Estimating Ad Clickthrough Rate through Query Intent Analysis

Azin Ashkan, Charles L.A. Clarke
School of Computer Science
University of Waterloo, Canada
{*aashkan, claclark*}*@cs.uwaterloo.ca*

Eugene Agichtein, Qi Guo
Mathematics and Computer Science Department
Emory University, United States
{*eugene, qguo3*}*@mathcs.emory.edu*

Abstract

Clickthrough rate and cost-per-click are known to be among the factors that impact the rank of an ad shown on a search result page. Hence, search engines can benefit from estimating ad clickthrough in order to determine the quality of ads and maximize their revenue. In this paper, a methodology is developed to estimate ad clickthrough rate by exploring user queries and clickthrough logs. As we demonstrate, the average ad clickthrough rate depends to a substantial extent on the rank position of ads and on the total number of ads displayed on the page. This observation is utilized by a baseline model to calculate the expected clickthrough rate for various ads. We further study the impact of query intent on the clickthrough rate, where query intent is predicted using a combination of query features and the content of search engine result pages. The baseline model and the query intent model are compared for the purpose of calculating the expected ad clickthrough rate. Our findings suggest that factors such as the rank of an ad, the number of ads displayed on the result page, and query intent are effective in estimating ad clickthrough rates.

1. Introduction

Clickthrough analysis is amongst the implicit feedback techniques that considers the history of user-submitted queries and user-selected results on search result pages. In sponsored search, ad clickthrough rate (CTR) is connected with a notion of ad quality [9]. Clickthrough rate and cost-per-click are known to be among the factors that impact the rank of an ad shown on a search result page. On the one hand, advertisers are charged based on user clicks (if any) on the displayed ads [5]. On the other hand, ad clickthrough rates are known to decrease as ads are displayed in lower ranks on result page, due to reduced visual attention [22]. This is also addressed by Joachims et al. [14] in terms of search engine organic clickthrough, and it is referred as the "trust bias" effect of displayed results based on their ranks.

Several methods for estimating ad CTR have been proposed in the literature [8], [9], [22]. The problem is mostly addressed using content of ads, while influence of factors like position of ads and queries for which the ads appear on the search result page is often not considered. In this regard, the purpose of our work in part is to study the potential of using query intent analysis in ad clickthrough estimation. Our hypothesis is that query intent will influence ad CTRs.

User query intent may correspond to any of the standard categories of Web query [6]: *navigational*, *informational*, and *transactional*. In sponsored search, information providers may also wish to know whether a user has the intention to purchase or utilize a commercial service, or what is called *online commercial intent* [7].

We focus on estimating ad CTRs with respect to history of ads and queries using a query and clickthrough log taken from a commercial Web search engine. The average CTR for search result pages (impression CTR), with respect to the number of displayed ads and rank position of ads, is the basis of our baseline model in estimating ad CTR. In order to study our hypothesis about the influence of query intent on ad CTRs, we address query intent classification, where the intentions underlying user queries are identified based on a combination of query features and the content of search engine result pages (SERPs). In this regard, we use a strategy to reliably and quickly produce a large batch of manually labeled queries. Finally, an intent-aware model (called the query intent model) is proposed to extend the baseline model with the information from query intents. Our findings suggest that factors, such as the intents underlying queries, the total number of ads displayed on a result page and the rank of ads, are effective in estimating ad CTRs.

2. Related Work

Information obtained from implicit feedback resources, such as user query logs [21] have been widely used to interpret and predict user Web search behavior and preferences [1], [14], [19], which may be extended to study query intent. Lee et al. [16] predict user query goals in terms of navigational and informational intent, based on past user-click behavior and anchor-link distribution. Rose and Levinson [23] conducted a study, developing a hierarchy of query goals with three top-level categories: informational, navigational and resource. Under their taxonomy, a transactional query as defined by Broder [6] might fall under either of their three categories, depending on details of the desired transaction. In [3], Baeza-Yates et al. establish three categories from the content of the queries for goals which motivate a user to make a search: informational, non-informational, and ambiguous. They develop models applying supervised and unsupervised learning techniques for query identification purpose. Semi-supervised and supervised learning techniques [4], [17], [18] have been applied over different data resources in order to address query intent classification. Most of the above work focuses on query intent with respect to the traditional navigational, informational, and transactional categories. More recently, there has been growing interest in commercial intent classification. In this regard, Dai et al. [7] propose a commercial query detector. They train machine learning models from search result pages and the contents of the top ranking pages.

In the area of ad clickthrough prediction, Richardson et al. [22] build a prediction model based on logistic regression using statistics of existing ads. They incorporate features from ads, such as their bid terms, the length of ads, the page ads reference, and statistics concerning related ads. In the work by Debmbsczynski et al. [8], the authors did not have access to ad contents and keywords. They approximated the title and the body of an ad by combining all queries for which the ad was displayed. They also used features based on the ad's target URL. They used these extracted features to build a prediction model based on decision rules, and generated recommendations on how to improve the quality of ads. Zhang et al. [25] examine correlation of features of query rewrites with their defined metric, ad clicks over expected ad clicks. They compare features which are predictive of relevance and clicks in user query logs. They show that features like rank position of ads have a high correlation with ad clickthrough . In [11], Ghose et al. study sponsored search at a keyword level. Regelson et al. [20] estimate clickthrough rate of new ads on a keyword basis by using clickthrough rates of existing ads with the same bid terms or topic clusters. In [13], Jansen et. al. study factors influencing ad clicks by searchers.

3. Data Set

The results reported in this paper are based on a data set obtained from Microsoft adCenter, consisting of search and click logs sampled over a few months. The data includes a sample of roughly 100 million search impressions, where an impression is defined as a single search result page. There is also a set of ad clicks associated with the impression data. We removed any extra space at the beginning and end of the queries, and between words of the queries for both impression and clickthrough files. We then case-normalized the queries. Impressions with a duplicate combination of impression id and user session id were removed in order to filter out repeated queries from the same user.

In order to prevent train-test contamination, we randomly split the impression and clickthrough data into three equal-sized sets (i.e. set A, set B, and set C) at the query level. We use set A to train classifiers for query intent detection; we use set B to study the average CTR for impressions with a varying number of ads; we use set C to estimate the CTR for ads with and without considering the user intent. All three sets contain approximately the same number of queries (about 800K) along with their impression and clickthrough information. There are many queries with very small number of ad clicks. Similar to Richardson et al. [22], as our analysis deals with empirical clickthrough, we removed queries with a small number of ad clicks, filtering the three sets to include only queries that have at least four ad clicks. After filtering, we ended up with 45032, 44941 and 44909 queries in sets A, B, and C respectively (134882 queries in total).

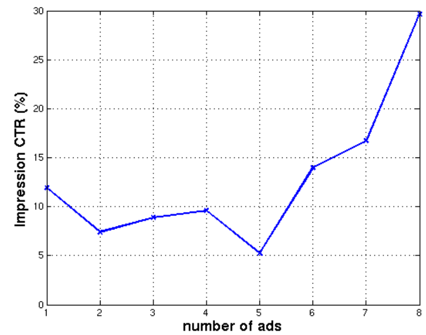

Figure 1. Average CTR for impressions with particular number of ads

4. Average CTR for Search Result Pages

In order to study whether the number and position of displayed ads influence clickthrough rate (CTR), the average number of clicks per impression for queries with a particular number of ads is examined in this section. First, impressions are sorted according to the number of ads displayed for each. The number of ads in the impression file varies from one to eight. Thus, impressions are divided into eight groups, each denoted as set A_i, where i is the number of displayed ads for the impressions in that set. The value $|A_i|$ indicates the number of impressions with i ads displayed. We use the unique id number for each impression (impression id) to determine whether it resulted in an ad click. Repeating this process for all impressions in the eight groups, we can calculate the total number of ad clicks for each.

Let $a_i^j \in A_i$ denote the unique impression id for the j^{th} impression in A_i. We define c_i^j to represent whether there was an ad click resulting from such an impression. In other words, $c_i^j = 1$, if there is an ad click associated with a_i^j in the clickthrough data, and $c_i^j = 0$ otherwise. Hence, the average number of ad clicks per impression, CTR_i, for queries with a particular number of ads i, is obtained as follows:

$$CTR_i = \frac{\sum_{j=1}^{|A_i|} c_i^j}{|A_i|} \qquad 1 \leq i \leq 8 \qquad (1)$$

We calculated the average CTR for the eight ad-based groups of set B, resulting in the plot depicted in Figure 1. For clarity of presentation, the points for each particular number of ads are connected. The lines do not imply interpolation.

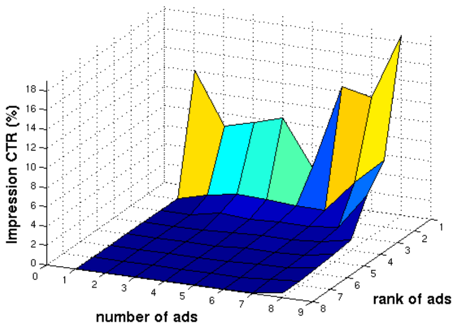

Figure 2. Average CTR at specific ranks for impressions with particular number of ads

Generally speaking, the more ads displayed, the more clicks they receive. This observation could indicate that the number of ads (in part) determines the number of ad clicks. The dip at rank 5 will be explained shortly. Figure 2 shows the same click-to-impression ratio trend when the rank of the clicked ads is also considered. Note that ad clicks mostly occur at the first and the second ranks, and most especially at the first rank. This observation confirms that CTR of ads decrease as they are displayed in lower ranks on the page, due to reduced visual attention [22].

4.1. Impact of Ad Location on CTR

In addition to rank of ads, their location (top or side of a result page) could also affect clickthrough rates. Unfortunately, our logs do not record the locations of ads. A single ad at rank 1, may appear either on the top or at the side. According to Jansen [12], top-listed ads are assumed to be more relevant than organic results and side-listed ads. This could affect the frequency of clicks for ads at different locations. In this regard, we hypothesize that peaks and valleys in the two plots (Figures 1 and 2) could result from the location of different ads for which the clicks are recorded.

As mentioned before, the total number of displayed ads varies from 1 to 8. A batch of approximately 43,000 queries (randomly selected from sets A, B, and C) were submitted to the Live search engine (www.live.com) in order to study different possibilities in terms of the number of ads and their locations on the search engine result pages (SERPs, and also known as impressions). The results are shown in Table 1 in percentage form. According to this experiment, the maximum number of ads displayed at the right side of a result page is 5, while the maximum number of ads displayed on top of a search result page is 3. It is also assumed the rank of ads are assigned in a way that the ads displayed on the top (if any) are ranked higher than the ones display at the side. As an example, if a_{t_1} and a_{t_2} are two ads displayed on top of the page as the first ad and the second ad respectively, while a_{s_1} is displayed at the side, the rank order of these three ads with respect to the page is considered as follows: a_{t_1}, a_{t_2}, and a_{s_1}. As ads receive clicks mostly at the first rank (Figure 2), we base our analysis on clicks at the first rank

Table 1. Percentage of SERPs with particular number of ads on the top and on the side

	$t=0$	$t=1$	$t=2$	$t=3$
$s=0$	11.7	3.8	1.1	0.38
$s=1$	6.8	3.4	1.2	0.68
$s=2$	4.1	2.8	1.3	0.88
$s=3$	2.7	2.2	1.4	1.2
$s=4$	1.8	1.5	1.1	1.3
$s=5$	3.6	5.1	5.4	34

in this section. We also base our analysis on the following assumption in order to simplify the calculations:

> **Assumption 1 -** Clicking on a top ad is *independent* of the number of ads displayed at the side of a result page.

Let R be a random variable characterizing the distribution of possible ranks of the ads at which clicks occur, and N represents the total number of ads displayed on a result page. The average CTR for result pages, in which $N = n$ ads are displayed and the ad at rank $R = r$ is clicked, can be denoted as $P(R = r|N = n)$. Let T and S represent the number of displayed ads on the top and at the side of a result page respectively, hence $N = T + S$. The probability of appearance of $T = t$ ads on the top and $S = s$ ads at the side of a result page conditioned on the total number of ads, $N = n$, displayed on the result page is $P(T = t, S = s|N = n)$. It is noted that, each cell in Table 1 represents the likelihood of SERPs with $T = t$ ads on the top and $S = s$ ads at the side (i.e. $P(T = t, S = s)$). Thus, the aforementioned conditional probability can be calculated from Table 1 as follows:

$$
\begin{aligned}
P(T=t,S=s|N=n) &= \frac{P(T=t,S=s,N=n)}{P(N=n)} \\
&= \frac{P(T=t,S=s)}{P(N=n)} \qquad (2)
\end{aligned}
$$

where $s + t = n$ and $P(N = n)$ can be calculated by the summation of the corresponding probabilities in Table 1 (e.g. $P(N = 1) = P(T = 0, S = 1) + P(T = 1, S = 0)$). The average CTR at the first rank for varying number of ads can be seen as $P(R = 1|N = n)$ for all possible values of n (i.e. $1 \leq n \leq 8$). In order to infer ad CTR with respect to their location (top or side) on the page, $P(R = 1|N = n)$ from Figure 2 and the estimation of different possibilities in terms of the number of ads and their locations on the search engine result pages from Table 1 are used. For $N = 8$, according to Table 1, there is only one possibility where $S + N = 8$ that is $T = 3$ and $S = 5$. Hence, $P(T = 3, S = 5|N = 8) = 1$, and therefore $P(R = 1|N = 8)$ is estimated as follows:

$$
\begin{aligned}
& P(R=1|N=8) \\
= \;& P(R=1|T=3,S=5)P(T=3,S=5|N=8) \\
\overset{(a)}{=} \;& P(R=1|T=3)\times 1 \\
\Rightarrow \;& \boxed{P(R=1|T=3)=P(R=1|N=8)=0.19} \qquad (3)
\end{aligned}
$$

where $P(R = 1|N = 8) = 0.19$ comes from Figure 2, and

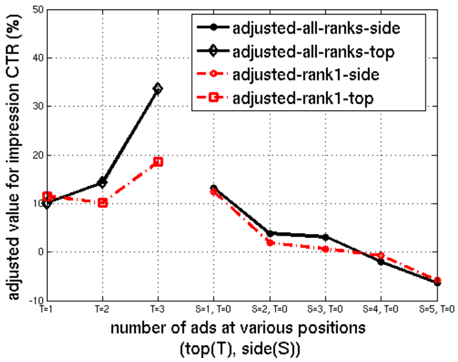

Figure 3. Adjusted plots for average CRT for impressions with a particular number of ads on top/ side of the page

the substitution of $P(R = 1|T = 3, S = 5)$ by $P(R = 1|T = 3)$ in (a) comes from Assumption 1.

Similarly, for impressions with $N = 7$ ads displayed, there are two possibilities according to Table 1: i) $T = 3$, $S = 4$ and ii) $T = 2$, $S = 5$. Hence, according to Equation 2, $P(T = 3, S = 4|N = 7)$ and $P(T = 2, S = 5, |N = 7)$ can be estimated as 0.2 and 0.8 respectively. We also have:

$$\begin{aligned}
& P(R=1|N=7) \\
= \;& P(R=1|T=3,S=4)P(T=3,S=4|N=7) + \\
& P(R=1|T=2,S=5)P(T=2,S=5|N=7) \\
= \;& P(R=1|T=3)\times 0.2 + P(R=1|T=2)\times 0.8 \\
= \;& (0.19\times 0.2) + (0.8\times P(R=1|T=2)) \\
\Rightarrow \;& P(R{=}1|T{=}2) = \frac{1}{0.8}\times (P(R{=}1|N{=}7) - 0.037) \\
\Rightarrow \;& \boxed{P(R=1|T=2) = 0.103} \qquad (4)
\end{aligned}$$

where $P(R = 1|N = 7) = 0.118$ comes from Figure 2. Similarly, we have:

$$\boxed{P(R=1|T=1) = 0.115} \qquad (5)$$

Using the estimated values of $P(R = 1|T = t)$ (for $1 \leq t \leq 3$) from Equations 3, 4, and 5 and also the values of $P(R = 1|N = n)$ (for $1 \leq n \leq 5$) plotted in Figure 2, we can estimate the average CTR at the first rank for cases where no ads are displayed on the top and at least one ad is displayed on the side. In other words, $P(R = 1|T = 0, S = s)$ is estimated for $1 \leq s \leq 5$, while the total number of ads displayed on a result page is at most 5 ($1 \leq n \leq 5$):

$$\begin{aligned}
& P(R=1|N=n) \\
= & \sum_{t=0}^{min\{n,3\}} P(R{=}1|T=t, S=n-t)P(T{=}t, S{=}n-t|N{=}n) \\
\Rightarrow \;& P(R=1|T=0,S=s) = \\
& \frac{1}{P(T=0,S=s|N=s)} \times [P(R=1|N=s) - \\
& \sum_{t=1}^{min\{s,3\}} P(R=1|T=t)P(T=t,S=s-t|N=s)]
\end{aligned}$$

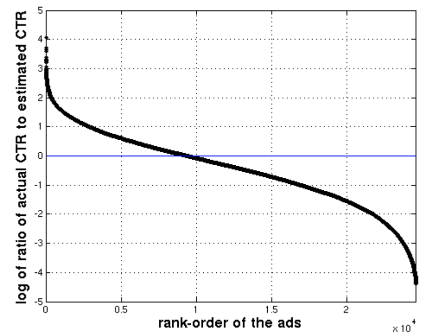

Figure 4. Log of ratio of the actual CTR to the estimated CTR for ads using the baseline model

The estimated values $P(R = 1|T = t)$ (where $1 \leq t \leq 3$) and $P(R = 1|T = 0, S = s)$ (where $1 \leq s \leq 5$) are referred as the *adjusted values* at the first rank. These values along with the corresponding values at all ranks (calculated similarly for all ranks in general) are plotted in Figure 3 in percentage form. Due to Assumption 1 which simplifies the calculations, there are a few values estimated as negative numbers, which we treat as zero.

The trend of changes in the adjusted values can be used to explain the dips in Figures 1 and 2. According to Figure 3, the least estimated value is for the case where there are 5 ads displayed on the side while no ads appear on the top (i.e. $S = 5$, $T = 0$). This can be viewed as a reason for the dip at 5 in Figures 1 and 2. As can be observed in Figure 1, the more ads displayed on the top, the more clicks they would receive (i.e. $T = 1$, $T = 2$, $T = 3$). Comparing the values for top ads (i.e. the first three on the plot) and the ones with no top ads (i.e. the last five on the plot), it is observed that ads on the top of a result page are more often the targets of clicks than the ads on the side. In the case where no ad is displayed on the top (i.e. the last five on the plot), a single side ad looks most likely to be the target of click.

5. Estimating Ad CTR: The Baseline Model

As depicted in Figure 2, the average CTR for search result pages depends not only on the rank position of ads, but also on the total number of ads displayed on the page. We introduce a *baseline model* based upon this observation in order to estimate the expected number of clicks for an ad.

In order to calculate the expected number of clicks for ads, we use set C. There are 83,550 unique ad ids with at least one click in this set, while they do not belong to set B (in order to avoid any common ad between train and test data). For a given ad a, its expected number of clicks is calculated as follows:

$$E_{click}(a) = \sum_{n=1}^{8}\sum_{r=1}^{n} P(R=r|N=n)\times Num(a,r,n) \quad (6)$$

where $Num(a, r, n)$ denotes the number of impressions in which n ads are displayed in total, while ad a appears at rank r ($1 \leq r \leq n$). Also $P(R = r|N = n)$ is the average CTR (Figure 2) at rank r when n ads are displayed on the result page in total. The expected CTR for ad a is simply obtained by dividing $E_{click}(a)$ by the total number of impressions in which ad a appears. The expected CTR for the $83,550$ ads in set C was calculated. The ratio of the actual number of clicks to expected number of clicks (which is the same as the ratio of estimated CTR to the actual CTR) was also calculated. The ads have then been sorted with respect to their ratio value. Figure 4 depicts the sorted values in logarithmic form.

In similar studies, ads with too few impressions are filtered: Richardson et al. [22] and Debmbsczynski et al. [8] select threshold values of 100 and 200 respectively in order to filter ads for the purpose of clickthrough estimation. Ads with too few impressions may lead to noise in the estimation process. Therefore, the sequence of ratios shown in Figure 4 is for the 24,771 ads that have greater than 100 impressions.

We evaluate the result of baseline model later in the paper, when we compare this model with the case where query intent is also considered in the estimation process.

6. Query Intent Analysis

Our purpose, in part, is to identify the intention underlying queries as *commercial* or *noncommercial*. The standard categories of Web queries by Broder [6] are also studied in this paper, however we subsume transactional queries under one of the two categories of either *navigational* or *informational*, as appropriate. In previous work [2], we address query intent classification with respect to two combinations of features: i) ad clickthrough features combined with query and SERP features, and ii) the combination of query and SERP features with no ad clickthrough features. As it will be explained later, by SERP features, we mean the features extracted from the contents of search engine result pages. These previously published results [2] indicate that the first combination outperforms the second one in terms of precision, recall, and accuracy. However, we use the later set of features in this paper in order to avoid any possible interaction between ad clickthrough features and the number of ad clicks.

6.1. Features

Our rationale for using content of result pages as a feature set is that pages can be viewed as representatives of the nature of a query. For instance, if a query such as “cheap shoes” is entered by user, the appearance of the terms “buy”, “free”, and “shipping” may be good indicators of the commercial nature of the query. Features extracted from the query itself are also considered to be helpful in understanding the intention underlying the query. Such query features used in our work include: number of characters and words in the query string, and a binary feature indicating whether the query string contains any URL elements (e.g. .com or .org).

To extract SERP features, we submitted each query to the Live search engine and downloaded the first result page for that query. Each SERP is then presented as a “bag of words” (i.e an unordered multi-set of terms frequencies). Note that the terms were extracted from the organic results only. Ads were removed, to avoid any possible bias that ad keywords might produce in the classification.

6.2. Labeling Process: Mechanical Turk

In previous work [2], we manually labeled a set of 1700 web queries with respect to their two dimensions of intent. Those queries were independently labeled by three researchers from our group. The annotators were responsible for judging the presumed intent of the queries from the perspective of a general user, requiring considerable time and effort. For this work we aimed to increase our set of labeled queries relatively easily, cheaply, and reliably. Hence, we employed Amazon Mechanical Turk (www.mturk.com/mturk). According to Amazon, “Mechanical Turk is based on the idea that there are still tasks that human beings can do much more effectively than computers, such as identifying objects in a photo or video”, or in our case manually labeling queries. Amazon calls these tasks *HITs* (human intelligence tasks). A HIT represents a single, self-contained task that a so-called *worker* can work on, submit an answer, and collect a reward for completing.

In order to train and evaluate the classifier, 3000 extra queries (among the 45K queries in set A) were selected in order to be manually labeled along the two dimensions of query intent. The original impression file was sorted based on the time of the impression. Starting from an arbitrary point in the file (approximately $1/5$ of the length of the file from the beginning), 3000 queries were selected for which the query was already contained in set A and it was not among the previously labeled 1700 queries. This selection approach was used in order to pick 3000 queries from a continuous period of time among the impressions. We refer to this set as the *MT set*. We randomly selected 1000 queries of the 1700 manually labeled queries as a *seed set* in order to be used to validate the results obtained from the experiment with Mechanical Turk. Consequently, we ended up with 4000 queries for labeling and eventually for training the classifier. The entire set of selected queries (i.e. 4000 queries) were then divided into 40 batches of 100 queries, with each batch containing 25 seed queries and 75 MT queries. These batches were submitted to Mechanical Turk, each as a single HIT, in order to be labeled according to instructions that we provided for the workers. They were asked to judge the presumed intent of the search queries from the perspective of a general user as follows:

> If the presumed purpose of submitting a query is to make an immediate or future purchase of a product or service, the query is labeled as “commercial”. Otherwise, it is labeled as “noncommercial”. If the presumed purpose of a query is to locate a specific Website, the query is labeled as “navigational”. Everything else is considered “informational”.

For each batch labeled and submitted by a worker, we compared the labels assigned to the seed queries of the

Table 2. Accuracy of the query intent classifiers

Query Intent	Precision	Recall	Accuracy
commercial/ noncommercial	0.89	0.85	88%
navigational/ informational	0.87	0.84	84%

batch with the actual labels of those queries (previously determined by three local annotators). If the agreement of the worker with our annotators was found to be above 60% (a threshold above the random case), the labels assigned by this worker were accepted. Otherwise, the labels were rejected and the same batch of queries were submitted for an extra round of labeling. If the agreement was found to be above 75% , we awarded a bonus to the worker. We continued this process until all batches were successfully labeled by five different workers. The final label of each query was assigned based on the majority of labels obtained for the query.

Fleiss' kappa [10] was used as a measure to evaluate the agreement among the workers. It is defined as the proportion of agreement corrected for chance and it works for a number of annotators, each giving categorical labels to the entire set of a fixed number of items. Hence, in each dimension of query intent, we measured Fleiss' Kappa for every submitted HIT which were labeled by five independent workers. The final Kappa value for the entire set and in regards to a particular query intent dimension is calculated as the mean of Fleiss' Kappa for each HIT. The mean Kappa was measured as 0.6028 and 0.5948 for commercial/ noncommercial and navigational/ informational dimensions respectively. Referring to the levels of interpretation for Kappa proposed by Landis and Koch [15], there are respectively *substantial* and *moderate* agreements among the workers in each dimension.

6.3. Results

As a result of labeling, 42% of queries were labeled as *commercial* and 58% were labeled as *noncommercial*, while 55% of queries were labeled as *navigational* and 45% were labeled as *informational*. The 4000 labeled queries, along their features were used to train an SVM classifier using the SVM-light package [24]. A separate classifier was trained for each dimension of query intent. The prediction accuracy of each classifier using 10-fold cross validation is presented in Table 2. The trained classifiers were used to classify queries from both sets B and C. These two sets along with the identified intents underlying them are the subjects of further investigation in the remainder of the paper.

Now that the apparent intention underlying each query has been determined, we follow a similar approach to that of Equation 1, calculating the average CTR for all the impressions with a particular number of ads. However, this time, we consider only the impressions for which the associated queries fall into a given class. The average CTR for the four possible combinations of query class in pairs (i.e. commercial- navigational, commercial- informational, noncommercial- navigational, and noncommercial- informational) against the number of ads are plotted in Figure 5. The plot from Figure 1 is also placed in Figure 5 in order

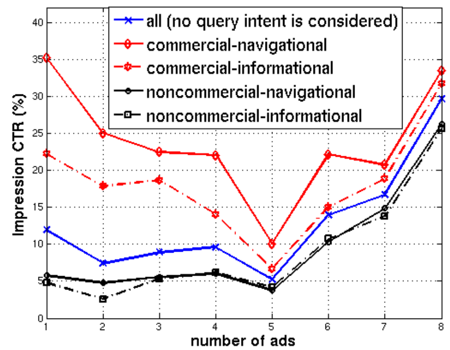

Figure 5. CTR for impressions with particular number of ads and associated with various query types

to provide a baseline for comparison. Note that the plots indicate the average CTR for impressions with a particular number of ads and associated with particular classes of query intent. As it is depicted in the figure, the average CTR differs for various classes of queries.

The peaks and valleys in the plots can be explained due to the location of different ads, similar to what we demonstrated for the general case. Similar to Figure 2, the 3d version of plots for pairs of query types can be obtained by considering the rank of the clicked ads. The plots show similar trends to those of Figure 2, but are omitted for space reasons. We note that commercial-navigational queries receive more ad clicks than commercial-informational queries, on average. Our intuition for explaining this would be the fact that a query is navigational restricts the top result links (either ad or organic results) to a particular Website. Therefore, the top results for a navigational query would more likely match the user's requirements. This match could result in more ad clicks for navigational queries in comparison to informational queries. An example for illustrating this difference is "American airlines" as a commercial-navigational query against "airline tickets" as a commercial-informational query. The chance that user would find a related ad for the former query is greater than the latter, because the former query is restricted by the airline name. As Jansen suggests [12], searchers approach e-commerce searching from two major perspectives, one to look for a specific product or service, and the other to receive information. We believe that the commercial-navigational queries fall mostly into the former category, while the commercial-informational queries fall into the latter one.

7. Estimating Ad CTR: The Query Intent Model

In this section, we study the impact of query intent on estimating ad CTR. The motivation is to consider the impact of identified query intent as an extra factor in the baseline model. In other words, the expected number of clicks for an ad is calculated with respect to: i) the total number of ads on the page, ii) the rank of the ad on the page, and iii) the intents underlying the query for which the ad is displayed. We call this model the *query intent model*.

Table 3. Performance measures of the results

	Filtering Setting	MSE (1e-3)	KLD (1e-2)	ME	Improvement w.r.t the three metrics
Baseline Model	-	34.98	28.16	0.59	-
Query Intent Model	-	34.41	27.75	0.56	1.62%, 1.45%, 5.08%
Baseline Model	$(imp > 100)$, $(actual\ CTR > 1\%)$	3.80	3.90	0.53	-
Query Intent Model	$(imp > 100)$, $(actual\ CTR > 1\%)$	3.09	3.46	0.51	18.68%, 11.28%, 3.77%
Baseline Model	$(imp > 100)$	2.64	2.33	1.08	-
Query Intent Model	$(imp > 100)$	2.18	2.06	0.99	17.42%, 11.58%, 8.33%
Baseline Model	$(imp > 200)$	2.58	2.12	1.37	-
Query Intent Model	$(imp > 200)$	2.09	1.83	1.25	18.99%, 13,67%, 8.75%
LR Model [22]	$(imp > 100)$	3.75	2.86	-	-
MLE Model [8]	$(imp > 200)$	2.24	-	-	-

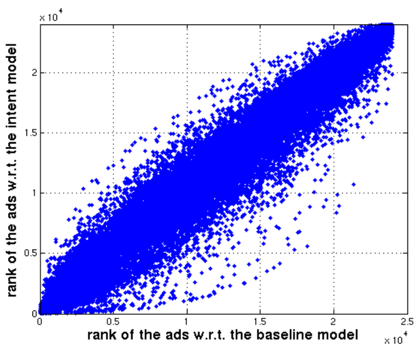

Figure 6. Movement in the ads rank order in the results of the two models

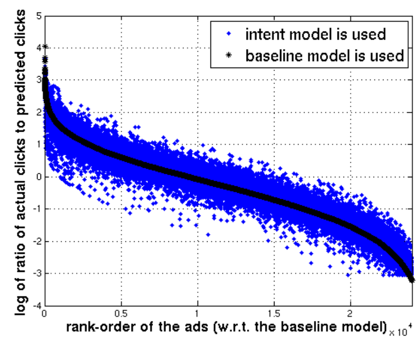

Figure 7. Changes in the ratio of the actual CTR to the estimated CTR per ad in the results of the two models

Let $P(R = r|N = n, I = i)$ be the average CTR at rank r when n ads are displayed on the result page in total and the impression belongs to a query with the intent i. The intent i can be one of the pairs: commercial- navigational (cn), commercial- informational (ci), noncommercial-navigational (nn), and noncommercial- informational (ni). Hence, $G = \{cn, ci, nn, ni\}$ is defined as the set of pairs of query intents. The expected number of clicks for a given ad a is then calculated as follows:

$$E_{click}^{int}(a) = \sum_{n=1}^{8} \sum_{r=1}^{n} \sum_{i \in G} P(R = r|N = n, I = i) \times Num(a, r, n, i) \tag{7}$$

where $Num(a, r, n, i)$ denotes the number of impressions that belong to queries with intent i and in which n ads are displayed, while ad a appears at rank r. The expected CTR for ad a is obtained by dividing $E_{click}^{int}(a)$ by the total number of impressions in which ad a appears.

The expected CTR for the $83,550$ ads in set C has been calculated based on this model. The ratio of actual number of clicks to expected number of clicks has been calculated per ad. Similar to what was done for the baseline model, ads with less than 100 impressions were filtered out. The ads were then sorted with respect to their ratio value, so that a rank number was assigned to each ad based on its position in the sorted result. Figures 6 and 7 depict the results of the baseline model (from Figure 4) and the ones obtained based on the query intent model together for comparison purposes.

In Figure 6, each point represents a single ad for which its ranks with respect to the baseline and intent models are respectively shown in the x-axis and y-axis. As may be observed from Figure 7, for almost all ads, the ratio value plotted using the intent model is different from the one based on the baseline model (the ratio value for each ad appears either above or below its corresponding value based on the baseline model). The observations from both figures indicate that the estimation results obtained based on the two models vary substantially from each other. In order to compare them, we collected some metrics that have been used for the same purpose in literature, and calculated them for each case separately. MSE (mean square error) has been used by both Richardson et al. [22] and Debmbsczynski et al. [8] as a performance metric. Richardson et al. [22] also reports the average KL-divergence (KLD) between their model's predicted CTR and the actual CTR. A perfect model would score 0 based on the KL-divergence metric. Our final metric is ME (mean error) as a measure of error used by Regleson and Fain [9]. ME is the average error, where the error E is calculated as $E = \frac{|estimated\ CTR - actual\ CTR|}{actual\ CTR + 1}$. As is explained by Regleson and Fain [9], this measure ensures that large errors on small rates are not neglected.

Table 3 reports values of the aforementioned metrics for the two proposed models with different filtering settings. As mentioned previously, there have been various filtering settings performed in literature in order to decrease the amount of noise in the results of the empirical estimation of CTR. The results based on those settings are reported in the table, along with a few extra settings that we used in our experiment. As mentioned before, the main purpose behind filtering is to avoid any possible noise from ads with too few impressions or with very low clickthrough rates. One can observe that within all settings of filtering and with respect to all three metrics, the proposed query intent model outperforms the baseline model. This suggests that the intention underlying a query can help to determine the number of clicks for ads displayed as the result of the query.

In Table 3, the best reported results from two previous papers in this area are presented: i) a model based on logistic regression using statistics of existing ads (LR model) proposed by Richardson et al. [22], and ii) a model based on maximum likelihood estimation (MLE) proposed by Debmbsczynski et al. [8]. A report of the results from the work by Regleson and Fain [9] is not listed in the table, because their results are based on term CTR prediction not on ad CTR prediction. Although details of our work may differ, the reported numbers suggest that our models, especially the query intent model, are promising for estimating ad CTRs.

8. Conclusions

We study the estimation of ad clickthrough rates with respect to history of ads and queries using a query and clickthrough log taken from a commercial Web search engine. Our contributions include: i) studying the impact of ad location (side or top of a result page) on the average CTR for impressions, ii) introducing a baseline model to estimate ad CTR, iii) demonstrating that large batches of manually labeled queries may be produced quickly and reliably, iv) training classifiers for the purpose of query intent detection, and v) introducing an intent-aware model in order to improve the proposed baseline model in estimating ad CTRs.

The estimation results obtained based on the two models vary substantially from each other. In order to compare them, we collected some that have been used for the same purpose in the literature, and calculated them separately for each case and for various filtering settings. Our findings suggest that the proposed models, especially the query intent model, are promising for estimating ad CTRs.

9. Acknowledgments

We would like to thank Microsoft Research and Microsoft adCenter for the Beyond Search data through the Beyond Search RFP, and for supporting this work.

References

[1] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting Web search result preferences. In *Proceedings of the* 29^{th} *ACM SIGIR*, pages 3–10, 2006.

[2] A. Ashkan, C. Clarke, E. Agichtein, and Q. Guo. Classifying and characterizing query intent. *Proceedings of the* 31^{st} *European Conference on Information Retrieval*, pages 578–586, 2009.

[3] R. Baeza-Yates, L. Calderán-Benavides, and C. González-Caro. The intention behind Web queries. In *Lecture Notes in Computer Science: SPIRE*, volume 4209, pages 98–109. Springer Berlin, 2006.

[4] S. Beitzel, E. Jensen, O. Frieder, D. Grossman, D. Lewis, A. Chowdhury, and A. Kolcz. Automatic Web query classification using labeled and unlabeled training data. In *Proceedings of the* 28^{th} *ACM SIGIR*, pages 581–582, 2005.

[5] R. Briggs and N. Hollis. Advertising on the Web: Is there response before clickthrough. *Journal of Advertising Research*, 37(2), 1997.

[6] A. Broder. A taxonomy of Web search. *ACM SIGIR Forum*, 36(2):3–10, 2002.

[7] H. Dai, L. Zhao, Z. Nie, J. Wen, L. Wang, and Y. Li. Detecting Online Commercial Intention (OCI). *Proceedings of the* 15^{th} *International Conference on WWW*, pages 829–837, 2006.

[8] K. Debmbsczynski, W. Kotlowski, and D. Weiss. Predicting ads clickthrough rate with decision rules. *Workshop on Target and Ranking for Online Advertising*, WWW 2008.

[9] D. C. Fain and J. O. Pedersen. Sponsored search: A brief history. *Bulletin of the American Society for Information Science and Technology*, 32(2):12–13, 2006.

[10] J. Fleiss and J. Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613, 1973.

[11] A. Ghose and S. Yang. An empirical analysis of sponsored search performance in search engine advertising. *Proceedings of the International Conference on Web Search and Web Data Mining*, pages 241–250, 2008.

[12] B. Jansen. The comparative effectiveness of sponsored and nonsponsored links for Web e-commerce queries. *ACM Transactions on the Web*, 1(1), 2007.

[13] B. Jansen, A. Brown, and M. Resnick. Factors relating to the decision to click on a sponsored link. *Decision Support Systems*, 44(1):46–59, 2007.

[14] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the* 28^{th} *annual international ACM SIGIR*, pages 154–161, 2005.

[15] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.

[16] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in Web search. *Proceedings of the* 14^{th} *International Conference on WWW*, pages 391–400, 2005.

[17] X. Li, Y. Wang, and A. Acero. Learning query intent from regularized click graphs. In *Proceedings of the* 31^{st} *ACM SIGIR*, pages 339–346, 2008.

[18] D. Nettleton, L. Calderán-Benavides, and R. Baeza-Yates. Analysis of Web search engine query session and clicked documents. In *Lecture Notes in Computer Science: Advances in Web Mining and Web Usage Analysis*, volume 4811, pages 207–226. Springer Berlin, 2007.

[19] B. Poblete and R. Baeza-Yates. Query-sets: using implicit feedback and query patterns to organize Web documents. 2008.

[20] M. Regelson and D. Fain. Predicting clickthrough rate using keyword clusters. *Proceedings of the* 2^{nd} *Workshop on Sponsored Search Auctions*, 2006.

[21] M. Richardson. Learning about the world through long-term query logs. *ACM Transactions on the Web*, 2008.

[22] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the clickthrough rate for new ads. *Proceedings of the* 16^{th} *International Conference on WWW*, pages 521–530, 2007.

[23] D. Rose and D. Levinson. Understanding user goals in Web search. *Proceedings of the* 13^{th} *International Conference on WWW*, pages 13–19, 2004.

[24] T. Joachims, SVMlight support vector machine, 2008. URL = http://svmlight.joachims.org.

[25] W. Zhang and R. Jones. Comparing click logs and editorial labels for training query rewriting. In *Query Log Analysis Workshop, WWW*, volume 7.