Relational Click Prediction for Sponsored Search

Chenyan Xiong[1,2,*]
[2]Graduate University
Chinese Academy of Sciences
Beijing, P.R.China
xiongcy@ios.ac.cn

Taifeng Wang
Microsoft Research Asia
Beijing, P.R.China
taifengw@microsoft.com

Wenkui Ding*
Tsinghua University
Beijing, P.R.China
dingwenkui@gmail.com

Yidong Shen[1]
[1]Institute of Software
Chinese Academy of Sciences
Beijing, P.R.China
ydshen@ios.ac.cn

Tie-Yan Liu
Microsoft Research Asia
Beijing, P.R.China
tyliu@microsoft.com

ABSTRACT

This paper is concerned with the prediction of clicking an ad in sponsored search. The accurate prediction of user's click on an ad plays an important role in sponsored search, because it is widely used in both ranking and pricing of the ads. Previous work on click prediction usually takes a single ad as input, and ignores its relationship to the other ads shown in the same page. This independence assumption here, however, might not be valid in the real scenario. In this paper, we first perform an analysis on this issue by looking at the click-through rates (CTR) of the same ad, in the same position and for the same query, but surrounded by different ads. We found that in most cases the CTR varies largely, which suggests that the relationship between ads is really an important factor in predicting click probability. Furthermore, our investigation shows that the more similar the surrounding ads are to an ad, the lower the CTR of the ad is. Based on this observation, we design a continuous conditional random fields (CRF) based model for click prediction, which considers both the features of an ad and its similarity to the surrounding ads. We show that the model can be effectively learned using maximum likelihood estimation, and can also be efficiently inferred due to its closed form solution. Our experimental results on the click-through log from a commercial search engine show that the proposed model can predict clicks more accurately than previous independent models. To our best knowledge this is the first work that predicts ad clicks by considering the relationship between ads.

*This work was done when the first and third authors were visiting Microsoft Research Asia.

WSDM'12, February 8–12, 2012, Seattle, Washington, USA.

Categories and Subject Descriptors

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

General Terms

Economics, Algorithms, Experimentation

Keywords

Online Advertising, Sponsored Search, Relational Click Prediction, Continuous CRF

1. INTRODUCTION

Online advertising has been a major business model for today's commercial search engines. When search engines deliver organic search results to a user, sometimes they also show sponsored search results, i.e., advertisements (ads). These ads are usually ranked according to two kinds of information[14]: the bid price on the query keywords given by the advertiser and the probability for the user to click on the ad predicted by the search engine. As a dominant industry practice, only when a user clicks on an ad, will the search engine charge the corresponding advertiser. This is referred to as cost per click (CPC). Generalized second price auction (GSP)[10] is a widely used pricing model for CPC, in which the price that an advertiser has to pay depends on the predicted click probability of his/her own ad as well as the bid price and predicted click probability of the ad ranked in the next position.

From the above introduction, we can clearly see that the accurate prediction of click probability is very important to sponsored search, since it impacts both ranking and pricing of the ads. In the literature, there have been several pieces of work on predicting the ad click probability. In [8] and [15], the click probability of an ad is predicted based on its historical click-through data. In real systems, however, there are many ads without (sufficient) historical click-through data and therefore the above work cannot be directly applied to these ads. To tackle the challenge, in [16][19], two kinds of new features are introduced: *semantic features* (e.g., the relevance of an ad to the query and the quality of the ad) and *aggregated click features* (e.g., aggregated at the advertiser or query levels). These new features can handle many ads

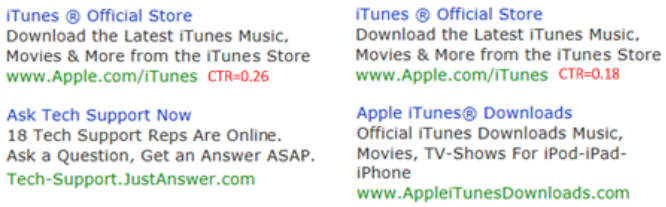


Figure 1: Two ad lists for query 'iTunes account'.

without historical data and can increase the data density of click prediction. To further improve the accuracy of click prediction, in some other work [4], the user demographic information is also used to personalize the prediction result.

Please note that all the aforementioned works predict the clicks on ads in an independent manner. That is, the click probability is computed only based on the information of a given ad, without considering the other ads shown together with it. However, our data study shows that it is inappropriate to ignore the influences of the other ads in real search scenarios. Fig. 1 shows two different ad lists from a commercial search engine triggered by the same query 'iTunes account'. We can see that ad 'iTunes Official Store' is shown in the same position of both lists. By analyzing the click-through logs, we find that its click-through rate (CTR) in the first list is 26%, while its CTR is just 18% in the second list. Actually this is not a rare phenomenon. Statistically, the CTR of the same ad, in the same position, and for the same query, can vary largely when surrounded by different ads (see Section 2.2).

We believe that the aforementioned observations will have profound impact on the accuracy of click prediction. However, it has not been well investigated in the literature. First, there are some previous studies [7, 11, 12, 21] on the mutual influence between ads. However, they are focused on mining user click behaviors or designing auction mechanisms, rather than predicting click probability. Second, there are not many attempts on understanding what kinds of factors account for the observed mutual influence between ads. Third, the experimental setting of some previous work was not specifically designed for click prediction, making their findings on the factors behind ad mutual influence not directly applicable to our scenario.

For example, in [21], the authors studied the user click behaviors on sponsored listings shown in the search results page. As part of the study, they claimed that when multiple ads are shown together, high-quality ads will influence low-quality ads in attracting user's attention (and therefore clicks). In their experiments to verify this claim, they used CTR as a surrogate of the ad quality. Given a CTR value at the first position (denoted by CTR-1), they averaged the CTR of the ads in the second position from different queries to get CTR-2, and showed that there is a strong dependency between CTR-1 and CTR-2. Please note that this dependency is observed when the query information is averaged out (and therefore unknown). However, in the context of click prediction, the query is always given and fixed. In this new setting, the conditional dependency between CTR-1 and CTR-2 may not exist any longer. Actually our experiments in Section 2 verified that the conditional dependency really does not exist. In other words, the CTR of an ad is not clearly influenced by the qualities of its surrounding ads when the query is given.

Considering the limitations of previous work, in this paper, we perform a new study on click data, with an experimental setting that targets the click prediction problems. Our study shows that the similarity between ads can explain the mutual influence much better than ad quality. Our statistics show that when the query is given, the CTR of an ad is (very well) negatively correlated with the similarity between it and the surrounding ads. Actually, this phenomenon can be intuitively explained. When the surrounding ads are similar to the given ad in their contents (or topics), it is very likely that they will distract user's attention since all these ads offer similar products or services. However, when the ads are dissimilar, they may become complementary to each other and will not distract user's attention by much. [1]

Based on our data study, we make the first proposal in the literature to predict ad click probability by considering the mutual similarity between ads. Specifically, we assume that the CTR of an ad is determined by two kinds of information. First, it is largely determined by the intrinsic properties of the ad itself, e.g., the relevance to the query, the static quality, and the historical CTR. Second, it is also affected by the similarity between this ad and the surrounding ads. We propose using a Conditional Random Fields (CRF)[20] model that can naturally leverage both kinds of information. Specifically, the feature functions on the vertices of the CRF capture the intrinsic properties of the ad, and the feature functions on the edges capture the similarity between ads. We show that this model can be effectively learned with maximum likelihood estimation and can be efficiently inferred due to its closed-form solution. We have conducted experiments using the ad click-through logs from a commercial search engine. The experimental results have shown that our proposed model can produce consistently better prediction accuracy on ad clicks than baseline models. This verifies the usefulness of leveraging the relationship between ads in click prediction.

To sum up, the contributions of our work include:

- We have performed a solid data analysis that certificates the existence of mutual influence between ads in the context of click prediction. In addition, we have found that the similarity between ads is a factor that can better explain the mutual influence than the quality of ads.

- We have proposed a CRF model to predict ad click probability by considering both the intrinsic properties of an ad and its similarity to other ads in same sponsored list. To the best of our knowledge, this is the first model that considers the relationship between ads in click prediction.

[1]Our finding is related to the diversity in organic search results [2, 9, 18, 22]. It has been shown that more diverse search results will have a higher probability of satisfying user's information needs, while similar search results will distract each other. There was some work[22] in the search space that diversifies search results by considering the similarity between web pages, however, as far as we know, there is no work in the online advertising area that considers the mutual similarity between ads when predicting their click probabilities.

The remaining parts of this paper are organized as follows. In Section 2, we present our data analysis. Before introducing our click prediction model in Section 4, we review the click prediction task and propose a new setting of click prediction called *relational click prediction* in Section 3. In Section 5, we present our experiment settings and compare our results with other baseline models. At last, we conclude the paper and discuss the future work in Section 6.

2. DATA ANALYSIS

As mentioned in the introduction, the focus of this work is to leverage the relationship between ads to improve click prediction. In this section, we will present some related data analysis. With the analysis, we would like to answer the following questions: 1) Is there mutual influence between the ads shown together that affects their CTR's, and furthermore in what manner and to what degree? 2) If the answer to the first question is yes, what kind of relationship will be the key factor behind this mutual influence?

In the remainder of this section, we will first introduce the setting of our study, and then introduce our findings with respect to the two questions.

2.1 Setting of the Study

We used the ad click-through log of one month from a commercial search engine in our study. The log contains all the ad information for each submitted query, including ad copy, displayed position, and click information. In our study, we focus on the mainline ads, which are shown above the organic search results, because the mainline ads contribute to the majority of search engines' ad clicks and revenue. We represent the data in terms of ⟨query, ad list⟩, where the ad list corresponds to the ordered group of ads shown for the query. We remove all those ⟨query, ad list⟩ with fewer than 10 occurrences in the data to make our study more reliable. After that, our data set contains over two millions of unique ⟨query, ad list⟩, about 700K unique queries, 600K unique ads, and over two hundred million impressions in total.

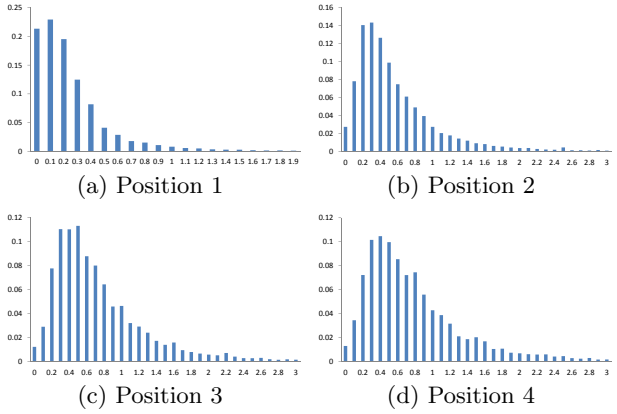
In order to study the mutual influence of ads on their CTR's, i.e., how the CTR of an ad is affected by its surrounding ads, we need to remove the influences of other factors, such as the different properties of ad and query. To this end, we put an ad in a specific context, i.e., a triple $T = \langle q, a, p\rangle$, where q, a, and p represent query, ad, and position respectively. In our data analysis, we select the triples T that appear in multiple ⟨query, ad list⟩ (for ease of reference, we denote these ad lists as $L = \{l_1, l_2, \cdots .l_n\}$). If we use $IMP = \{imp_1, \cdots, imp_n\}$ to denote the number of impressions, and $CLK = \{clk_1, \cdots, clk_n\}$ the number of clicks of T in these lists, we can compute the CTR of T in l_i as follows:

$$CTR_{T,l_i} = \frac{clk_i}{imp_i}.$$

And the average CTR of T will be

$$CTR_T = \frac{\sum_i imp_i \times CTR_{T,l_i}}{\sum_i imp_i}.$$

By comparing CTR_{T,l_i} for the same T but in different l_i, we will be able to examine the influence of other ads in l_i on the CTR of T.



(a) Position 1 (b) Position 2

(c) Position 3 (d) Position 4

Figure 2: NSDev of $\langle q, a, p\rangle$ triple at different Positions.

2.2 Mutual Influence on CTR

Previous work on click prediction assumes that the CTR of T is independent of other ads shown together with it. That is, the CTR of T will not change (or not change by much) in different ad lists. In order to verify whether this assumption holds, we check the normalized standard deviation (NSDev) of T's CTR in different ad lists:

$$NSDev_T = \frac{SDev\{CTR_{T,l_1}, \cdots, CTR_{T,l_n}\}}{CTR_T},$$

where $SDev\{CTR_{T,l_1}, \cdots, CTR_{T,l_n}\}$ is the standard deviation of $\{CTR_{T,l_1}, \cdots, CTR_{T,l_n}\}$.

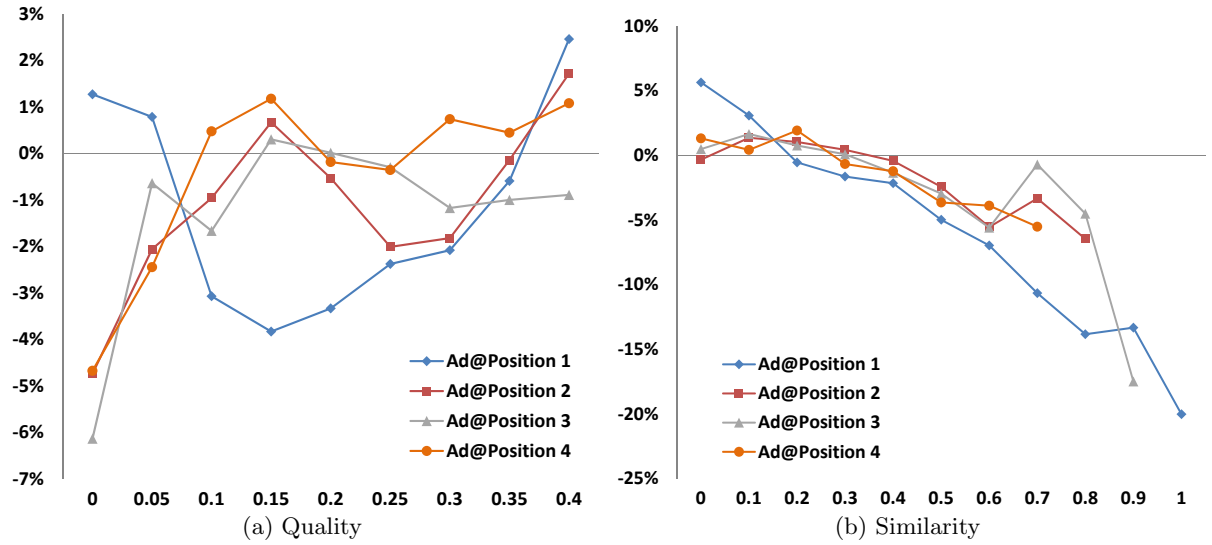
Fig. 2 shows the $NSDev_T$ with respect to different positions of the ads. The horizontal axis corresponds to the $NSDev_T$ value, and the vertical axis corresponds to the percentage of triples having this value. From the figure, we can see that the CTR of an ad varies largely among different lists, indicating that the surrounding ads will heavily influence the CTR of a given ad. Therefore, if we estimate the CTR of an ad without considering its surrounding ads, the best we can get will be the average CTR over different lists. It is clear that for most cases, the real CTR will be very different from this estimation given the large variance. This motivates us to perform deep investigation on the relationship between ads, and take the relationship into consideration when predicting click probabilities. This is exactly the focus of our paper.

2.3 Factors of Mutual Influence

In this subsection, we study two kinds of factors that may potentially account for the mutual influence: quality of ads and similarity between ads.

2.3.1 Ad Quality

Inspired by [21], we first study whether the CTR of an ad will decrease if the other ads shown together with it are of high quality. We also use the historical CTR of an ad as the surrogate of its quality. Given a triple T and an ad list l_i containing T, we test whether the CTR of T in l_i (calibrated by its average CTR on all the ad lists in order to better reflect the influence of other ads) correlates with the CTR's of the others ads in l_i. Here we use ΔCTR to



(a) Quality (b) Similarity

Figure 3: ΔCTR with ad quality and similarity. Y axis is ΔCTR. X axis is surrounding ads' CTR in (a) and similarity with surrounding ads in (b).

denote the calibrated CTR:

$$\Delta CTR_{l_i} = \frac{CTR_{T,l_i} - CTR_T}{CTR_T}.$$

Then we plot the ΔCTR_{l_i} for all the triples T with respect to different CTR values of their surrounding ads in Fig. 3.a. Surprisingly, as the CTR (quality) of other ads increases, the ΔCTR of the target triple T does not decrease as expected. That is, given the query, the CTR of an ad is not clearly influenced by the quality of other ads shown together with it.

At first glance, our finding seems to be contradictory to the results given in [21]. In our opinion, however, they are not opposite to each other, and the difference mainly comes from different experimental settings.

In [21], for each <query, ad list>, the CTR's of the ads ranked in position 1 and position 2 are computed. Then all the ad lists whose first ads have similar CTR (denoted as CTR-1) are put together, no matter they are from the same query or not, and the average value of the CTR's of their second ads are computed (denoted as CTR-2). They observed that when CTR-1 increases, CTR-2 drops accordingly. Actually, we have rerun their experiments on our data, and have obtained very similar results (see Fig. 4). However, please note that according to the experimental setting in [21], the dependency between CTR-1 and CTR-2 are observed when the query information is averaged out (and thus unknown). Therefore, it is unclear whether the dependency comes from CTR-1 and CTR-2 themselves, or from the unknown query. Further study is needed to make it clear. By carefully looking at the data, we find that it is the navigational queries that account for the major part of the above dependency. Specifically, a lot of ads with 20% or even higher CTR were shown for navigational queries. Usually, for navigational queries, user will only click on one ad whose display URL is from the web site created for the query. As a result, the CTR's of all the other ads will be low. In such case, CTR-2 is low not because CTR-1 is high. The true reason lies in the navigational nature of the query.

In the setting of click prediction, the mutual influence between ads is always studied when the query is given and fixed. In this setting, the conditional dependency between CTR-1 and CTR-2 may not exist any longer. This is easy to understand according to the basic principles of probability theory. That is, when two random variables are dependent, they can become conditionally independent when a third random variable is observed. In our experimental setting, we consider the query as given, and therefore our finding can be used to verify whether the CTR's of the first ad and the second ad are conditionally independent. According to our results, they seen to be independent of each other (at least, there is no clearly dependency observed).

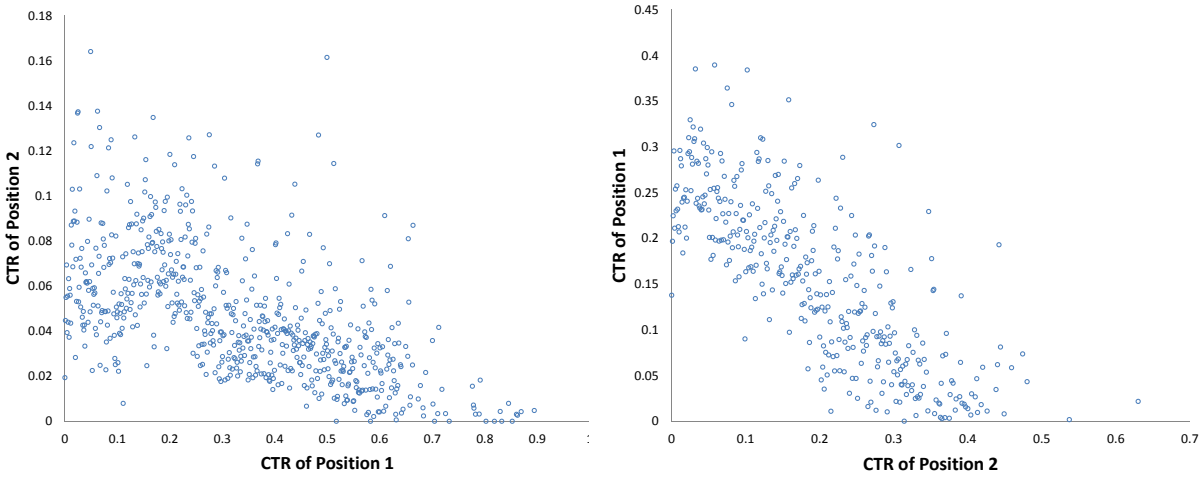
2.3.2 *Similarity between Ads*

While the mutual influence between ads is not well evidenced by ad quality, we have found another factor, i.e., the similarity between ads, which seems to be more relevant to the mutual influence between ads. Our finding is based on the investigation on those triples whose CTR variances in different ad lists are relatively large among all the triples. We find that when a triple T is shown in a list l_i, its CTR_{T,l_i} will be relatively lower than its average CTR_T when the other ads in l_i have similar contents to it. Please refer to several such examples in Fig. 5.

To see whether the observations from examples are statistically reliable, we have conducted the following experiments. For simplicity and without loss of generality, we use term overlap in the title and description of two ads as the similarity measure. Denote $S(str_1, str_2)$ as the term overlap of two strings str_1, str_2, whose definition is as below:

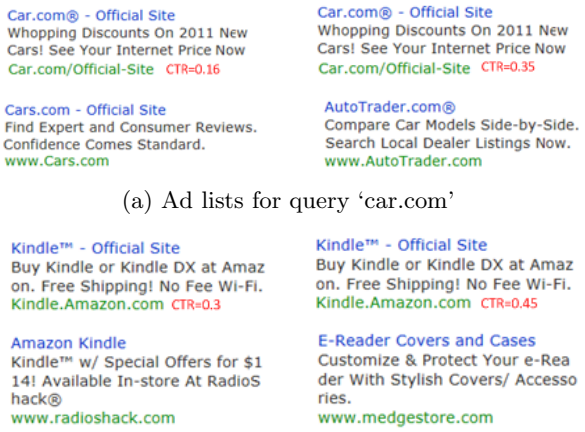
$$S(str_1, str_2) = \frac{2\sum_{t_i \in str_1, t_j \in str_2} I(t_i, t_j)}{|str_1| + |str_2|},$$

Where t_i, t_j are terms in str_1 and str_2, $I(\dot{})$ is an indicator function whose value is one only when $t_i = t_j$, and $|str_1|, |str_2|$ are the lengths of the two strings. Furthermore,



(a) Position 2 via position 1 (b) Position 1 via position 2

Figure 4: CTR pairs: The left one is CTR of ad in Position 2 bucketed by Position 1; the right one is CTR of Position 1 bucketed by Position 2.



(a) Ad lists for query 'car.com'

(b) Ad lists for query 'kindle'

Figure 5: Examples for ads' different CTR's with different surrounding ads.

we define the similarity between two ads a_1, a_2 as follows,

$$S(a_1, a_2) = \frac{1}{2}(S(str_{t1}, str_{t2}) + S(str_{d1}, str_{d2}))$$

where str_{ti}, str_{di} are the strings of a_i's title and description.

For a triple T, we get all the ⟨query, ad list⟩ from our data set, denoted as $L = \{l_1, l_2, \cdots, l_n\}$. We then compute the average similarity S_i between the ad in T and the other ads in l_i. After that, we check whether S_i has some correlation with $\Delta CTR_{T,l_i}$, and the result is plotted in Fig. 3.(b).

From the figure, we can clearly see that $\Delta CTR_{T,l_i}$ is negatively correlated [2] with S_i. This indicates that for a particular triple T, when it is shown in an ads list where the other ads are more similar to it, the CTR of T will become lower.

In our opinion, the above finding can be explained in the following intuitive manner. When the surrounding ads are similar to the given ad in their contents (or topics), it is very likely that they will distract user's attention since all these ads offer similar products or services. However, when the ads are dissimilar, they may become complementary to each other and will not distract user's attention by much.

Motivated by our findings, we propose a new click prediction model which takes the similarity between ads into consideration. Details will be given in the next two sections.

3. RELATIONAL CLICK PREDICTION

In this section, we first review the sponsored search system and the click prediction task. Then we propose a new setting of click prediction, which we call *relational click prediction*.

3.1 Overview of Sponsored Search System

Given a query and a user, the sponsored search system tries to find an optimal subset of ads that achieves the best expected revenue from its inventory of ads. Roughly speaking, this process can be divided into three steps:

1. Ad selection: select a set of candidate ads for a search, based on the bid keywords and the match type (e.g., exact match and broad match).

2. Ad ranking: rank the candidate ads based on their expected revenue[14]: $E(revenue) = pClick \times b$, where $pClick$ is the output of a click prediction model and b is the bid price. According to the ranking results and the expected revenue, the system also determines where to show these ads (e.g., in the mainline or sidebar of the search result page).

3. Ad pricing: charge the corresponding advertiser a certain amount of money if a user clicks on his/her ad. Usually the general second price auction (GSP) is used to determine the price. Specially, the price for ad a_i is calculated as follows: $\frac{pClick_{i+1} \times b_{i+1}}{pClick_i}$, where a_{i+1} is the

[2]The Pearson product-moment correlation coefficient r of S_i and $\Delta CTR_{T,l_i}$ at position 1 to 4 are $-0.99, -0.74, -0.73, -0.64$ ($r < -0.5$ usually means a strong correlation.)

ad displayed in the next position to a_i, and $pClick_{i+1}$, b_{i+1} are the $pClick$ and bid price of a_{i+1} respectively.

We can see that the click prediction plays a fundamental role in the above process: it determines both ranking and pricing of the ads. Previous works on click prediction [1, 4, 5, 19] usually treat each ad in an independent manner, without considering the influence of other ads shown together with the ad. However, based on the analysis in the previous section, the CTR of an ad is highly dependent on its relationship with other ads. Therefore, it would be meaningful to consider all the ads showed in the same ad list together and reformulate the click prediction problem in a relational manner.

3.2 Relational Click Prediction

Let u denote a user and q denote a query submitted by the user. Let $\{a_1, \cdots, a_n\}$ be the set of ads in the ad list l, c_i a binary variable indicating whether a click on ad a_i happens, and p_i the position of ad a_i.

Conventional click prediction models aim to predict the probability of $c_i = 1$ given user u, query q and position p_i for each individual ad a_i, in an independent manner:

$$p(c_i = 1|u, q, p_i, a_i) = f(x_i),$$

where x_i is the feature vector extracted from $\langle u, q, p, a_i \rangle$, and f is the predictive model to be learned.

In this paper, we treat the ads in a list no longer independent with each other and propose a new setting of click prediction, referred to as *relational click prediction*. Instead of predicting each $p(c_i = 1|u, q, p_i, a_i)$ individually, relational click prediction leverages the relationship between ads and predicts the CTR's of all the ads in an ad list simultaneously using a unified model:

$$(p(c = 1|u, q, p_1, a_1, l), \cdots, p(c = 1|u, q, p_n, a_n, l))^T = F(X, R)$$

where $X = \{x_1, x_2, \cdots, x_n\}$ includes all the feature vectors x_i extracted from $\langle u, q, p_i, a_i \rangle$ and R is the relationship between ads, and F is a model whose output is a n-dimension vector (the i-th dimension corresponds to the CTR of a_i). We refer to F as the relational click prediction model, and f the local model. Also, we call X the local features and R the relational information.

Specially in this paper, we choose Conditional Random Fields (CRF) as the relational click prediction model $F(X, R)$. CRF is widely used to model relationships between variables. It can capture the local features using its vertexes and the relational information using its edges. In the next section, we will present the details of the CRF model as well as its learning and inference processes.

4. CRF FOR RELATIONAL CLICK PREDICTION

In this section, we introduce our CRF model for relational click prediction. Since we require the model to output a click probability, which is a continuous value, we choose to use the continuous CRF model proposed in [17]. Fig. 6 shows the graph representation of the continuous CRF model.

4.1 Continuous CRF Model

Let $X = \{x_1, x_2, \cdots, x_n\}$ denote the input feature vectors of the ads displayed for a query, and $Y = \{y_1, y_2, \cdots, y_n\}$ denote the log mods[3] of the CTR's for the ads. The probability distribution of the output Y conditioned on the input X is defined as

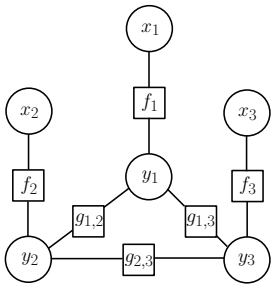


Figure 6: A continuous CRF model for relational click prediction.

$$P(Y|X) = \frac{1}{Z(X)} \exp\left\{ \sum_i h(y_i, X; w) + \sum_{j>i} \beta g(y_i, y_j, X) \right\}, \quad (1)$$

where h is the vertex feature function representing the dependence of the CTR on the input vector of ads, g is the edge feature function representing pairwise relationship between ads, and $Z(X)$ is a normalization factor (also known as the partition function).

$$Z(X) = \int \left\{ \sum_i h(y_i, X; w) + \sum_{j>i} \beta g(y_i, y_j, X) \right\} \mathrm{d}Y.$$

For simplicity, we define the vertex feature function as follows,

$$h(y_i, X; w) = -(y_i - f(x_i; w))^2 \quad (2)$$

where $f(x_i; w)$ can be any conventional click prediction model which only depends on the feature vector of each individual ad.

According to our findings in Section 2, we use similarity between ads as the relationship. The intuition is that if two ads are very similar to each other, their click probabilities will both become lower. To encode this intuition, we define the edge feature function as below.

$$g(y_i, y_j, X) = -s_{i,j}(y_i + y_j), \quad (3)$$

where $s_{i,j}$ is the term similarity between ads i and j: $s_{i,j} = S(a_i, a_j)$. This feature function implies a penalty imposed on the CTR's of a similar ad pair (a_i, a_j) and the penalty strength is determined by the degree of their similarity. Please note that although we only use one edge feature function here, the model allows multiple relationships by adding more edge functions.

By combining all the feature functions, we obtain the overall conditional probability distribution:

$$P(Y|X) = \frac{1}{Z(X)} \exp\left\{ \sum_i -(y_i - f(x_i; w))^2 + \sum_{j>i} -\beta s_{i,j}(y_i + y_j) \right\}, \quad (4)$$

[3]We use log mods for numerical convenience in learning.

where

$$Z(X) = \int \exp\Big\{ \sum_i -(y_i - f(x_i;w))^2 + \sum_{j>i} -\beta s_{i,j}(y_i + y_j) \Big\} \mathrm{d}Y.$$

It is clear that if we remove the edge feature function, this new model will reduce to the conventional click prediction model as its special case.

4.2 Learning

Given training data $\{X^q, Y^q\}_{q=1}^N$ where $X^q = \{x_1^q, x_2^q, ..., x_N^q\}$ is a set of input feature vectors of ads shown for query q, and $Y^q = \{y_1^q, y_2^q, ..., y_N^q\}$ is the set of the corresponding log mods of CTR's, we estimate the parameters of the continuous CRF model by Maximum Likelihood Estimation (MLE). Suppose the parameters of the continuous CRF model are $\theta = \{w, \beta\}$, we can write the conditional log likelihood of the training data as follows.

$$\begin{aligned} L(\theta) &= \sum_{q=1}^N \log P(Y^q|X^q;\theta) \\ &= \sum_{q=1}^N \Big(-\sum_i (y_i^q - f(x_i^q;w))^2 - \beta \sum_{j>i} s_{i,j}(y_i^q + y_j^q) \\ &\qquad - \log Z(X^q) \Big). \end{aligned} \tag{5}$$

With some mathematical deduction, the normalizer $Z(X)$ becomes

$$Z(X) = \pi^{\frac{n}{2}} \exp\{-c + \frac{1}{4}\alpha^T\alpha\}, \tag{6}$$

where

$$c = \sum_i f(w, X_i)^2$$

and α is a n-dimensional vector whose ith dimension is

$$\alpha_i = 2f(w, X_i) - \beta \sum_{j \neq i} s_{i,j}.$$

We adopt the gradient ascent method to maximize this log likelihood. The gradients of $L(\theta)$ with respect to w and β can be computed as follows.

$$\frac{\partial L(\theta)}{\partial w} = \sum_{q=1}^N \left(\sum_i 2(y_i^q - f(x_i^q;w)) \frac{\partial f(x_i^q;w)}{\partial w} \right) - \sum_{q=1}^N \frac{\partial \log Z(X^q)}{\partial w}$$

$$\frac{\partial L(\theta)}{\partial \beta} = -\sum_{q=1}^N \sum_{i,j} s_{i,j}(y_i^q + y_j^q) - \sum_{q=1}^N \frac{\partial \log Z(X^q)}{\partial \beta}$$

The partial derivatives $\frac{\partial \log Z(X^q)}{\partial \beta}$ and $\frac{\partial \log Z(X^q)}{\partial w}$ are computed in the following way.

$$\frac{\partial \log Z(X^q)}{\partial \beta} = \frac{1}{2} \sum_i \left(\sum_{j \neq i} s_{i,j} \right) \left(\beta \sum_{j \neq i} s_{i,j} - 2f(x_i^q;w) \right)$$

	Query-list pair	query	ad	impression
Training	482,903	140,435	255,788	24,352,968
Testing	514,878	136,525	254,407	24,943,320

Table 1: Statistics of the data set.

and

$$\frac{\partial \log Z(X^q)}{\partial w} = -\sum_i \beta \sum_{j \neq i} s_{i,j} \frac{\partial f(x_i;w)}{\partial w}.$$

To further speed up the learning process, we adopt the Stochastic Gradient Ascent method, which updates the parameters for each training query in an iterative manner.

4.3 Inference

In inference, given parameters θ, we choose the Y^* that maximizes the conditional probability $P(Y|X)$ as the predicted log mods of CTR.

$$\begin{aligned} Y^* &= \arg\max_Y P(Y|X;\theta) \\ &= \arg\max_Y \Big\{ -(Y - f(X;w))^T (Y - f(X;w)) \\ &\qquad - \beta S^T Y \Big\} \end{aligned} \tag{7}$$

where

$$f(X;w) = (f(x_1;w), \cdots, f(x_n;w))^T$$

and

$$S = \left(\sum_{j \neq 1} s_{1,j}, \cdots, \sum_{j \neq n} s_{n,j} \right)^T.$$

Please note that the objective function in Eqn.(7) is concave with respect to Y, and therefore the optimization is easy. Actually, we can get the closed-form solution of the optimization problem, as shown below.

$$Y^* = f(X;w) - \frac{1}{2}\beta S. \tag{8}$$

Due to this closed-form solution, the time complexity of the inference is the same as the conventional click prediction models. This makes the adoption of the relational click prediction model feasible in practice.

5. EXPERIMENTS

In this section, we first describe the settings of our experiments, and then report the experimental results.

5.1 Experimental Settings

We use the same data set as in our data analysis, and split it into three parts, each containing the data of ten days. We use the first part for feature extraction, the second part for training, and the last part for testing. We remove those ad lists with only one ad since there is no need to do relational prediction for them. The detailed statistics of our data set are listed in Table 1.

Feature extraction is not the focus of our paper, and therefore we simply use some representative features according to previous work [4][19]: history COEC (position normalized CTR) for query-ad pair, query, and ad respectively,

smoothed COEC according to query term, ad term, and advertiser's bid information; relevance of ad to query; attractiveness of ad title and description; reputation of advertiser; etc.. As for the relational information, we use the term similarity as introduced in Section 2.

5.2 Baseline Algorithms

To demonstrate the power of relational click prediction, we compare our CRF model with conventional click prediction models. Specially, we implement the logistic regression based model proposed in [4][19] as a baseline. This algorithm only uses the local features that we extracted. For ease of reference, we denote it as LOCAL. In our CRF model, we use the linear regression function as our local model, i.e., $f(x, w) = w^T x$, for simplicity and without loss of generality. We denote our method as CRF. To further examine the power of relational information, we test the performance of a variant of our CRF model with β set to 0. We treat this algorithm as another baseline, denoted as CRF-NR. Note that in CRF-NR, the CRF model has no edge feature function and reduces to a local prediction model that is very similar to logistic regression, because the Y in the model represents the log mods of CTR's.

Following [19], we use the CTR computed from the log data as the ground truth for training.[4] For evaluation, we choose the relative information gain (RIG) [15] and mean square error [19] as the metrics. For ease of discussion, we normalize the MSE of CRF and CRF-NR by the MSE of LOCAL (denoted as NMSE).

5.3 Experimental Results

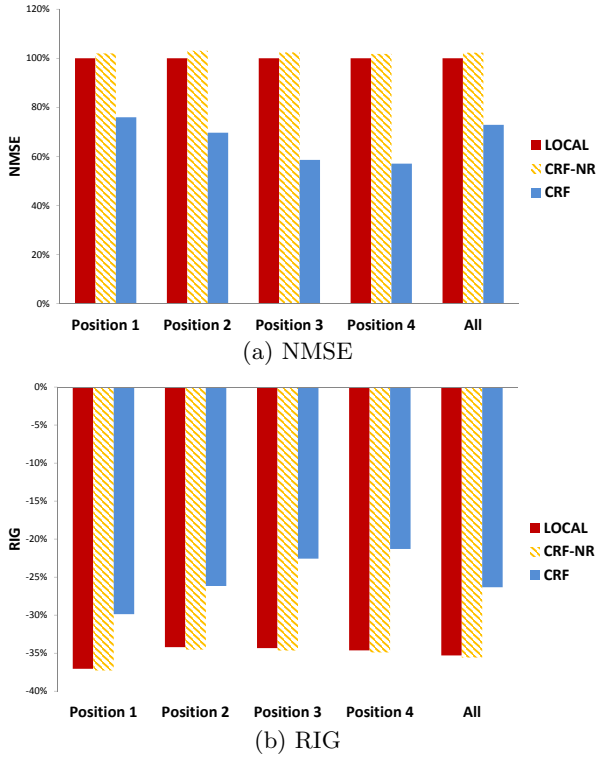
5.3.1 Overall Performance

The parameter β learned by our CRF model from the training data is 0.6810, which is consistent with our intuition and data analysis: a positive β indicates that the CTR of an ad will decrease when the similarity increases. Fig. 7 shows the MSE and RIG of the click prediction algorithms on the test set with respect to different ad positions. Higher RIG and lower MSE mean better prediction accuracy. From the figure, we can see that for all the positions our proposed model significantly outperforms the baselines. The best result is achieved by CRF at slot 4, where it reduces MSE and increases RIG both by about 40% relatively as compared to LOCAL. Even in the worst case, CRF at position 1 still reduces MSE by 25% relatively and increases RIG by 20% relatively. These results clearly show that using relational information in click prediction does improve the prediction accuracy.

Also, the performance of CRF-NR has no significant difference from the conventional click prediction model, but is much worse than that of CRF. As the only difference between CRF-NR and CRF is that our CRF model considers the relational information in its edges, this result demonstrates that the advantage of our method does come from modeling the relational information.

Furthermore, by comparing the results in different ad slots, one could see that our model performs better in lower positions than higher positions. This result consists with the cascade assumption [23] that most users tend to examine search results and ads from top to bottom. So the influence

[4]Please note that our model is also able to deal with $0-1$ ground truth (skip or click).



(a) NMSE

(b) RIG

Figure 7: Results of NMSE and RIG.

of the ads in higher position on the ads in lower position will be stronger, and weaker vice versa. As a consequence, when predicting the CTR of ads in lower positions, the relational information will help more.

5.3.2 Performance at Different Similarity Level

We further study the performance of click prediction with respect to different levels of similarities in the ad lists. The results are shown in Fig. 8. We bucket the ads according to their similarities to others ads in the same ad lists. Then we check the prediction performance in different buckets. As we can see from Fig. 8, when the similarity between ads increases, the performance of CRF also increases, while the performance of baselines does not. This result indicates that our CRF model can effectively handle the relationship information and provide better click prediction, especially when the relationship gets stronger.

To sum up, by comparing the proposed CRF model with a local model, our experiments show that modeling the relationship between ads can significantly increase the accuracy of click prediction.

6. CONCLUSION AND FUTURE WORK

In this paper, we have studied the problem of click prediction in sponsored search. We have reported our data analysis that verifies the mutual influence between ads in their click probabilities. Based on the analysis, we have found that the similarity between ads is an important type of relationship that accounts for the mutual influence, and should be modeled in click prediction. Accordingly, we have proposed a relational click prediction model based on continuous Conditional Random Fields. Our experimental results

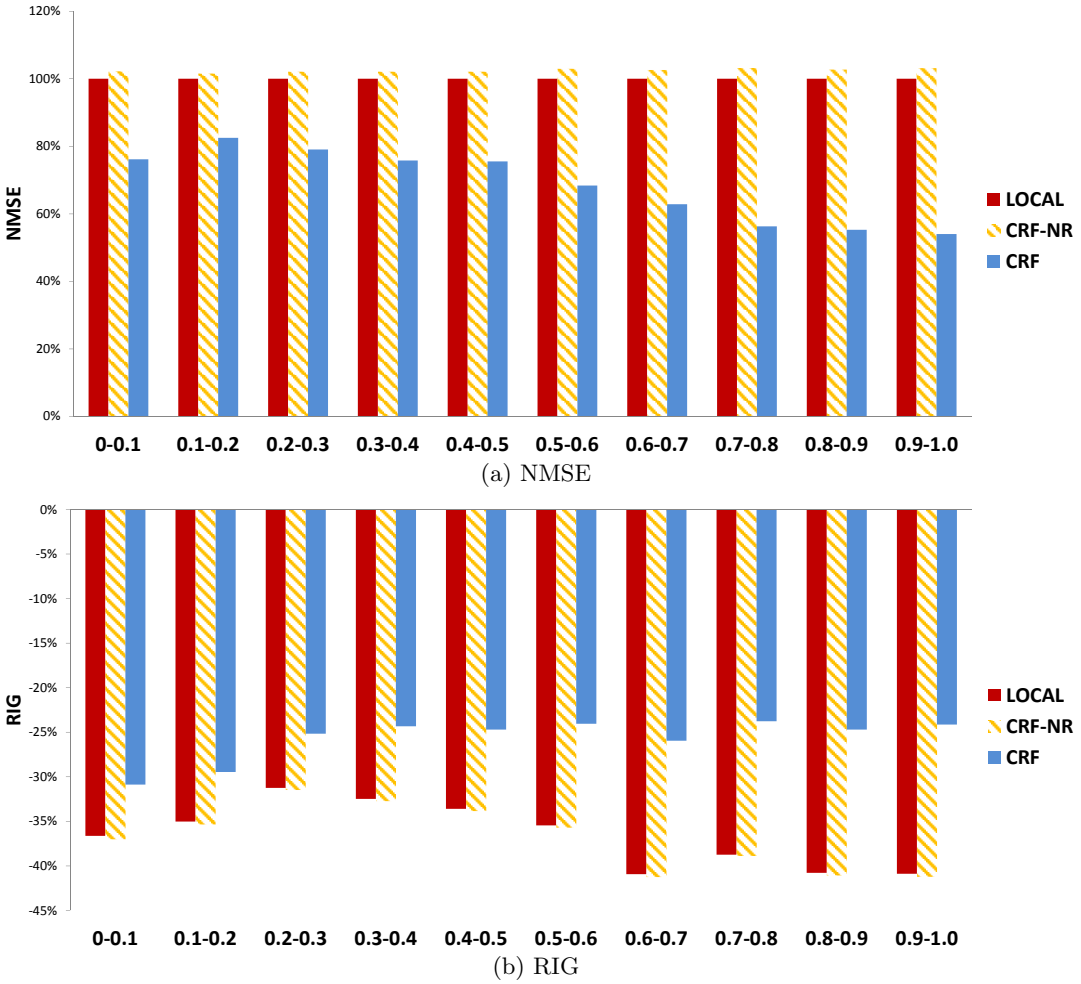


Figure 8: NMSE and RIG results at different similarity levels.

have shown that modeling the relational information can significantly increase the accuracy of click prediction.

For the future work, we plan to investigate on the following challenging but important issues.

- We will study how the mutual influence works for the sidebar ads, i.e., the ads shown in the right hand side of the search result page. The click-through rate on sidebar ads is usually much lower than that on mainline ads. It is highly possible that the mutual influence between sidebar ads will also be different from that between mainline ads. We need to perform a new data analysis to verify this assumption, and adjust our relational click prediction model accordingly.

- We will study more types of relationships. For example, in addition to the term similarity, we can also explore the similarity computed from the click-through bipartite graph. For another example, in addition to the relationship between ads, the relationship with the organic search results [3, 6, 13] should also be considered when predicting the ad click probability. In this case, the relationship will become heterogeneous, and we may need a more powerful mathematical tool to deal with it.

- We will study how to leverage the output of relational click prediction to refine the ranking and pricing components in the sponsored search system. This task turns out to be non-trivial. First, since the relational click prediction is performed at the level of ad list, it will become very time consuming to find the ranked list of ads with the largest expected revenue. Second, it is not clear how the price for each ad should be determined. We believe the solution to the aforementioned task will significantly improve the existing sponsored search systems, and we will have a careful look at it in our future work.

Acknowledgments

We would like to thank our colleagues Ying Zhang, Di He, Wei Chen, Tao Qin, and Bin Gao for their helpful discussions with us on this work.

7. REFERENCES

[1] D. Agarwal, B.C. Chen, and P. Elango. Spatio-temporal models for estimating click-through rate. In *Proceedings of the 18th international conference on World wide web*, pages 21–30. ACM, 2009.

[2] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14. ACM, 2009.

[3] G. Buscher, S.T. Dumais, and E. Cutrell. The good, the bad, and the random: An eye-tracking study of ad quality in web search. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49. ACM, 2010.

[4] H. Cheng and E. Cantú-Paz. Personalized click prediction in sponsored search. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 351–360. ACM, 2010.

[5] M. Ciaramita, V. Murdock, and V. Plachouras. Online learning from click data for sponsored search. In *Proceeding of the 17th international conference on World Wide Web*, pages 227–236. ACM, 2008.

[6] C. Danescu-Niculescu-Mizil, A.Z. Broder, E. Gabrilovich, V. Josifovski, and B. Pang. Competing for users' attention: on the interplay between organic and sponsored search results. In *Proceedings of the 19th international conference on World wide web*, pages 291–300. ACM, 2010.

[7] K. David and M. Mohammad. A cascade model for externalities in sponsored search. In *Proceedings of the 4th International Workshop on Internet and Network Economics*. ACM, 2008.

[8] K. Dembczynski, W. Kotlowski, and D. Weiss. Predicting ads click-through rate with decision rules. In *Workshop on Targeting and Ranking in Online Advertising*, volume 2008. Citeseer, 2008.

[9] Z. Dou, S. Hu, K. Chen, R. Song, and J.R. Wen. Multi-dimensional search result diversification. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 475–484. ACM, 2011.

[10] B. Edelman, M. Ostrovsky, and M. Schwarz. Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords. 2005.

[11] A. Ghosh and M. Mahdian. Externalities in online advertising. In *Proceeding of the 17th international conference on World Wide Web*, pages 161–168. ACM, 2008.

[12] A. Ghosh and A. Sayedi. Expressive auctions for externalities in online advertising. In *Proceedings of the 19th international conference on World wide web*, pages 371–380. ACM, 2010.

[13] S. Gollapudi, R. Panigrahy, and M. Goldszmidt. Inferring clickthrough rates on ads from click behavior on search results. pages 1–5, 2011.

[14] Google. How are ads ranked. In *http://www.google.com/support/grants/bin/answer.py ?hl=en&answer=98917*.

[15] T. Graepel, J.Q. Candela, T. Borchert, and R. Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoftŠs bing search engine. In *Proc. 27th Internat. Conf. on Machine Learning. Morgan Kaufmann, San Francisco, CA*. Citeseer, 2010.

[16] Dustin Hillard, Eren Manavoglu, Hema Raghavan, Chris Leggetter, Erick CantĺŠ-Paz, and Rukmini Iyer. The sum of its parts: reducing sparsity in click estimation with query segments. In *Information Retrieval Journal*. Springer, 2011.

[17] T. Qin, T.Y. Liu, X.D. Zhang, D.S. Wang, and H. Li. Global ranking using continuous conditional random fields. In *Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems (NIPS 2008)*, 2008.

[18] D. Rafiei, K. Bharat, and A. Shukla. Diversifying web search results. In *Proceedings of the 19th international conference on World wide web*, pages 781–790. ACM, 2010.

[19] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pages 521–530. ACM, 2007.

[20] H.M Wallach. Conditional random fields: An introduction. In *Technical report MS-CIS-04-21, University of Pennsylvania*, 2004.

[21] W. Xu, E. Manavoglu, and E. Cantu-Paz. Temporal click model for sponsored search. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 106–113. ACM, 2010.

[22] C.X. Zhai, W.W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 10–17. ACM, 2003.

[23] Z.A. Zhu, W. Chen, T. Minka, C. Zhu, and Z. Chen. A novel click model and its applications to online advertising. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 321–330. ACM, 2010.