine arranges the position of documents with respect to their similarities to the query, a higher positioned document is more likely to have a close connection to the query, and a higher probability of being clicked by the user.

On the other hand, Figure 10 shows that the last-click position decreases in the range of $[0, 0.4) \cup (0.85, 1]$ and increases in the range of $[0.4, 0.85]$. The reason for its increase may be due to another characteristic of the search intent: with a high intent bias, the user tends to continue browsing the search result page and click more documents. However, why does the last-click position dramatically decrease in the range of $(0.85, 1]$? This phenomenon is caused by the existence of navigational queries. In fact, users who submit navigational queries usually have their intent biases close to zero, or value of μ close to 1. Most of them would be satisfied by the top document and leave. Thus, the last-click position in sessions with a low intent bias is often equal to the first-click position.

The curve of the average number of clicks can also be interpreted in a similar way. For intent biases in the range of $[0, 0.85)$, most of the queries are informational, so the users click more documents if they have higher intent biases. However, for intent biases in the range of $[0.85, 1]$, the navigational queries occupy a large proportion, which makes the average number of clicks close to 1.

The interesting connection between the intent bias and other click features makes it a valuable attribute for the

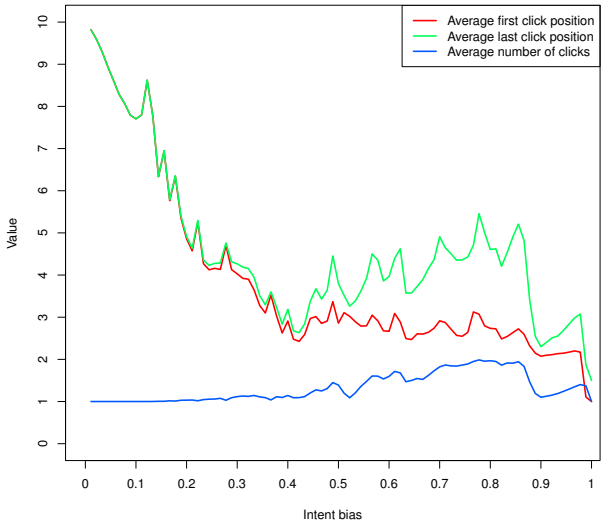


Figure 10: The relation between the intent bias and three click patterns such as the average first click position, the average last click position and the average number of clicks.

training of ranking functions. In addition, it can also be used for the recognition and classification of click patterns.

7. CONCLUSION

In this paper, we have investigated the relationship between intent, document and query and make a deep exploration of the gap between click-through rate and document relevance. We have found that the widely adopted concepts of the position bias and the examination hypothesis fail to completely explain the actual bias between click-through rate and relevance because of the gap between user search intent and input query. In order to characterize the diversity of search intent, we propose the intent hypothesis as a complement to the original examination hypothesis. The new hypothesis is very general and can be fit into most of the existing click models to improve their capacities for learning unbiased relevance. Under the concept of intent bias introduced in this paper, we have successfully modeled the actual bias between click-through rate and relevance, whose rationality has been verified both theoretically and empirically. Furthermore, we have demonstrated how to infer the click models with the consideration of the intent hypothesis. The experiments on large scale click through log show that the models with the intent hypothesis consistently and significantly perform better than the original versions of click model under the examination hypothesis.

Besides user clicks, other useful information can be derived from in click through logs, such as the user's history of input queries and visited pages. This kind of information is related to the user's current search intent and can be used to better identify the search intent behind the query. Our next step is to include such information into the click model with the intent hypothesis to further improve click model accuracy.

Acknowledgment

Botao Hu would like to thank Prof. Roger Olesen of Tsinghua University for polishing the language. Botao Hu was supported in part by the National Basic Research Program of China Grant 2007CB807900, 2007CB807901, the National Natural Science Foundation of China Grant 60604033, 60553001, 61073174, 61033001 and the Hi-Tech research and Development Program of China Grant 2006AA10Z216.

8. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. SIGIR '06, pages 19–26.

[2] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36:3–10, September.

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

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

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

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

[7] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. SIGIR '04, pages 478–479.

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

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

[10] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20:422–446, October.

[11] T. Joachims. Optimizing search engines using clickthrough data. KDD '02, pages 133–142.

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

[13] C. Liu, F. Guo, and C. Faloutsos. Bbm: bayesian browsing model from petabyte-scale data. KDD '09, pages 537–546.

[14] T. P. Minka. Expectation propagation for approximate bayesian inference. UAI '01, pages 362–369.

[15] B. Piwowarski, G. Dupret, and R. Jones. Mining user web search activity with layered bayesian networks or how to capture a click in its context. WSDM '2009, pages 162–171.

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

[17] R. Srikant, S. Basu, N. Wang, and D. Pregibon. User browsing models: Relevance versus examination. KDD '10, pages 223–232.

[18] Y. Zhang, D. Wang, G. Wang, W. Chen, Z. Zhang, B. Hu, and L. Zhang. Learning click models via probit bayesian inference. CIKM '10, pages 439–448.

[19] Z. A. Zhu, W. Chen, T. Minka, C. Zhu, and Z. Chen. A novel click model and its applications to online advertising. WSDM '10, pages 321–330.