Classifying and Characterizing Query Intent

Azin Ashkan[1], Charles L.A. Clarke[1], Eugene Agichtein[2], and Qi Guo[2]

[1]University of Waterloo, Canada
[2]Emory University, United States

Abstract. Understanding the intent underlying user queries may help personalize search results and improve user satisfaction. In this paper, we develop a methodology for using ad clickthrough logs, query specific information, and the content of search engine result pages to study characteristics of query intents, specially commercial intent. The findings of our study suggest that ad clickthrough features, query features, and the content of search engine result pages are together effective in detecting query intent. We also study the effect of query type and the number of displayed ads on the average clickthrough rate. As a practical application of our work, we show that modeling query intent can improve the accuracy of predicting ad clickthrough for previously unseen queries.

1 Introduction

Understanding the intent underlying user queries may help personalize search results and therefore improve user satisfaction. User intent may correspond to any of the standard categories of Web query [2]: *navigational*, *informational*, or *transactional*. In the context of sponsored search, information providers may also wish to know whether a user intents to purchase or utilize a commercial service, or what is called *online commercial intention* [4]. In this regard, we define a *commercial* query as a query with the underlying intention to make an immediate or future purchase of a specific product or service. We place all other queries into the *noncommercial* category. Furthermore, a *navigational* query is defined as a query with the underlying intention to locate a specific Web site or page, while an *informational* query is everything else.

In the first part of this paper, query intent detection is addressed based on two different settings of features: i) the ad clickthrough features are combined with the query and SERP features (obtained from the content of Search Engine Result Pages), and ii) the combination of query and SERP features are used while no ad clickthrough features are involved. By "ad(s)" throughout the paper, we mean advertisement(s) that are presented on top or right side of the result page returned by a search engine. Although it is shown that a classifier based on the former setting outperforms the later one, we use the later set of features in the rest of the paper (for studying characteristics of different query intents) in order to avoid any possible distortion the ad clickthrough features could create. The paper consolidates and substantially extends previous work [1] where the aim is to distinguish between characteristics of different query types according to their ad clickthrough behavior. We characterize query intent based on the average clickthrough rates, which suggests that these clickthrough rates depend on factors such as the nature of the information need, the number of ads displayed for the query, and possibly the position of ads on the page. We also show that factors like query intent and number of displayed ads can be used to estimate ad clickthrough for previously unseen queries.

2 Related Work

Dai et al. [4] propose a commercial query detector. Their findings indicate that frequent queries are more likely to have commercial intent. In the context of query intent detection based on clickthrough data, Lee et al. [8] predict user query goals in terms of navigational and informational intent. They show that prediction can be performed based on two types of feature sets: past user-click behavior and anchor-link distribution.

Regelson and Fain [9] estimate the clickthrough rate of new ads by using the clickthrough rates of existing ads with the same bid terms or topic clusters. Similar work by Richardson et al. [10] incorporates features that depend on more than just the bid terms, including information about the ad itself, such as the length of the ad, the page the ad points to, and statistics concerning related ads. Debmbsczynski et al. [5] approximate the title and the body of each ad by combining all queries for which a given ad was displayed. They also use features based on the search result page and on the ad's target URL in order to build a prediction model based on decision rules, generating recommendations on how to improve the ad quality. All three efforts focus on ad-based features to predict the clickthrough rate and quality of new ads.

Jansen et. al. [7] study the factors influencing the ad clicks by searchers. They report that searchers have a bias against sponsored links (ad results) as compared to non-sponsored links (organic results).

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. Personally identifying information was removed from this data set. The data includes a sample of roughly 100 million search impressions, where an impression is defined as a single search result page. Queries are assumed to be in the English language. We removed any extra space at the beginning and end of the queries, and between words of the queries for both the impression and the clickthrough files. All queries were case-normalized. We found about 27 million queries occurring only once in the impression file, mostly with no ads. Such queries were removed from the impression data. 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 partitioned the impression and the clickthrough data into three equal-sized sets (i.e. training, test, validation) at a query level, so that each query appears in only one set. The training set is used to train the classifier, the test set is used to study the characteristics of different query intents. Finally, the validation set is used to estimate the number of ads for queries. All the three sets contain approximately

800K queries. We found many queries with very small number of ad clicks. Similar to Richardson et al. [10], since our analysis deals with empirical ad clickthrough of queries, it may be wildly different from the true clickthrough rate for queries with few number of ads, leading to noise in the three processes. Hence, we further filtered the three sets to include only those queries that have at least four ad clicks. After the filtering, we ended up with approximately 45K unique queries in each set (135K queries in total).

4 Classifying Query Intent

There are two dimensions of query intent studied in this work: commercial/ noncommercial and navigational/ informational. There are three types of features used at this stage of the work: i) query based features, ii) content of search result pages, and iii) ad clickthrough features. The intuition behind using clickthrough and query based features, like other implicit feedback techniques, is to represent user behavior and preferences in terms of the clickthrough behavior for different queries. In addition, we would hypothesize that the content of search engine result pages (displayed to the user) can be representative of the nature of a query. For instance, if a query like “shoe sale” is entered by user, the appearance of keywords like “buy”, “free”, and “shipping” in the result page may be good representatives of the commercial nature of such a query.

The features have been extracted for all query sets (training, test, and validation). The clickthrough features have been extracted according to the impression and clickthrough data recorded for each query. The query-specific features have been extracted from the query strings and also from the content of search engine result pages returned for them. We submitted each query to the Live search engine [1] and downloaded the first search engine result page (SERP) for that query. Each SERP is then represented as an unordered multi-set of term frequency ratios (a “bag of words”). The terms are extracted from the organic results only. Ad text is not included, avoiding any possible distortion that ad keywords might produce in the classification.

4.1 Classification Set-Up

We used Matlab toolbox for SVM and kernel methods [3] to classify queries in two dimensions: commercial/noncommercial and navigational/informational. In order to train and evaluate the classifier, 1700 queries (among the 45K queries in the training set) have been selected for manual classification. 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), 1700 queries were selected for which: i) the query was contained in training data, and ii) the ad click frequency of the query was greater than or equal to 11. Each selected query was then manually labeled as commercial/ noncommercial and navigational/ informational by three independent annotators.

The annotators were responsible for judging the presumed intent of the search queries from the perspective of a general user. If the presumed purpose of submitting a query is to make an immediate or future purchase of a product or

[1] http://www.live.com

service, the query is labeled as "commercial". Otherwise, it is labeled as "noncommercial". On the other hand, if the presumed purpose of a query is to locate a specific Web site or page, the query is labeled as "navigational". Everything else is considered "informational". There were 81% and 87% agreements (i.e. queries for which all annotators assigned the same label) among the annotators in commercial/ noncommercial and navigational/ informational labeling respectively. The final label of each query has been assigned based on the majority agreement among the annotators. As a result of the labeling, 42% of the queries were labeled as *commercial* and 58% were labeled as *noncommercial*, while 60% of the queries were labeled as *navigational* and 40% were labeled as *informational*.

4.2 Performance Measures

The prediction accuracy of each classifier, using 10-fold cross validation, is presented in Table 1. In both dimensions, including ad clickthrough features results in better accuracy. These seem more effective in the commercial/noncommercial classifier compared to the navigational/informational classifier, possibly due to the nature of ad clickthrough feature in effectively detecting commercial/ noncommercial intention. However, due to possibility of distorting the results, the ad clickthrough features are taken out and the classifiers based only on query and SERP features are used for the purpose of query intent detection. The results of such an intent detection will be used in the rest of the paper.

Table 1. Prediction Accuracy using SVM and 10-Fold Cross Validation

Features	Query Intent	Precision	Recall	Accuracy
SERP + Query + Clickthrough	Commercial	0.90	0.83	90%
	Noncommercial	0.89	0.94	
SERP + Query	Commercial	0.85	0.80	85.5%
	Noncommercial	0.86	0.90	
SERP + Query + Clickthrough	Navigational	0.86	0.87	84.5%
	Informational	0.81	0.80	
SERP + Query	Navigational	0.83	0.84	83.7%
	Informational	0.79	0.81	

5 Characteristics of Different Query Intents

In this section, we study how the average clickthrough rate for a particular number of ads varies for different query intents. We first calculate the average clickthrough rate for queries without considering the intention underlying them. Then, the same calculation is performed with respect to particular query intents. Finally, the impact of navigational and informational intent is studied on commercial intent by calculating the average clickthrough rate for commercial-navigational and commercial-informational intents.

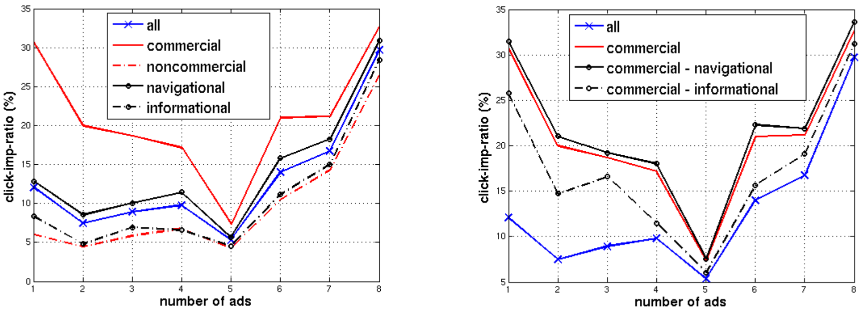
5.1 Click to Impression Ratio: General Case

In order to study the general case (where the intent is not considered), the impressions corresponding to the queries of the test set, 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, such 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. The unique id number for each impression (the impression id) is used 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 resulting from the impressions in each group. Let $id_i^j \in A_i$ denote the unique 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 id_i^j in the clickthrough data, and $c_i^j = 0$ otherwise. Hence, the average number of ad clicks per impression (clickthrough rate), CTR_i, for queries with a particular number of ads i is obtained as follows:

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

5.2 Click to Impression Ratio: The Impact of Query Intent

A similar approach can be followed here, however this time, only queries of a particular type are considered. The average clickthrough rates for the four general types of queries (i.e. commercial, noncommercial, navigational, and informational) with the corresponding number of ads are plotted in Figure 1-a. For clarity of presentation, we connect the points for each particular number of ads, but the lines do not imply interpolation. The same analysis is performed for two pairs of query types (either commercial-navigational or commercial-informational), resulting in the two plots depicted in Figure 1-b. The plot for the general case is also placed in Figures 1-a and 1-b to provide a baseline for comparison.



(a) **(b)**

Fig. 1. Average Click to Impression Ratio for Impressions with Particular Number of Ads (lines do not imply interpolation)

According to Figure 1-a, the commercial query type is the leader in terms of click ratio for all numbers of displayed ads. This ratio is larger (Figure 1-b) when the commercial query is also navigational rather than informational. There are some peaks and valleys in the plots of both figures that could result from the location of different ads (top or side of the result page) for which the clicks are recorded. According to Jansen [6], top-listed ads are assumed to be more relevant than organic results and side-listed ads. This could impact the frequency of clicks for ads at different locations and could be the cause of bumps at some points of the plots (especially the dip at 5). The locations of ads (top or side) are not available to us, but their impact should be the subject of further study. As depicted in Figure 1-a, navigational queries receive more ad clicks than informational queries on average. Similarly, Figure 1-b shows that commercial-navigational queries receive more ad clicks than commercial-informational queries on average.

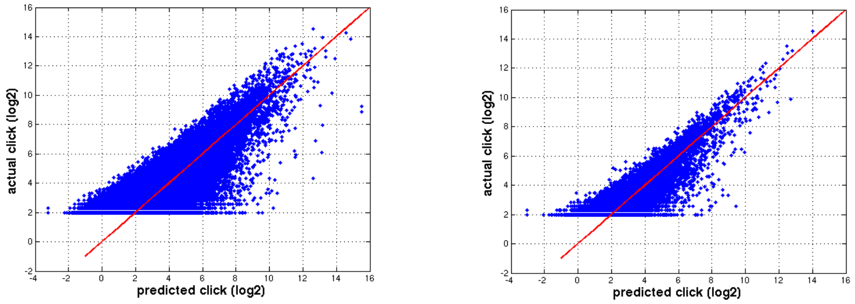
6 Estimating Number of Ad Clicks for Queries

The average clickthrough rate for each particular type of query intent can be used towards calculating the number of ad clicks for previously unseen queries. This is possibly due to the observation that the number of displayed ads actually affects the number of ad clicks for each category of query intent differently. In this regard, using the average clickthrough rate previously obtained from the *test* set, the number of ad clicks for queries in the *validation* set is estimated in this section, where: i) the ratio values for queries in general are used, or ii) the ratio values associated with particular number of ads and query intent are used.

In the general case, the number of ad clicks for a given query q can be estimated based on the number of ads displayed for q (thus, the average clickthrough rate corresponding to that ad#) and the number of unique impressions in which the query appears. The average clickthrough rate in this case is the one obtained according to Equation 1, where the intentions underlying queries are not considered. Hence, for the query q in the validation set, let imp_q^i denote the number of times query q appears in the impressions with i number of ads. In Equation 1, we estimated the average ad clickthrough rate for such a query as CTR_i. Thus, the estimated number of ad clicks for such a query is calculated as follows:

$$click_q = \sum_{i=1}^{8} CTR_i \times imp_q^i \qquad (2)$$

Figure 2-a depicts the actual number of clicks versus the estimated number of clicks in log-log basis and for all queries in general.



(a) all queries in general **(b)** commercial queries

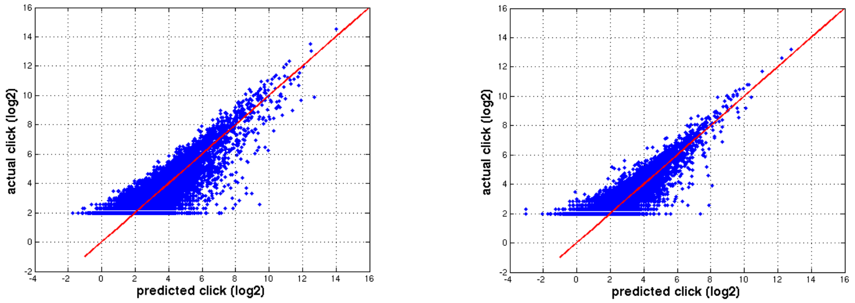
Fig. 2. The Actual Number of Clicks vs. the Estimated Number of Clicks: All Queries in General vs. Commercial Intent

Let CTR_i^{cn} and CTR_i^{ci} be the average clickthrough rates for impressions with i ads that belong to commercial-navigational and commercial-informational queries respectively. Similarly, let CTR_i^{nn} and CTR_i^{ni} be the average clickthrough rates for impressions with i number of ads that belong to noncommercial-navigational and noncommercial-informational queries respectively. For a given query $q \in Q$, where Q is the set of all queries, we define function t as $t : Q \to T$. T is the set of pairs of query intents: commercial-navigational (cn), commercial-informational (ci), noncommercial-navigational (nn), and noncommercial-informational (ni). According to our previous notation, we consider $T = \{\text{cn, ci, nn, ni}\}$. Based on what we just defined, the proposed prediction strategy obtains the query intents using the function t. It then uses the average rate corresponding to that category of intents in order to calculate the estimated number of clicks by going through all the impressions (similar to the Equation 2) of the query:

$$click_q^{int} = \sum_{i=1}^{8} CTR_i^{t(q)} \times imp_q^i \tag{3}$$

where $click_q^{int}$ is the estimated number of clicks based on the proposed prediction model which considers the average clickthrough rate for different query intents.

The plots for the commercial queries are presented in Figure 2-b. As is shown in the figure, the predicted number of clicks and the actual number of clicks are more correlated than the baseline depicted in Figure 2-a. We measure KL-divergence [10] between our model (predicted number of ads) and the actual number of ads on the validation set, which can be seen as the amount of information needed to encode the number of ad clicks for a new arriving query using the prediction model. Note that a perfect model would score 0. KL-divergence for the plot in Figure 2-a is calculated as 0.69, while the one for commercial queries (Figure 2-b) is 0.29. This may indicate that the number of ads represents a major factor in determining the number of clicks for commercial queries.



(a) commercial-navigational queries **(b)** commercial-informational queries

Fig. 3. The Actual Number of Clicks vs. the Estimated Number of Clicks: Impact of Navigational/Informational Intent on the Commercial Intent

To further study the effectiveness of the number of ads in such an intention-based prediction, we plotted the actual number of clicks versus the predicted clicks for commercial-navigational and commercial-informational queries in Figures 3-a and 3-b respectively. It is worth mentioning that the number of queries plotted on these two graphs are nearly the same (about 9K each), however the one for commercial-informational seems more correlated than the other. Moreover, KL-divergence measure is higher for the commercial-navigational one compared to the commercial-informational (0.33 versus 0.19). This could indicate that the number of ads determine the number of ad clicks for commercial-informational queries more effectively than queries that are commercial-navigational.

7 Conclusions and Future Directions

In this paper, we develop a methodology to use the combination of ads clickthrough and query features with the content of search result page in order to determine the intention underlying queries. The findings of our study suggest that ad clickthrough features improve the accuracy of detecting different query intents comparing to the case that only query and SERP features are used. The average ad clickthrough rate is then estimated for different intents. Our findings show that the number of displayed ads affects the number of ad clicks for each category of query intent differently, and therefore it might be usable as a means to study the characteristics of different query intents. The obtained clickthrough rates have been used to estimate the number of ad clicks for previously unseen queries with particular intentions and various number of ads (one to eight) displayed as the result of each query.

8 Acknowledgments

We would like to thank Microsoft Research and Microsoft adCenter for the Beyond Search data sets, and for partially supporting this work. We would also like to thank Bradly Lushman and Mike Patterson for their help with this work.

References

1. A. Ashkan, C. Clarke, E. Agichtein, and Q. Guo. Characterizing query intent from sponsored search clickthrough data. *Proceedings of the SIGIR Workshop on Informational Retrieval for Advertising*, pages 15–22, 2008.
2. A. Broder. A taxonomy of Web search. *ACM SIGIR Forum*, 36(2):3–10, 2002.
3. S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy. SVM and kernel methods matlab toolbox. Perception Systmes et Information, INSA de Rouen, Rouen, France, 2005.
4. 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 World Wide Web*, pages 829–837, 2006.
5. K. Debmbsczynski, W. Kotlowski, and D. Weiss. Predicting ads clickthrough rate with decision rules. *Workshop on Target and Ranking for Online Advertising*, WWW 2008.
6. B. Jansen. The comparative effectiveness of sponsored and nonsponsored links for Web e-commerce queries. *ACM Transactions on the Web*, 1(1), 2007.
7. 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.
8. U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in Web search. *Proceedings of the* 14^{th} *International Conference on World Wide Web*, pages 391–400, 2005.
9. M. Regelson and D. Fain. Predicting clickthrough rate using keyword clusters. *Proceedings of the* 2^{nd} *Workshop on Sponsored Search Auctions*, 2006.
10. M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the clickthrough rate for new ads. *Proceedings of the* 16^{th} *International Conference on World Wide Web*, pages 521–530, 2007.