Characterizing User Search Intent and Behavior for Click Analysis in Sponsored Search

by

Azin Ashkan

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2013

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Azin Ashkan

Abstract

Interpreting user actions to better understand their needs provides an important tool for improving information access services. In the context of organic Web search, considerable effort has been made to model user behavior and infer query intent, with the goal of improving the overall user experience. Much less work has been done in the area of sponsored search, i.e., with respect to the advertisement links (ads) displayed on search result pages by many commercial search engines. This thesis develops and evaluates new models and methods required to interpret user browsing and click behavior and understand query intent in this very different context.

The concern of the initial part of the thesis is on extending the query categories for commercial search and on inferring query intent, with a focus on two major tasks: i) enriching queries with contextual information obtained from search result pages returned for these queries, and ii) developing relatively simple methods for the reliable labeling of training data via crowdsourcing. A central idea of this thesis work is to study the impact of contextual factors (including query intent, ad placement, and page structure) on user behavior. Later, this information is incorporated into probabilistic models to evaluate the quality of advertisement links within the context that they are displayed in their history of appearance. In order to account for these factors, a number of query and location biases are proposed and formulated into a group of browsing and click models.

To explore user intent and behavior and to evaluate the performance of the proposed models and methods, logs of query and click information provided for research purposes are used. Overall, query intent is found to have substantial impact on predictions of user click behavior in sponsored search. Predictions are further improved by considering ads in the context of the other ads displayed on a result page. The parameters of the browsing and click models are learned using an expectation maximization technique applied to click signals recorded in the logs. The initial motivation of the user to browse the ad list and their browsing persistence are found to be related to query intent and browsing/click behavior. Accommodating these biases along with the location bias in user models appear as effective contextual signals, improving the performance of the existing models.

Acknowledgements

First of all, I would like to thank my advisor, Charles Clarke, from who I learned a lot throughout these years. His expertise, unique insights, continuous support, and positive attitude have been extremely helpful to me. I am forever indebted to Charlie for his friendship and for providing by far the best research collaboration and experience of my career.

I would like to express my sincere gratitude to my examining committee, Charles Clarke, Gordon Cormack, Rosie Jones, Mark Smucker, and Olga Vechtomova for taking the time and the effort to participate in my committee and for their constructive feedback.

My sincere gratitude goes to Eugene Agichtein and Qi Guo for their valuable comments and collaborations in part of this work. I would also like to thank Gordon Cormack for his kind and encouraging attitude towards me, and for the insightful discussions.

Many thanks to my great colleagues in the IR / PLG group for providing such a friendly and wonderful environment in the lab. In particular, I would like to thank John Akinyemi, Jun Chen, Adriel Dean-Hall, Ashif Harji, Maheedhar Kolla, Bradley Lushman, Nomair Naeem, Adam Roegiest, Bahareh Sarrafzadeh, and Luchen Tan. I will never forget the fun we had together during our game nights, dinner events, and curling matches.

I hereby want to thank my beloved family for their unconditional love and care. In particular, my love and gratitude go to my Mom for always supporting me and helping me towards my dreams. Her insightful, caring, and strong personality has always been my inspiration in life. My love and respect go to my Dad for always believing in me and for passionately introducing me to the world of Mathematics and Physics. Many thanks to my friend to the spirit, Aida, for always being there for me.

My deepest appreciation goes to the love of my life, my husband, Ali. The best and sweetest part of this journey was having you beside me at all time. Our love, respect, and understanding for each other are the precious things I cherish the most in life.

At the end, I would like to acknowledge the Natural Sciences and Engineering Research Council of Canada for the NSERC Postgraduate Scholarship, the University of Waterloo for the President's Graduate Scholarship, and the Cheriton School of Computer Science for the David R. Cheriton Graduate Scholarship.

To my beloved Ali!

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Traditional information retrieval methods do not explicitly consider goals, interests, and preferences of the user (Belkin, 1993). Accommodating these factors requires us to understand the intent of the user when issuing a query. In order to implement this requirement, user behavior must be captured and modeled, allowing us to make meaningful inferences regarding intent. These inferences, in turn, allow us to improve the user experience and satisfying their information need.

This thesis is specifically concerned with *sponsored search*: the selection and display of advertisement links (ads) by a commercial search engine (e.g., Bing or Google) in response to a user query. To explore and model user behavior, we work with a log of queries, result pages, and clicks provided by a commercial search engine for research purposes. Although this log is heavily anonymized, with substantial redaction, the impact of search context is demonstrated on user behavior. As part of this context, the nature of the query, the content of the organic (i.e., non-sponsored) results, the page structure, and the positioning of ads on the page are considered. By considering this context, improved models for user behavior and new methods for measuring ad quality are presented.

1.1 Background and Problem Statement

To study user intent, two major approaches have been employed in prior work. The first focuses on the manual examination of user queries to establish major categories of user intent (Bodoff, 2006; Carmel *et al.*, 1992; Choo *et al.*, 1998; Morrison *et al.*, 2001; Navarro-Prieto *et al.*, 1999; O'Day and Jeffries, 1993; Sellen *et al.*, 2002; Teevan *et al.*, 2004). The second applies machine learning methods to automatically classify user intent (Ashkan and Clarke, 2013; Ashkan *et al.*, 2009a; Baeza-Yates *et al.*, 2006; Beitzel *et al.*, 2005; Jansen *et al.*, 2008; Lee *et al.*, 2005; Li *et al.*, 2008; Nettleton *et al.*, 2007; Tan and Peng, 2008). Most early work was concerned only with user intent as it related to the organic search results (Broder, 2002; Kang and Kim, 2003; Rose and Levinson, 2004). More recent work has considered sponsored search, particularly the problem of predicting commercial intent (Ashkan *et al.*, 2008; Dai *et al.*, 2006; Hu *et al.*, 2008), i.e., the intent to purchase a product or service.

According to Rose and Levinson (2004), studying user goals in information access involves three primary tasks: i) creating a conceptual framework for user goals, ii) finding a way to associate user goals with queries, and iii) modifying engines to exploit user preferences via the goal information. The problem of inferring user search intent, as well its application to the improvement of Web search experiences, can be understood with respect to these three tasks. In what follows, the discussion is organized around these tasks.

The first task can be seen as *understanding* and *modeling* user search intent. With respect to this task, one direction of study could aim at extending the traditional categories of query intent by considering different domains of Web search and by the manual examination of query/click logs.

Traditionally, user intent corresponded to any of the standard categories of Web query defined by Broder (2002): *navigational*, *informational*, and *transactional*. A navigational query is defined as a query through which the user has an immediate intent to reach a particular Webpage. The user intent behind an informational query is to acquire some information assumed to be present on one or more Webpages. The intent underlying a transactional query is to perform some Web-mediated activity, such as buying a product or playing a game.

In the context of 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* (OCI) (Dai *et al.*, 2006). In this respect, this thesis studies and extends the traditional categories of Web queries in the commercial domain, i.e sponsored search.

The second task, that of *inferring* user intent, involves using the information obtained from any available source, such as a user search history, for the purpose of identifying the intention behind a query. A natural source of information for this purpose is the query itself. However, the query is short and it may not reveal much about the user (Broder *et al.*, 2007). Therefore, additional sources of information may be used to enrich the query.

Earlier work mostly infers intent by considering the anchor text associated with the links clicked by the user and the content of the pages reached from these links. This work approaches query intent detection using the information obtained from the combination of query itself and the relation between the query string and URLs presented on the page that the search engine returns as a result of the query. In addition, search result pages returned for queries are considered as representatives of the nature of the queries, and therefore the caption (Clarke *et al.*, 2007) of search results displayed on these pages are used as means to identify the intent underlying queries.

The third task concerns the application of user intent studies to improve Web search. Understanding and inferring the intent underlying user queries can help search providers improve search personalization, and ideally increase user satisfaction. There has been a growing interest in employing query intent analysis to improve various Web search experiences, while there have been less efforts made in the sponsored search domain. This thesis considers the use of the contextual information, particularly the query intent, for the purpose of click analysis in the online advertising domain.

Sponsored search provides the major source of revenue for commercial search engines. In its simplest form, for a given query, a set of candidate advertisement links (a.k.a. ad(s)) are obtained through an ad retrieval algorithm that matches the query terms with the ads' bid terms — keywords that are selected by the advertisers (Fain and Pedersen, 2006). These ads are then ranked in decreasing order based on the expected revenue estimated for each individual ad, a, typically as: $REV(a) = b_a \times Q_a$ where b_a is the bid placed by an advertiser, to be paid once the ad's target page receives a click, and Q_a is the quality score of the ad as a notion of its performance predicted by the search engine.

A click on an ad is viewed as a potential purchase opportunity for the product or service that is offered by the advertiser. Hence, the expected ad clickthrough rate (CTR) is the classic measure of ad quality, determining the expected revenue for the search engine.

Information obtained from implicit feedback resources, such as query logs and the click history of ads recorded in the logs (Richardson, 2008), have been widely used to interpret and predict user's future click behavior over these links. The existing models adopt the common assumption of a *trust bias* (Joachims *et al.*, 2005) with respect to displayed results, which is the basis of the *examination model* (Richardson *et al.*, 2007). Under this model, it is assumed that click rate decreases towards the lower ranks on result pages due to the reduced visual attention from the user.

The link examination behavior of users appears to be similar to their click behavior, where most of the links examined by users are found to be the top-listed results on search result pages for both sponsored and non-sponsored links (Jansen *et al.*, 2007). Hence, the main factors determining an ad click under the examination model are the relevance of each individual ad to the user's need and the rank position of the ad on the page. This thesis, in part, investigates the result page contextual factors, such as the number of ads and the intent underlying the queries, as factors that could impact the ad's click.

Further efforts in click models consider the influence of the co-appearance of ads with each other; a relatively strong ad appearing prior to a particular ad on a page may distract the user from the lower ads, while a weakly related ad appearing at a relatively high position may annoy the user into abandoning the list, regardless of the quality of the later ads. This idea is introduced by Ghosh and Mahdian (2008) as the *externality effects* of ads in advertising auctions, and it is based on the assumption that users visually scan the ad list from top to bottom. Once an ad is examined by the user, ad-specific factors will determine the click decision and continuation probability. In contrast to the examination model, the probability of clicking on an ad is considered to be dependent on the other ads shown above it on the page as well as on its own quality factors.

The browsing and click behavior of users over the advertisement links are targeted in this thesis where the linear browsing behavior of users, the impact of their queries, and the structure of the result pages are among the factors that are taken into account.

1.2 Research Approach

This research work explores query intent characterization and user browsing and click behavior within the sponsored search domain. The contributions of the thesis can be summarized as follows:

- Extending the traditional categories of Web queries.
- Enriching queries with contextual information obtained from search result pages returned for these queries.
- Exploring simple methods for the reliable labeling of training data.
- Studying characteristics of commercial intent in sponsored search domain.
- Employing such contextual information as the query intent for click analysis in sponsored search.
- Exploring user browsing behavior in sponsored search, eventually using the information for the better understanding of the click behavior in this domain.
- Analyzing patterns and common signals found in user browsing and click behavior with respect to various context.

We aim at extending the traditional categories of Web queries for commercial search. In the initial steps towards this goal, queries that exhibit commercial/ noncommercial and navigational/ informational intent are studied. The commercial intent category is further extended to reference specific products, brands and retailers (Ashkan and Clarke, 2009), avoiding the more traditional topical categories, such as sport and news.

This work tackles the query intent inference using mostly the context in which a query appears. With respect to this approach, the content of search engine result pages (SERPs) returned for queries are considered as representatives of the nature of the queries, and they are used as means to identify the intent underlying queries.

A relatively easy and reliable approach is presented to label a set of queries through crowdsourcing (Doan *et al.*, 2011) in order to be used for training purposes. The queries are consistently labeled by assessors in different dimensions, and the labels are fed to classifiers in order to identify different categories of query intent for a larger set of queries.

Feature sets are used to train five binary classifiers to recognize different dimensions of the commercial intent: i) commercial/ noncommercial, ii) navigational/ informational, iii) product oriented, iv) brand oriented, and v) retailer oriented. Classification accuracies confirm that distinctions among these categories are reasonably distinguishable, and also the query based features, along with the content of search engine result pages, are effective in detecting such categories.

Throughout this thesis, it is demonstrated that click behavior is consistent with the query intent categories identified and inferred earlier. While ad click analysis is primarily addressed in the literature by considering the *content* of ads and their associated bid terms, this work aims at tackling the problem through the *context* in which an ad appears. Consider an ad as it appears in the context of a search engine result page. This context may strongly influence whether or not the user will click on the ad, and includes the display location of the ad, the rank of the ad, the user's query, the organic search results, and other ads displayed along with it.

A SERP represents a result page that the search engine returns as the result of a query. When a user clicks on an ad displayed on a SERP, they are transferred to the ad's landing page, and the context and click information are typically stored in a log by the search engine. An ad *impression* represents a pair of ad and SERP where the ad appears on the

SERP as a result of a query. An ad's impression number refers to the number of times the ad is displayed in its history of appearance in the log, whether it is clicked on or not. Commercial search engines generate and store this information as part of their normal operations. Historical clickthrough rate (CTR) for an ad can be computed directly from this log information as the ratio of the total number of clicks recorded for the ad to the ad's impression number. One of the purposes of this thesis is to use the click log information to study the relationship between user ad click behavior and characteristics of queries and their corresponding sponsored links shown on search result pages.

In order to validate characteristics and differences of the query categories, the sponsored search domain is employed due to two reasons: i) this domain has not been much addressed in previous query intent studies, and ii) this domain is more heavily targeted by the commercial intent of users. This step can also be seen among the primary efforts to address the third task introduced by Rose and Levinson (2004) as indicated previously; search engines can utilize the information obtained from user goals and preferences in order to improve the existing algorithms in Web search and information access.

Models for click behavior are presented in order to estimate ad clickthrough rate (CTR) with respect to the history of ads while the contextual information about the search result pages, such as the query intent category, forms the basis of these models. The aggregated click behavior of search result pages (i.e. context CTR), with respect to the number of displayed ads and rank position of ads, forms the basis of the proposed *baseline model*. In order to further study the influence of the query intent on ad CTR, the previously determined intents underlying the users' queries are used to extend the baseline model, forming the proposed *query intent model*. These context-based ad click analyses suggest that contextual factors such as the intent behind user query have correlations with the performance of ads in sponsored search. Comparing the baseline model against the query intent model suggests that ad clickthrough prediction techniques could benefit from the query intent information and other contextual factors.

The above models consider only the aggregated click probability of an individual ad, whereas the list of ads appearing prior to a particular ad on the page and the user's browsing behavior do not have any influence on the prediction. They are also limited with respect to their need for explicit judgements of query intent. In a further attempt, a group of cascade-based click models, called *location– and query–aware models*, are studied that aim at using variability of user behavior and differences found between user browsing behavior in dealing with sponsored links and organic links to better model click behavior in sponsored search. The main goal of these models is to gain insight into user browsing and click behavior in sponsored search, which in turn improves one's ability to infer the probability of clicks on the advertisement links.

We augment context by adding biases that are based upon known properties of sponsored search, and upon the differences that exist in user browsing behavior over advertisement links as opposed to organic links. According to Jansen and Resnick (2006), users exhibit a strong bias against ads, as opposed to organic links. Thus, users have a stronger tendency to consider organic links rather than ads. As for the ads themselves, users are known to pay more attention to the top-listed ads (i.e., those appearing above the organic results) as opposed to the side-listed ads (i.e. those appearing to the right of the organic results) (Jansen and Resnick, 2006). The nature of the query may also influence the probability that the user will initiate browsing of the ad list, and continue browsing the ad list once they start. A user who issues a commercially oriented query may be assumed to have a greater tendency to purchase or utilize a commercial service, and thus they click on ads more frequently.

A notion of *location bias* is formally modeled in this work in order to account for top-listed and side-listed ads separately. Furthermore, *query biases* are introduced and parameterized in order to account for the probability that the user will initiate browsing of the ad list, and for their persistence (patience) in continuing to browse through the list. The *initiation* probability with respect to a particular query is defined as the chance that the user who issues this query will eventually initiate browsing the ad list. The *persistence* or *transition* probability with respect to a query is defined as the chance that the corresponding user who examined a particular ad at rank i will continue on to examine the ad at rank $i+1$. Both the initiation probability and the transition probability are determined separately for different display locations of ads on result pages.

The parameters of the location- and query- aware models and also their original settings are learned from the click signals recorded for the advertisement links appearing in a log of search result pages. To evaluate the performance of the models and to compare them with state-of-the-art performance, standard evaluation metrics, including log-likelihood and perplexity are applied. The evaluation results indicate that, through the incorporation of query and location biases, significant improvements can be achieved in predicting browsing and click behavior in sponsored search.

In part of the experimental study, the extent to which these biases actually reflect varying behavioral patterns is explored, confirming that correlations exist between the biases and user search behavior. These sorts of observations are interesting and at the same time may be helpful in the sense that they are obtained from independent experiments. One group of experiments empirically calculates the bias parameters with respect to the click signals recorded in the search engine log. Another group of experiments matches these values against query types that have been obtained independently, and finds distinctive patterns of user behavior based on them. These patterns not only shed light on user behavior, they may also suggest the development of user dependent properties to be used as signals for ad click analysis in sponsored search.

1.3 Data Set

The empirical study described throughout this thesis work is based on a data set obtained from the Beyond Search program (Beyond Search Data, 2007) developed by Microsoft Research and Microsoft adCenter in 2007. This set of ad search and click logs was sampled over the course of three months with personally identifying information removed. It includes a sample of 101 million SERPs. A set of ad clicks (about 8 million) is also associated with the SERPs.

Each SERP and each click is described by a set of attributes. The list of attributes used from the SERP data is as follows: *date and time of when the SERP was displayed*, *user query*, *number of ads displayed on the SERP*, *the list of ad ids displayed on the SERP*,

Table 1.1: Statistics of the data

	Original	Setting #1 (Click≥4)	Setting #2 (Frequency≥3, Click≥1)
Unique SERPs	98M	20M	41M
Unique ad impressions	360M	107M	206M
Unique queries	25M	135K	2.8M

user session ID, and *SERP ID*. The list of attributes used from the click data is as follows: *date and time of the click*, *user query*, *rank of the clicked ad on the corresponding SERP*, *user session ID*, and *corresponding SERP ID*.

We removed any extra white space at the beginning and the end of the queries, and between words of the queries for both SERP and click files. We then case-normalized the queries. SERPs with a duplicate combination of SERP id and user session id were removed in order to filter out repeated queries from the same user, resulting in approximately 98 million unique SERPs. We used the above attributes to create the set of ad impressions (about 360 million) from the SERPs. Remember an impression is represented by the pairs of ad and SERP where the ad appears on the SERP as a result of a query. Each impression obtains its unique id as the combination of <*user session ID*, *SERP ID*, *ad id* >, and it records all the information from the corresponding SERP. Each click instance was then associated with the appropriate ad impression in order to track the click status of each impression. A report of the statistics of the data is presented in Table 1.1.

There are various experiments conducted in this thesis in order to provide evidence for the proposed hypotheses or to empirically evaluate the performance of the proposed models and techniques. For these purposes, various subsets of the original data are required, which are prepared through different filtering settings. We categorize these subsets under two settings that will be explained next.

1.3.1 Setting #1

The first setting prepares subsets of the original data which will be used by those experiments exploring query intent categories. In order to prevent train-test contamination across those experiments, we randomly split the impression and click data into three equal-sized sets (set $A^{(1)}$, set $B^{(1)}$, and set $C^{(1)}$) at the query level. All the impressions and click data for a given query are put into the same set. This process is achieved by randomly assigning each query (with all its impression and click information) into one of the three sets.

Set $A^{(1)}$ is used to train classifiers for query intent detection; set $B^{(1)}$ is used to study characteristics of different categories of query intent; set $C^{(1)}$ is used to study ad click-through with respect to various contexts of result pages. All three sets contain approximately the same number of queries (about 800K) along with their SERP, impression, and click information.

There are many queries with a very small number of clicks (less than four). The upcoming analyses deal with empirical ad click rates, and estimating click rates from small numbers of clicks may lead to estimates that are wildly different from the true click rates for these queries, producing problems with the analysis. Following the approach of Richardson *et al.* (2007), the sets have been filtered to include only queries that have at least four ad clicks. After filtering, we ended up with 45032, 44941 and 44909 queries in sets $A^{(1)}$, $B^{(1)}$, and $C^{(1)}$ respectively ($134,882$ queries in total).

In the remainder of the thesis, we will refer to the case/space/user normalized SERP data, impression data and their corresponding click information as the **original** data. Otherwise, by **SERP** data, **impression** data or **click** data from the **setting #1**, we mean one of the prepared sets (i.e. $A^{(1)}$, $B^{(1)}$, or $C^{(1)}$) as described above.

1.3.2 Setting #2

The second setting prepares data for those experiments that involve modeling user browsing and click behavior in sponsored search. Two subsets of the data are prepared for this purpose.

The first subset is referred to as set $A^{(2)}$ in the remaining of the thesis. There are approximately 206 million impressions (resulted from 41M SERPs) in this set which were sampled over three months. These impressions correspond to a sample of about 2.8M queries, all of which were required to appear at least 3 times and to have at least one ad click recorded in the log.

For another group of experiments, we will require access to the location of the ads on the SERPs. Since the precise location of the ads is not recorded in the original data set, we had to separately select a subset of the data for the second step such that the location of ads can be identified with certainty. This selection is possible due to the ad placement strategy of the search engine at the time, i.e. for each SERP, at most three ads were displayed on the top and a maximum of five ads were displayed at the side.

Given this constraint, and assuming that the ads displayed on the top are ranked higher than the ones displayed at the side, one can be assured that SERPs with eight ads belong to result pages with three top-listed ads and five side-listed ads. All such SERPs along with their impression and click information from the set $A^{(2)}$ have therefore been placed into a second set, called $B^{(2)}$. Approximately 56 million impressions (resulted from 7M SERPs) appear in the set $B^{(2)}$, which corresponds to a sample of about 712K queries.

For both sets $A^{(2)}$ and $B^{(2)}$ from the **setting #2**, we make sure that their **SERP** data is sorted according to their time stamp. This is required due to the final set of the experiments that focuses on the online learning of ad click probabilities, and therefore the order in which the SERPs appear in the log matters.

1.4 Bibliographical Notes

Preliminary versions of the material presented in various parts of this thesis have appeared in the following publications:

- Azin Ashkan and Charles L.A. Clarke, Impact of Query Intent and Search Context on Clickthrough Behavior in Sponsored Search, Knowledge and Information Systems (KAIS) Journal, Volume 34, Issue 2, pp. 425-452, February 2013.

- Azin Ashkan and Charles L.A. Clarke, Modeling Browsing Behavior for Click Analysis in Sponsored Search, In Proceedings of the 21^{st} ACM Conference on Information and Knowledge Management (CIKM), pp. 2015-2019, Maui, Hawaii, November 2012.

- Azin Ashkan and Charles L.A. Clarke, Characterizing Commercial Intent, In Proceedings of the 18^{th} ACM Conference on Information and Knowledge Management (CIKM), pp. 67-76, Hong Kong, China, November 2009.

- Azin Ashkan, Charles L.A. Clarke, Eugene Agichtein, and Qi Guo, Estimating Ad Clickthrough Rate through Query Intent Analysis, In Proceedings of the IEEE / WIC / ACM International Conference on Web Intelligence (WI), pp. 222-229, Milan, Italy, Sept. 2009.

- Azin Ashkan and Charles L.A. Clarke, Term-Based Commercial Intent Analysis, In Proceedings of the 32^{nd} ACM International Conference on Research and Development in Information Retrieval (SIGIR), pp. 800-801, Boston, USA, July 2009.

- Azin Ashkan, Charles L.A. Clarke, Eugene Agichtein, and Qi Guo, Classifying and Characterizing Query Intent, In Proceedings of the 31^{st} European Conference on Information Retrieval (ECIR), pp. 578-586, Toulouse, France, April 2009.

- Azin Ashkan, Charles L.A. Clarke, Eugene Agichtein, and Qi Guo, Characterizing Query Intent From Sponsored Search Clickthrough Data, In Proceedings of the SIGIR Workshop on Information Retrieval for Advertising (IRA), pp. 15-22, Singapore, July 2008.

Other related publications are as follows:

- Azin Ashkan and Charles L.A. Clarke, On the Informativeness of Cascade and Intent-Aware Effectiveness Measures, In Proceedings of the 20^{th} International Conference World Wide Web (WWW), pp. 407-416, Hyderabad, India, March-April 2011.

- Charles L.A. Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan, A Comparative Analysis of Cascade Measures for Novelty and Diversity, In Proceedings of the 4^{th} ACM International Conference on Web Search and Data Mining (WSDM), pp. 75-84, Hong Kong, China, February 2011.

- Azin Ashkan, John S. Whissell, and Charles L.A. Clarke, Unsupervised Learning of Result Page Context for Clickthrough Analysis in Sponsored Search, In Proceedings of the 5^{th} International Workshop on Data Mining and Audience Intelligence for Advertising (ADKDD), pp. 44-52, San Diego, USA, August 2011.

- John S. Whissell, Charles L.A. Clarke, and Azin Ashkan, Clustering Web Queries, In Proceedings of the 18^{th} ACM Conference on Information and Knowledge Management (CIKM), pp.899-908, Hong Kong, China, November 2009.

- Qi Guo, Eugene Agichtein, Charles L.A. Clarke, and Azin Ashkan, In the Mood to Click? Towards Inferring Receptiveness to Search Advertising, In Proceedings of the IEEE / WIC / ACM International Conference on Web Intelligence (WI), pp. 319-324, Milan, Italy, September 2009.

- Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon, Novelty and Diversity in Information Retrieval Evaluation, In Proceedings of the 31^{st} ACM International Conference on Research and Development in Information Retrieval (SIGIR), pp. 659-666, Singapore, July 2008.

1.5 Thesis Organization

The remainder of this thesis is organized as follows:

In Chapter 2, a background review of the query intent analysis is provided with a focus on traditional query intent and the commercial intent categories. As well, concepts from the sponsored search domain are briefly discussed in this chapter. Various clickthrough studies are reviewed where the main purpose is to address the lack of contextual information in this area. Towards the end of the chapter, an overview of the common cascade-based browsing and click behavior models is provided.

Details of query modeling techniques (manual annotation and crowdsourcing) are discussed in Chapter 3. The evaluation results of the annotation process are also reported. The remainder of the chapter discusses the details of the query inference procedure within the five dimensions of query intent studied in this work. A preliminary study of the query intent is explained first which accounts for the click information of the query terms in order to determine the intent of the query. It is concluded that click information is costly and noisy, and therefore the rest of chapter focuses on developing a features set from the result page contextual information. The chapter ends by reporting the results of the classification of the query intent using the collected set of features.

In Chapter 4, characteristics of search result pages are studied with respect to a group of factors: the position of ads, the location of ads, the number of displayed ads, and query intent. It is concluded that the placement of ads has a substantial impact on the number of clicks they receive. This impact is more obvious when the intent underlying the queries for which those ads are displayed is also taken into account. User's click behavior in sponsored search is modeled in the remaining of the chapter. The models are based on aggregated click behavior over the ads displayed within different contexts on the result pages.

Another group of click models, called location– and query–aware models, are studied in Chapter 5. Rather than taking the individual performance of ads into account (like the models studied in Chapter 4), this group of models considers each ad in the context of the ones appearing prior to it on a result page. They also account for variability of user behavior and differences that exist between user browsing behavior in dealing with sponsored links and organic links to better model click behavior in sponsored search. The evaluation results of the models are presented ate the end along with an overall comparison of the performance of all models.

Chapter 6 further analyzes the results of the previous chapter and reports interesting patterns found in user browsing and click behavior in sponsored search.

Finally, a summary of the contributions of this thesis, along with findings, and future directions are provided in Chapter 7.

Chapter 2

Background and Related Work

Implicit feedback has been widely used to interpret and predict search behavior and user preferences (Agichtein *et al.*, 2006; Joachims *et al.*, 2005; Oard and Kim, 2001; Oard *et al.*, 1998; Poblete and Baeza-Yates, 2008; Richardson, 2008; Teevan *et al.*, 2008), including user intent. As Broder *et al.* (2007) point out, queries are often short. A query by itself may not reveal much about the user, and implicit feedback can enhance this information. Compared with explicit feedback, such as a user questionnaire, implicit feedback has the advantage that it can be collected at much lower cost, in much larger quantities, and without placing an unwanted burden on the user (Joachims *et al.*, 2005). However, implicit feedback is more difficult to interpret than explicit feedback, is potentially noisy (Joachims *et al.*, 2005), and is generally thought to be less accurate (Kelly and Teevan, 2003).

2.1 Query Intent Studies

While the earliest Web search engines were primarily based on document content and link structure (Broder, 2002), they now depend heavily on an understanding of user behavior, including user intent. A synthesis of prior work on query intent analysis is depicted in Table 2.1. Prior efforts reported in literature can be categorized from two perspectives:

i) the method of inference (*manual* or *automatic*), and ii) the query intent categories addressed (traditional or commercial).

Table 2.1: Synthesis of prior work on query intent analysis.

	Manual	**Automatic**
Traditional	Mostly the traditional categories, such as navigational, informational, transactional, and resource, are addressed Much effort has been made in manual examination and survey review	Limited training data is used Low performance results or trade-off between precision and recall are reported Past click behavior and anchor text information are mostly used
Commercial	Limited study on the commercial categories has been conducted	High level query categories (commercial and noncommercial) are mostly addressed There is lack of study on the relation of this newer dimension of query categories with the traditional categories Characteristics of various query categories are not validated

Early work on query intent analysis includes a variety of controlled studies and manual examinations from various perspectives of user search, such as browsing (Carmel *et al.*, 1992; Teevan *et al.*, 2004), search strategies, tasks, tactics (Choo *et al.*, 1998; Morrison *et al.*, 2001; Navarro-Prieto *et al.*, 1999; O'Day and Jeffries, 1993), and information seeking (Bodoff, 2006; Sellen *et al.*, 2002). More recently, efforts have been made to tackle query intent modeling and detection through automatic approaches (Ashkan and Clarke, 2013; Ashkan *et al.*, 2009a; Baeza-Yates *et al.*, 2006; Beitzel *et al.*, 2005; Jansen *et al.*, 2008; Lee *et al.*, 2005; Li *et al.*, 2008; Nettleton *et al.*, 2007; Tan and Peng, 2008) mostly through the application of machine learning techniques. In what follows, we address a group of manual and automatic efforts in query intent analysis, for both the traditional and commercial categories of query intent.

2.1.1 Traditional Query Intent

A fundamental study of the traditional query categories was performed by Broder (2002), who proposed three broad user intent categories for Web queries: *navigational*, *informational*, and *transactional*. A navigational query is defined as a query through which the user has an immediate intent to reach a particular Webpage. The user intent behind an informational query is to acquire some information assumed to be present on one or more Webpages. The intent underlying a transactional query is to perform some Web-mediated activity. Using survey results, Broder reports that approximately 26% of queries are navigational, nearly 73% are informational, and an estimated 36% are transactional. Based solely on the log analysis, Border reports that 48% of the queries are informational, 20% navigational and 30% transactional.

Broder discusses the evolution of search engines with respect to these categories of query intent. In the early stage, engines mostly supported the informational queries, and they used to be heavily based on classic IR techniques over text and formatting of documents. Later, search engines targeted both informational and navigational queries by employing information other than content of documents, such as link analysis, anchor text, and clickthrough information. Finally, the emerging generation of search engines supports informational, navigational, and transactional queries and seeks answer to "the need behind the query" by employing various sources of information, such as semantic and context analyses.

Kang and Kim (2003) define query classification schemes using heuristics to distinguish two types of queries: topic finding (i.e. informational intent) and homepage finding (i.e. navigational intent). They placed the documents acting as entry points for a particular homepage into a database called DB_{HOME}, while the remaining documents from their collection were assigned to a database called DB_{TOPIC}. From these databases, they extracted various information, such as the frequency of appearance of an individual term and the mutual information of term pairs, and combined them linearly to formulate various query types. The best results obtained from their settings had a precision of 91.7%, while the recall was reported as 61.5%.

Rose and Levinson (2004) conducted an extensive study of user intent 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 (2002) might fall under either of their three categories, depending on the details of the desired transaction. They repeatedly revise these goal categories and suggest that the goals naturally fall into a hierarchical structure:

- The definition of a navigational goal by Rose and Levinson (2004) follows that of Broder's work (Broder, 2002). They consider a query to be navigational if the user has a single authoritative Webpage in mind when they issue the query. Hence, most queries consisting of names of companies, universities, or well-known organizations are considered to be navigational.

- Following Broder's definition, they express the informational goal as: i) a directed question to learn something in particular about a topic, ii) an undirected question to learn anything/everything about a topic, iii) a request for advice, ideas, suggestions, or instructions, iv) a request to locate something in the real world, or v) a request to obtain a list of suggestions for further research. Most queries consisting of topics in science, medicine, history, or news qualify as undirected informational queries.

- The resource category represents the goal of obtaining some resource other than information. This category can be expressed as: i) a goal to download a resource to be used on the computer or on other device, ii) an entertainment intent to be satisfied by viewing items available on the result page, iii) a goal to interact with a resource using another program/service available on the Website the user finds, or iv) a goal to obtain a resource that does not require a computer to use (e.g. printing a document).

They manually judged a sample of 500 queries with respect to the above categories and by looking at four sources of information: the query itself, the results returned by the search engine, the results on which the user clicked, and further searches or other actions (e.g. query refinement) performed by the user. Their observations indicate that roughly 62% of queries are informational while nearly 13% and 25% of the queries are navigational and resource, respectively. A large fraction of the queries appeared to be attempts to locate a product or service rather than to learn about a subject. With respect to these queries, over 35% appeared to have general research goals (i.e. questions and undirected requests for information) for which traditional information retrieval systems were designed.

The approaches described above primarily determine intent category from manual classification techniques, surveys, and other sources, using adhoc thresholds and parameters setting (Brenes *et al.*, 2009). One of the first attempts at automatic query intent classification was conducted by Lee *et al.* (2005) in order to predict user query goals in terms of navigational and informational intents. Their approach is based on past user click behavior and anchor-link distribution. They studied the 50 most popular queries issued to Google from the UCLA (University of California, Los Angeles) Computer Science Department. They computed the click distribution for these queries using click log data. Their findings suggest that the click distribution for a navigational query is highly skewed towards one or just a few domains, while the distribution is relatively flat for an informational query. Their approach correctly identifies the intent underlying 90% of these queries.

Liu *et al.* (2006) addressed the task of separating navigational queries from informational/ transactional ones. In part of their work, they studied the same features used by Lee *et al.* (2005): user-click behavior and anchor-link distribution. They extracted anchor text information from over 202 million crawled Chinese Webpages. Given a page and a query, if the page shares the same anchor text as the query, they consider it as an anchor match for the query. After filtering for spam, duplicate pages, and other noise, they calculated the percentage of queries matching a certain number of anchor texts. As a result of this study, they found that the percentage of matching queries does not vary with time, and that there are only 16.24% matching queries per day, on average. From a similar study, they found that as time goes by, the percentage of newly-appearing queries drops to about 10%, implying that click data can be applied to about 90% of queries. Based on these experiments, they conclude that the past user click behavior can be seen as prior information for about 90% of queries, while anchor text evidence may be effective for a much smaller percentage of queries. Hence, their work mainly focuses on using click data sources. With respect to these sources, they incorporate two features (extracted from click data) into a decision tree based algorithm for identifying the intent underlying user queries: the n Clicks Satisfied (nCS) and the Top n Results Satisfied (nRS).

The nCS feature is the proportion of sessions containing a given query in which the user clicked on at most n results. The underlying assumption for this feature is that users issuing navigational queries click on fewer results than users submitting informational or transactional queries. Hence, when using a small value for n (e.g. 2 clicks) navigational queries would exhibit larger nCS values than informational/ transactional queries.

The nRS feature is based on the assumption that users submitting navigational queries tend to click on the top results. Thus, nRS is the proportion of sessions containing a given query in which the user clicked on, at most, the top n results. Navigational queries exhibit higher nRS values (e.g. for $n = 5$).

Liu et al. compare their work with that of Lee *et al.* (2005). Their test set consists of 81 informational/transactional queries and 152 navigational queries. The informational/transactional queries were obtained from a Chinese search engine, while the navigational queries were obtained from a widely-used Chinese Web directory `hao123.com`. Their findings indicate that, although past user-click behavior is quite effective for query type identification, adding the new nCS and nRS features achieves improved performance. They report an average precision and recall of 81.5% on their classification. Moreover, their decision tree method was found to outperform that of Lee *et al.* (2005).

Jansen *et al.* (2008) develop a methodology to classify user search intent in terms of navigational, informational, and transactional categories. They consider user query intent with respect to the type of content specified by the query and other user expressions, and they operationalize these categories by defining characteristics for them. Examples of the characteristics for each query intent include: i) in navigational searches the query may contain domain suffixes, ii) in transactional searches the query may relate to image, audio, or video collections, and iii) in information searches the query may contain question words, such as *ways to*, *how to*, *what is*, etc.

They implement these characteristics using a decision tree approach that automatically classifies Web search engine queries. For their experimental study, they use a transaction log from `Dogpile.com`. They selected a random sample of 400 queries from this transaction log and manually classified these queries. Their results indicate that more than 80% of Web queries are informational, with navigational and transactional queries each representing about 10% of Web queries. They relate the variation in reported percentage of navigational and transactional queries from the earlier work to the relatively small size of the samples used in prior studies. Their findings also suggest that about 70–80% of the queries can be classified into one category with a high degree of confidence, while the remaining queries appear to be more problematic and may represent multiple intents. By achieving an average accuracy of 74% in their approach, they conclude that automatic classification of user intent is achievable using data that is readily available to Web search engines.

Furthermore, machine learning techniques have been borrowed by researchers in this area in order to automatically infer the intent underlying user queries. For instance, Baeza-Yates *et al.* (2006) develop models applying supervised and unsupervised learning techniques for query identification, primarily using the content of landing pages. From the content of the queries, they establish three categories for the *goals* which motivate a user to initiate a search: *informational*, *non-informational*, and *ambiguous*. They further consider the *topic* category of queries, which is defined as the type of user information need. One group of their topics belong to the *general* categories from ODP (Open Directory Project)[1] listed as: *arts*, *games*, *kids and teens*, *reference*, *shopping*, *world*, *business*, *health*, *News*, *society*, *computers*, *home*, *recreation*, *science* and *sports*. They also consider three more topics: *various*, *other*, and *sex*. The *various* topical category contains the queries that can be placed into more than one category, while *other* refers to the queries that can be put into any category. Finally, the *sex* topic contains adult queries.

They employed a set of 6042 manually labeled queries, originally obtained from a Chilean search engine, for their experimental evaluation. In the user goal classification

[1] `http://www.dmoz.org`

task, their reported precision varies from 30% to 70%, while the recall for various goals falls between 20% to 90%. The informational goal is found to be the most distinguishable one among the three user goals. The precision is reported to be between 55% to 70% for the topic classification task, while the recall for this task falls in between 20% to 50%.

Various semi-supervised and supervised learning techniques (Beitzel *et al.*, 2005; Li *et al.*, 2008; Nettleton *et al.*, 2007; Tan and Peng, 2008) for query intent detection have been applied mainly over click data (e.g click distribution and query session) and over anchor text information resources. While click data appears to be noisy and anchor text appears to cover a low percentage of the queries (Brenes *et al.*, 2009), other sources of information such as the query string itself and the context in which the query appears are often not considered in the literature.

In part of their work,. Ashkan *et al.* (2008, 2009a) address query intent inference for the traditional categories of query intent, informational and navigational, using two groups of features: i) clickthrough features combined with query features and search engine result page (SERP) features, an ii) a combination of query and SERP features alone. They trained classifiers based on these two sets of features in order to automatically identify the intent underlying 1700 queries selected from the logs of a commercial search engine (and labeled manually). They report an average precision, recall, and accuracy of 89%, 88%, and 90%, respectively, using the first set of features; and an average precision, recall, and accuracy of 83%, 83%, and 84.5% using the second set of features. Their findings suggest that query and SERP features are effective in identifying the intent underlying users' queries. Moreover, a classifier trained on query and SERP information plus the clickthrough features outperform one where only query and SERP features are used. Chapter 3 of this thesis is, in part, based on the preliminary results we reported in these earlier studies.

2.1.2 Commercial Intent

Most existing work focuses on query intent with respect to the traditional navigational, informational, and transactional categories defined by Broder (2002). More recently, there has been growing interest in commercial intent classification. Jansen (2007) employs a user evaluation technique to manually organize commercial queries into five categories: *intent to buy*, *product specific*, *location specific*, *company specific*, and *general*. Most of the queries fall into the product specific category, while a very small percentage focus on companies and locations.

Dai *et al.* (2006) formalize the problem as a binary classification problem: deciding whether or not a search query is intended for commercial purposes, such as intending to buy a product or finding product information in a research stage. In their experimental setup, they picked a random sample of 1408 US English search queries from one day of MSN search log, and manually labeled these queries. They trained their commercial query detector using search result pages and the contents of the top ranking pages returned for the queries. They report 86% precision, 82%recall, and 84% accuracy.

Hu *et al.* (2008) improve the commercial intent detection method of Dai *et al.* (2006) through a skip-chain conditional random field model and a set of additional features. To detect the commercial intent of a query, their method considers two types of features. One contains the generalized OCI intention features of Dai *et al.* (2006), which are extracted from the content of the top result pages returned for the queries. The other consists of historical similarity features, which take past queries into consideration. These two feature sets are used in a conditional random field model which takes a sequence of queries (rather than a single query) issued by a particular user into consideration for personalized commercial intent detection.

Hu *et al.* (2008) limit their empirical evaluation to the search history of four users picked from the logs of a commercial search engine. These users had substantially more queries than other users (a total of nearly 28K non-unique queries in their log). They manually labeled the corresponding queries as commercial and noncommercial taking the clicked pages of the queries into consideration. Their experimental results indicate that their algorithm can improve the performance (in terms of F1 measure) of a baseline OCI detector (Dai *et al.*, 2006) by an average of 8.7% across the four selected users.

Shen *et al.* (2009) take the product query classification into the consideration by classifying Web queries into a predefined product taxonomy. They study the impact of query expansion and the size of training data in their work by comparing the classification performance of different combinations of query representation and training data. There are two approaches used in their work for the enrichment of the Web queries. One approach uses search snippets for query enrichment, and the other expands queries by their similar queries automatically discovered from clickthrough log data.

They use three settings for training data. In the first setting, a set of labeled queries (as well as their enriched representations) is used to train classifier. In the second, the classifiers are built using labeled product names from MSN shopping. However, due to the difference of languages for Web queries and product names, classifiers trained over product names are much worse than those trained over queries. To remove the gap between product names and Web queries, they produce "pseudo" queries by translating product names to Web queries, to give their third setting. Then they train classifiers using the pseudo queries. Although these classifiers cannot beat those trained over labeled queries, they appear to be much better than the classifiers trained over product names directly.

On the other hand, using search snippets to represent Web queries achieves the best results compared to other methods. As retrieving search snippets is time consuming and it may be infeasible for a large scale query set, they minimize the dependency over search snippets by building a "hybrid" classifier that uses the original queries in order to classify Web queries directly. Only when the classifier cannot classify the queries with high confidence, it uses the snippet information. With this solution, they classify more than 70% queries directly without retrieving the search snippets, while maintaining the same accuracy as using search snippets for all Web queries.

Ashkan *et al.* (2009a) and Ashkan and Clarke (2009) address the commercial intent categorization in their work with respect to three goals: i) enriching the queries with contextual information, an issue that has been previously tackled in literature mostly through the click and anchor text information; ii) obtaining reliable training data through a relatively easy approach; and iii) studying characteristics of commercial intent in sponsored search domain. In order to enrich the queries, they employ the information from the query string combined with the features obtained from the content of search result pages returned for those queries. They also consider the relation among the string of the queries and the document URLs listed on the result pages (e.g. whether the query is the substring of any

URL listed among the results).

To extend a set of manually labeled training queries in a relatively easily and reliably way, they employ Amazon Mechanical Turk (2009). A set of 4000 labeled queries was obtained and validated through this process. Finally, they study the commercial intent category with respect to the impact of navigational and informational intent of the queries. Their results indicate that the commercial-navigational queries receive more ad clicks than commercial-informational queries. In other words, the ads that reflect the intent of commercial-navigational queries seem to be more of a target for clicks than the ones that reflect the intent of commercial-informational queries. They further study the commercial category in terms of the specificity of brands, products, and retailers by characterizing these sub-categories with respect to their clickthrough behavior in sponsored search. Parts of Chapter 3 of this thesis is based on the results we reported in these earlier studies.

Understanding the intent underlying user queries can help search providers for search personalization and therefore improve user satisfaction. There has been a growing interest in employing query intent analysis to improve various search experiences such as: Web spam detection (Benczúr *et al.*, 2007), usability of interactive applications (Fourney *et al.*, 2011), and general Web search experiences (Dai *et al.*, 2011). Current research work considers the use of this technique for the purpose of clickthrough analysis in sponsored search, a direction that has not been previously addressed in literature.

2.2 Sponsored Search

In sponsored search, advertisers promote their commercial product or service once they are allocated slots on search engine result pages to show their advertisement links (a.k.a. ad(s)). In order to acquire a slot, an advertiser submits to the search engine their ad and a set of keywords describing the ad, called *bid terms* or *bid phrases* (Bendersky *et al.*, 2010; Fain and Pedersen, 2006; Ghose and Yang, 2008). They also submit a *bid* value for each of the keywords indicating the price they are willing to pay once their ad is shown on a result page for a query matching the bid term(s).

In its simplest form, for a given query a set of candidate ads are obtained through an ad retrieval algorithm (Bendersky *et al.*, 2010). The search system uses an auction mechanism (Edelman *et al.*, 2007) to find candidate ads that can maximize the expected revenue. These ads are ranked in decreasing order based on the expected revenue estimated for each individual ad.

A sample search result page containing the organic results (a.k.a. Web search results or non-sponsored results) and ad results (a.k.a. sponsored results) is depicted in Figure 2.1. As shown in the figure, ads may appear in two different locations on a result page: typically at the top (north) or the right (east) side of the organic results. An ad displayed in either of these locations is represented by three components: the *title*, the *creative* (the text describing the ad) and the *URL of the ad*. Once a user clicks on the ad, they will be directed to a page, called the *landing page* of the ad.

Once a set of N ads are retrieved for a given query q and ranked on the N slots [available at the top and/or side] of the search result page, the expected revenue of the sponsored results can be computed as follows (Hillard *et al.*, 2010):

$$REV = \sum_{i=1}^{N} cost(q', a_i, i) \times P(C = 1|q, a_i)$$

where $cost(q', a_i, i)$ is the cost of a click for ad a_i at position i, computed according to the bid placed by the corresponding advertiser, to be paid once the ad's URL receives a click for the phrase q'. In case of a *standard match*, $q' = q$, in which the ad is shown for a query if the query finds an exact match among the ad's bid phrases. Otherwise, $q' \neq q$ which is referred to as an *advanced match*. For an advanced match, the query is found to be semantically related to the bid phrases, rather than exactly matching them.

Search engines typically rank the ads retrieved for a given query q according to the product of the estimated clickthrough rate, $P(C = 1|q, a_i)$, and the bid amount in an attempt to maximize revenue for the search engine. This *cost-per-click model* is among the most common revenue models, charging an advertiser per click according to their bids.

From the user's point of view, a good quality ad should be trustworthy and relevant to their interests. From the search engine point of view, a click on an ad is viewed as a

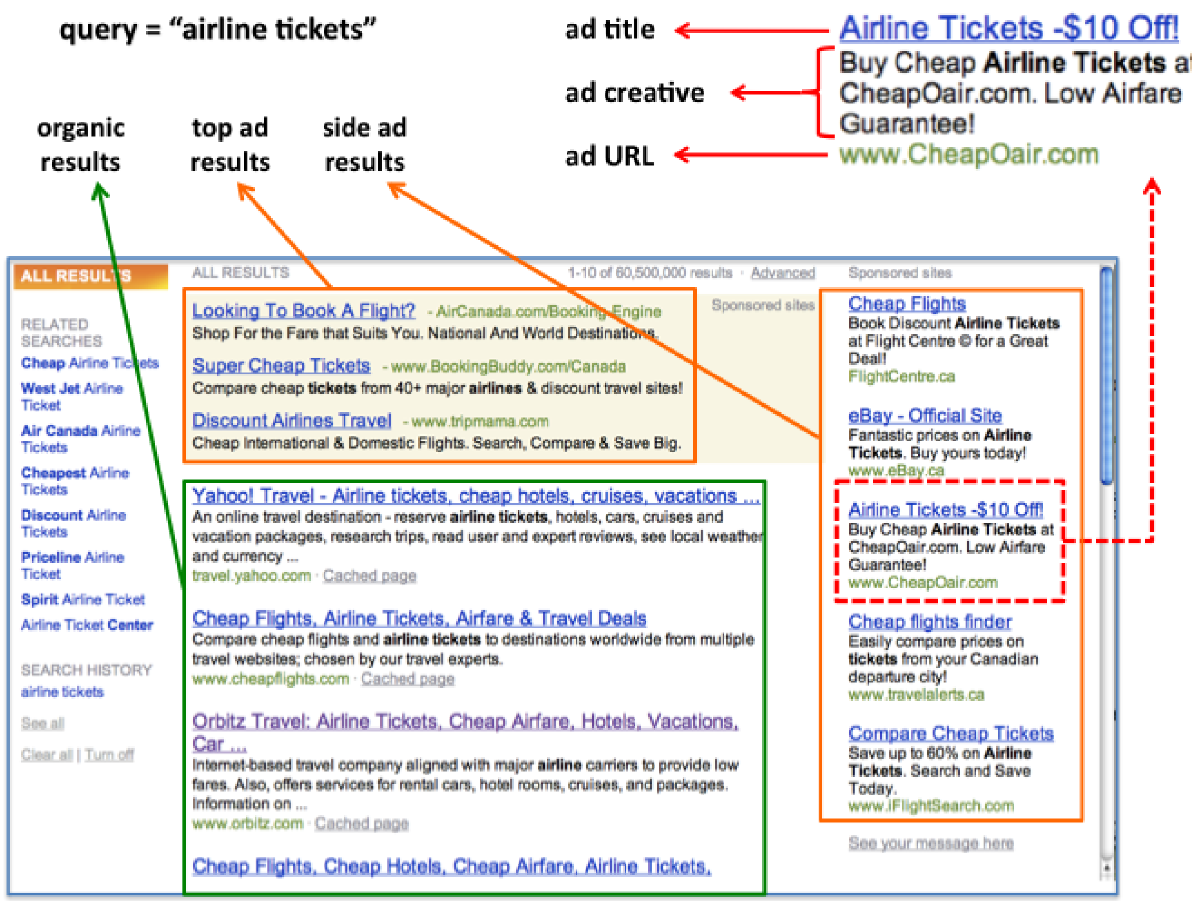


Figure 2.1: Sample search engine result page for the query “airline tickets” that contains ad results (top-listed and side-listed) displayed along with the organic results. The three components of a sample ad are displayed.

potential purchase opportunity for the service or product offered by the advertiser. Hence, the expected ad clickthrough rate (CTR) is the classic measure of ad quality, determining the expected revenue for the search engine. Clickthrough rate is defined as the ratio of the total number of clicks recorded for an ad to the total number of impressions of the ad (AdWords, 2013; Yahoo! Search Marketing, 2013). While an impression may lead to the click on the ad, it may or may not lead to a meaningful action favored by the advertiser, such as the actual purchase. Hence, the conversion rate is another measure of ad quality, and it is defined as the ratio of the number of clicks on an ad that resulted in a meaningful action, known as the conversion (e.g. purchase or sign-up), to the total number of clicks recorded for the ad (AdWords, 2013).

2.2.1 Ad Clickthrough

The probability of click on an ad, representing the ad clickthrough rate (CTR), is targeted in the literature as the basis for comparing the quality of competing ads. As an example, let $P(C = 1|a, pos)$ represent the probability of click on ad a given the positional property pos with respect to the result page on which the ad is displayed. According to Richardson *et al.* (2007), pos denotes the rank position of the ad on the page. Debmbsczynski *et al.* (2008) consider the result page number as well as the position of the ad on this page.

Given another ad, a', one should be able to compare the quality of the two ads given they share the same positional property, and the result of the comparison should be the same once this property changes; if $P(C = 1|a, pos) < P(C = 1|a', pos)$ for property pos, it is expected that $P(C = 1|a, pos') < P(C = 1|a', pos')$ for property pos'. In other words, if two ads with the same positional properties are compared, the results of this comparison should not depend on the particular property, assuming they are the same for both ads (Debmbsczynski *et al.*, 2008). Remember that CTR is used as a representative of the ad quality, and it is mostly assumed to depend on the ad's content.

Note that the above discussion becomes more precise when the effect of the query is also taken into account. Once the query changes, the quality factor of the ad will be evaluated differently according to the user's query, especially with respect to the intent underlying the query. For example, ad a may outperform ad a' within the same positional property, when they appear for query q. However, with respect to another query, q', ad a' may target the user intent underlying q' better than a, and therefore it outperforms a' in the context of q'.

The probability of a click is considered to be dependent on a hidden factor, representing the probability that user views the ad (Becker *et al.*, 2007; Debmbsczynski *et al.*, 2008; Richardson *et al.*, 2007). This is referred to as the examination state of the user, denoted by variable E:

$$P(C=1|a,pos) \quad = \quad \sum_{E\in\{0,1\}} P(C=1|a,pos,E)P(E|pos,a) \tag{2.1}$$

It is intuitively reasonable that the probability that an ad is clicked is zero, given that it is not viewed by the user (i.e. $P(C=1|a,pos,E=0)=0$). In other words, the introduction of the variable E implies that $E=1 \Leftrightarrow C=1$. Substituting this implication into Equation 2.1 results in the following formulation for the probability of clicking on ad a with the positional property pos:

$$P(C=1|a,pos) \quad = \quad P(C=1|a,pos,E=1)P(E=1|pos,a) \tag{2.2}$$

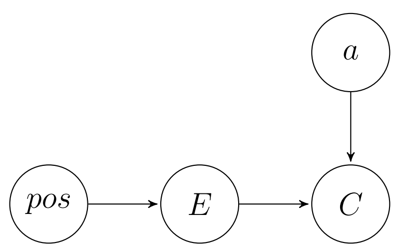


Figure 2.2: The graphical presentation of ad clickthrough model (Becker *et al.*, 2007).

The impact of the hidden variable E can be visualized through the graphical representation presented by Becker *et al.* (2007) and depicted in Figure 2.2. As it can be seen in the figure, and according to the *explaining away* effect in graphical models (Koller and Friedman, 2009), E is marginally independent of a; i.e., the probability an ad is viewed is independent of the ad. Similarly, Richardson *et al.* (2007) makes the same assumption, that the probability an ad is viewed is independent of the ad given the position, and also independent of the other ads shown on the page: i.e. $P(E=1|pos,a)=P(E=1|pos)$.

As it can be seen in Figure 2.2, the *Markov chain* property implies that clicking on an ad depends on the positional property only through the hidden variable E (Becker *et al.*, 2007; Debmbsczynski *et al.*, 2008); i.e. the probability an ad is clicked is independent of the position given it is viewed: $P(C = 1|a, pos, E = 1) = P(C = 1|a, E = 1)$. As a result, Equation 2.2 can be written as follows:

$$P(C = 1|a, pos) = P(C = 1|a, E = 1)P(E = 1|pos) \tag{2.3}$$

The probability that an ad will be clicked can therefore be justified in terms of two factors: i) the probability of clicking on the ad given that it is viewed (examined) by the user, and ii) the probability that the ad is viewed by the user.

In the literature, $P(E = 1|pos)$ is usually estimated based on the results of eye tracking studies. For instance, Richardson *et al.* (2007) consider the heat map of search result page viewership intensity for different ad positions obtained from the eye scan activity of users over the result pages (Hotchkiss *et al.*, 2005). As another example, Debmbsczynski *et al.* (2008) model the positional property in terms of two factors, the result page number and the rank position. They compute $P(E = 1|pos)$ as a factor of the page number, s, and the rank position, r, as follows:

$$P(E = 1|pos) = P(E = 1|s, r) = p_s(s) \times p_r(r)$$

For the normalization of these factors, they set $p_r(1) = p_s(1) = 1$, such that $P(E = 1|s = 1, r = 1) = 1$, implying that an ad is always seen at the first position of the first result page.

The probability of a click on an ad, given that it is viewed, $P(C = 1|a, E = 1)$, serves as an interpretation of clickthrough rate (CTR) for ad a, and it is usually addressed in the literature by way of features related to ad content, landing pages, and bid terms. The following section presents a review of CTR prediction studies.

2.2.2 Clickthrough Prediction

Regelson and Fain (2006) estimate the CTR of ads on a keyword basis using the click rates of existing ads with the same bid terms or topic clusters. Their underlying hypothesis states that the more closely terms are related, the closer their click rate will be. They use a hierarchical clustering of terms based on the partitioning of a keyword-advertiser matrix (Carrasco *et al.*, 2003) in order to measure the semantic proximity of terms as the relatedness factor. For each term assigned to a cluster, they predict future click rate from a combination of the historical click rate on the term and the historical click rate at each level up in the hierarchy. Incorporating information at increasing levels of generality is assumed to help in refining the estimate for terms with a small quantity of historical data. Their results indicate that using cluster membership can substantially improve prediction accuracy for the term based click rates of advertisement links.

Zhang and Jones (2007) examine the correlation between ad clicks and the features generated from query rewrites. These features include syntactic features, such as edit distance, word disagreement ratio, and the length difference between a query and its rewrite. They normalize the click rate by the expected click rate at particular ranks (i.e. ad clicks over expected ad clicks) in order account for the position bias that they refer to as the rank effect. They build a prediction model based on logistic regression using these features. The results of this study indicate that the query rewrite features are predictive of ad clicks, suggesting that future models may be trained based on click log data in place of human relevance judgements.

Richardson *et al.* (2007) indicate that even within the same term there can be a large variation in ad clickthrough rates. They argue that in some cases, the clickthrough rate of the best ad on a particular topic can be ten times that of the average ad on the same topic. Hence, further features that depend on more than just terms are needed to account for the within-keyword variations. The results of a user study conducted by Jansen and Resnick (2005) also suggest that Web searchers consider the creative, the title, and the URL of an ad in deciding whether to click it or not.

Richardson *et al.* (2007) hypothesize that there are at least five rough categories of influence on the user that cause a person to decide to click (or not to click) on an ad: i) The *appearance* of an ad determines if it is aesthetically pleasing; ii) the *attention capture* determines whether the ad draws the user in or not; iii) the *reputation* of an advertiser indicates whether they are known to carry a reputable brands, or on the other hand, may indicate whether the user guesses that the advertiser carries good brands given they are not familiar with the advertiser; iv) the *landing page quality*, judged based on the previous click experiences of the user, may be indicative of the probability that the user will click on an ad associated with the page, where it is hypothesized that many ad clicks go to advertisers that a user is already familiar with (e.g. eBay and Amazon); and v) the *relevance* of an ad to the query.

For each category, Richardson *et al.* (2007) derive a number of features to indicate the quality of the ad for that category. These features are built from the ad's title, the content of its creative, the corresponding bid terms, the landing page, and the relation of the bid terms with the ad features. They build a prediction model based on logistic regression using these features. Their results report a 30% of reduction in the prediction error over a baseline model that considers only the historical click rate of ads. Moreover, their findings suggest that users prefer to click on ads from more reputable, established entities, whereas they tend to avoid clicking on ads with various free offers and trials.

Debmbsczynski *et al.* (2008) approximate the title and the body of an ad by combining all queries for which the ad was displayed. They use these ad features along with the ad's landing page to build a prediction model based on decision rules, generating recommendations on how to improve the quality of ads. Their prediction results report 13% of improvement over a baseline that computes the average CTR over the training examples.

Ghose and Yang (2008) study sponsored search at the level of bid terms (i.e. ad keywords). They empirically estimate the impact of keyword attributes, such as the presence of retailer information, brand information and the length of the keyword, on user clickthrough and purchase behavior. They use a dataset of 1799 unique keywords collected from a nationwide retailer for their experimental study.

The results of the study by Ghose and Yang (2008) indicate that while retailer-specific information in ad keywords increase clickthrough rate, brand-specific information in ad keywords increases the conversion rate. As for the length attribute, they find that longer keywords tend to experience lower clickthrough rate. A longer keyword is less frequent and typically tends to suggest a more directed or specific search, as opposed to a shorter keyword that is more frequent and typically suggests a more generic search. A user with no directed search (as a result of shorter keywords) has a wide consideration set, suggesting that the user is likely in the surfing mode (Danaher and Mullarkey, 2003). Such a user may click on several links before they find a product to purchase. On the other hand, the probability of a goal directed user (as a result of longer queries in a more directed search) clicking on a retailer's ad is low unless the retailer carries the specific product that the user is searching for. As a result, it is argued that longer queries and longer query terms are more common for less focused searches, which could end up with a lower chance of a click or a conversion from the user.

The findings of this study also confirms the effect of the rank of ads on their clickthrough rates; i.e. the lower the rank of an ad (displayed higher on the result page), the higher the clickthrough rate. An interesting observation is that for keywords that contain retailer-specific information, a lower rank leads to even higher clickthrough rate. On the other hand, longer keywords appear to moderate the effect of the rank on click-through; i.e. as the number of words in an ad increases, a given rank leads to relatively lower clickthrough rate than an ad with a fewer number of words.

Motivated by prior successes in ad click prediction, Sculley *et al.* (2009) study the *bounce rate* for ads, which is defined as the fraction of users who click on the ad but almost immediately move on to other tasks. The problem of bounce rate prediction is addressed using ad-content-based features, along with landing page features and bid terms. A correlation is shown between ad click rate and ad bounce rate, where ads with very low bounce rates have very high click rates. Their observations also indicate that navigational and commercial queries result in a very low bounce rate

Hillard *et al.* (2010) train binary classifiers to predict the relevance of ads given particular query terms. Their baseline model uses query length and a group of features that compare the query to three properties of ad content (title, creative and display URL). These features include word overlap (unigram and bigram), character overlap (unigram and bigram), cosine similarity, and the number of bigrams in the query that had the order of the words preserved in the ad zone (ordered bigram overlap).

They use a data set obtained from a particular advertiser database with a TF-IDF based ad retrieval system that retrieves 20 ads per query. The data contains 7K unique queries sampled from the search engine traffic. A set of about 80K query-ad pairs were judged by a group of professional editors, from which about 40K pairs were used for the evaluation purposes.

The primary experimental studies performed by Hillard *et al.* (2010) report the maximum performance over their baseline to be based on a classifier trained using Gradient Boosting Decision Trees (GBDT (Zheng *et al.*, 2007)). The maximum precision, recall, and F-measure for this classifier are reported as 0.671, 0.551, and 0.605, respectively.

The baseline relevance model operates based on simple text overlap features, failing to detect relevant ads if no syntactic overlap is present between the query and ad properties. The authors introduce two approaches for employing user click data to learn semantic relationships between queries and ads. The former one uses historical click rate from query-ad pairs as an extra features used in the relevance model. Their experimental results confirm that the observed click history is helpful in predicting relevance when sufficient observations are available. However, such information is available for a portion of the ads that has seen sufficient search traffic.

Hence, the latter approach uses an aggregation of the historical click information from other parts of the sponsored search system in order to compensate for the missing information for those query-ad pairs that do not have click history. These aggregations benefit from observed click behavior on similar ads (e.g., from the same ad campaign). The experimental results report 7% improvement in recall comparing to the baseline model, with 2% reduction in precision. Combining this approach with the former approach (using the observed click history features) recovers precision over the baseline while maintaining improved recall.

Most of these existing efforts study ad clickthrough in sponsored search through factors related to *ad content*, *landing pages*, and *bid terms*. Ashkan and Clarke (2013); Ashkan *et al.* (2009b) provide insights concerning the relationship between user ad clickthrough behavior and the characteristics of queries and their corresponding ad links shown on a search result page. They propose models to predict the aggregated clickthrough behavior of ads by considering the influence of such factors as the location of ads and the rank of ads, along with query intent. The intuition behind this idea is that ads positioned at the top of a page may receive more clicks, even if they are less relevant than other ads (i.e. examination model). Also, a weakly related ad appearing with the results of a commercially oriented query may receive more clicks than a strongly related ad appearing with the results of a less commercially oriented query. Chapter 4 of this thesis is based on the results we reported in these earlier studies.

All these studies adopt the common assumption of a *trust bias* (Joachims *et al.*, 2005) for higher ranked results, which has been confirmed through eye-tracking studies (Joachims *et al.*, 2007). The trust bias provides a basis for the *examination hypothesis* proposed by Richardson *et al.* (2007). Under this hypothesis, it is assumed that the clickthrough rate decreases towards lower positions due to reduced visual attention from the user. Jansen *et al.* (2007) report that the link examination behavior of users is similar to their click behavior, with users preferring top-listed results for both sponsored and non-sponsored links. Hence, two main factors influencing an ad's clickthrough rate under this examination model are the relevance of the individual ad to user [commercial] need and the rank position of the ad on the page.

A more intuitive approach considers the influence of the co-appearance of ads. A relatively strong and compelling ad appearing before a second ad may distract the user from the second ad, regardless of its quality (Xu *et al.*, 2010). On the other hand, a weak ad may annoy the user into abandoning the list, regardless of the quality of the rest of the list. This concept is described by Ghosh and Mahdian (2008) as the *externality effect* of ads in advertising. In considering these externality effects, they work from a linear browsing assumption, in which users visually scan the ad list from top to bottom (Aggarwal *et al.*, 2008). Once an ad is examined by the user, ad-specific factors (e.g., relevance of the ad as perceived by the user) will determine the click decision and continuation probability. In contrast to the examination model, the probability of clicking on an ad depends on the quality of the ads shown above it on the page.

Chapter 5 of this thesis is based on our preliminary results reported in (Ashkan and Clarke, 2012), where a linear browsing assumption is adopted that considers the externality effect of ads. This assumption is related to the *cascade model* of user behavior borrowed from the organic search domain (Craswell *et al.*, 2008). Considering the sponsored search domain, under this model, an ad is examined only if the user first scans over all the previously displayed ads. In the next section, we present a review of the cascade model, and a few other models based upon it.

2.3 Cascade Model of User Behavior

The cascade model of user behavior (Craswell *et al.*, 2008) was originally proposed in the context of organic search. Given that a user issues query q, the binary hidden variable E_i indicates whether the user examines the document d_i displayed at rank i in the result list. Similarly, C_i is defined as a binary variable representing whether the user clicks on d_i given they viewed its caption.

According to the cascade model, the probability of examining d_i (i.e., $P(E_i = 1)$) is known as the examination probability, which is assumed to be dependent on the quality of the documents shown prior to it (i.e., listed at earlier ranks) on the page:

$$P(E_i = 1) = \prod_{j=1}^{i-1}(1 - \omega_{d_j}^{q})$$

where $\omega_{d_j}^{q}$ represents the attraction probability of the document d_j with respect to the query q. This probability represents the chance that the user perceives d_j to be relevant to their information need and clicks on it, given they examined it. The attraction probability is sometimes known as *perceived* relevance (Chapelle and Zhang, 2009)).

The cascade model makes the following assumptions about user browsing and click behavior: i) the user performs a linear scan of the result list starting from the top ii) there is at most one click per search; hence, the model cannot explain multiple clicks, and iii) if the user does not click on a viewed link, they continue examining links, i.e., the user is infinitely persistent:

$$
\begin{aligned}
&P(E_1 = 1) = 1 \\
&P(E_{i+1} = 1|E_i = 0) = 0 \\
&P(C_i = 1|E_i = 1) = \omega^q_{d_i} \\
&P(E_{i+1} = 1|E_i = 1, C_i = c_i) = 1 - c_i
\end{aligned} \tag{2.4}
$$

where c_i presents the click event value (0 or 1) for the variable C_i.

There are other biases and factors addressed in related research work. The user browsing model (UBM) (Dupret and Piwowarski, 2008) and Bayesian browsing model (BBM) (Liu *et al.*, 2009) are among the click models that do not employ the cascade assumption. They extend the examination hypothesis by considering the dependency on the positional distance to the previous click in the query session. The task-centric click model (TCM) (Zhang *et al.*, 2011) considers the sequences of queries and clicks in a session as a task and characterizes user behavior related to a task as a collective whole. It formalizes user behavior with respect to two biases; one is query reformulation and the other is the user's desire for unseen documents in a session. Hu *et al.* (2011) study the impact of query intent diversity on the existing click models. They argue that user click can not be explained only by the relevance and position of the document, but also by the diversity of the user's queries. The whole page click model (Chen *et al.*, 2011) differs from the previous approaches as the authors explore the whole search result page including all the click blocks (e.g., organic results, sponsored results, etc) on the page as an integrated entity. Their findings include that if there is a click in a given block, a user is less likely to examine the next block.

There have been further efforts to extend the cascade model through various hypotheses and assumptions, all aiming at modeling user browsing/click behavior in a more realistic fashion, in which multiple clicks are permitted. All of these models share a notion of user *patience* and *persistence* as they move from document to document. Zhu *et al.* (2010) define a group of user and URL specific attributes, such as query, browser type, local hour, and the position to model the relevance and examination transitions effects as random variables. The Click Chain Model (CCM), proposed by Guo et al. (Guo *et al.*, 2009a), defines the transition probability from document i to $i + 1$ in the cascade model through three global parameters. These parameters are fixed and independent of the users and URLs.

The dynamic Bayesian network (DBN) model, proposed by Chapelle and Zhang (2009), defines a persistence factor that is assumed to be fixed and shared across query sessions. According to DBN, a user starts from the first document and keeps on examining d_{i+1}, given they already examined d_i in two cases: i) either they do not click on d_i and skip it with a probability of λ, or ii) they click on d_i and find it un-satisfying (non-relevant), so they move on to the next document with a probability of λ. In both cases, a binary variable S_i, which indicates the satisfaction status, becomes 0. In case of a click on d_i, this variable will be set to 1 if and only if d_i satisfies the user. Here, $\nu^q_{d_i}$ represents the *satisfaction* probability, also known as the post-click relevance:

$$\begin{aligned}
&P(E_1) = 1 \\
&P(E_{i+1} = 1|E_i = 0) = 0 \\
&P(E_{i+1} = 1|E_i = 1, S_i = 0) = \lambda \\
&P(S_i = 1|C_i = 1) = \nu^q_{d_i}
\end{aligned} \tag{2.5}$$

where λ represents the persistence of the user in browsing, and it is considered to be a fixed parameter that is constant across all query sessions.

The dependent click model (DCM) (Guo *et al.*, 2009b) is also based on the cascade assumption but it models the user persistence in a different fashion. Here, a position-dependent form of the λ parameter, denoted by λ_i, is defined as the chance that the user would be willing to see more results after a click at position i. It is assumed that the user starts from the first document and continues examining the next document with a probability that depends on their click action at rank i. The next document is examined with a probability of one or λ_i given that the user skips or clicks the document at rank i, respectively:

$$\begin{aligned}
&P(E_1 = 1) = 1 \\
&P(E_{i+1} = 1|E_i = 0) = 0 \\
&P(E_{i+1} = 1|E_i = 1, C_i = c_i) = \lambda_i^{c_i}
\end{aligned} \tag{2.6}$$

where the maximum likelihood estimate of the λ_i values is empirically computed for the various positions on the result pages.

Following from the related work in click and browsing modeling in Web search, Ashkan and Clarke (2012) propose *query* biases in the domain of sponsored search in order to better cope with the actual user behavior in this domain. This effort is extended in this thesis by introducing and formulating the *location* bias in sponsored search as well as the query biases. A query- and location- aware browsing/click model is proposed in Chapter 5. Inference of parameters of the model are detailed as well as the experimental studies performed to evaluate the performance of the model under various settings. The experimental studies are further extended to explore whether the introduced biases reflect varying behavioral patterns for different users.

Chapter 3

Understanding and Inferring Query Intent

Understanding user intent in Web search requires us to identify the information need underlying the query. As defined by Shneiderman *et al.* (1997), *an information need is the perceived need for information that leads to someone using an information retrieval system in the first place.*

Different users may issue the same query but have varying information needs, suggesting that subjectivity and ambiguity exists in query intent identification. For instance, a query such as "Niagara falls" represents multiple intents. A user may be looking for the history of the city, pictures of the falls, or accommodations. The work of researchers such as Clarke *et al.* (2008), Radlinski *et al.* (2009), and Welch *et al.* (2011) address this issue by explicitly considering the novelty and diversity of user information needs. In this thesis, query intent is considered in terms of the popularity of user needs (Radlinski *et al.*, 2010). In other words, query intent is defined as the most popular understanding of the information need of a typical user in the system.

This chapter studies several aspects of query intent, particularly commercial intent. A query is considered to have *commercial* intent if the user issuing the query is most likely to plan an immediate or future purchase of a specific commercial product or service.

Otherwise, the intent is considered to be *noncommercial*. We also introduce three sub-categories of commercial intent: i) *product* specific intent, ii) *brand* specific intent, and iii) *retailer* specific intent. These categories were identified from a manual examination of query logs. They are validated through consistency of annotation results and later through click analysis. In addition, the standard categories of Web queries as described by Broder (2002) are addressed in this chapter. However, for simplicity, the transactional queries (Kang, 2005; Li *et al.*, 2006) are subsumed under the categories of *navigational* or *informational*, as appropriate. The intent behind a *navigational* query is to locate a specific Webpage (which may be for transactional purposes). If the intent is not navigational, it is assumed to be *informational* (even if the ultimate goal is transactional).

The manual annotation of a set of queries is described later in the chapter. It is followed by a semi-automatic approach for labeling a relatively large batch of training queries in various dimensions of query intent by using crowdsourcing. The results of evaluations over the labeled queries are discussed later. In the remainder of the chapter, query intent inference is addressed. The contribution of query terms and their corresponding ad clicks on commercial queries are studied first. Finally, a methodology is developed for using query specific information and the content of search engine result pages (SERPs) to identify the intent underlying user queries in various dimensions. There are two settings addressed here: i) query features are combined with SERP features, and ii) SERP features are used alone.

Although clickthrough information is found among the common sources of features for query intent analysis, we do not use it in our query intent classifiers. There are a number of reasons discussed later to justify this decision. The main reason here is to avoid any circularity effect or possible distortion that ad click features could create in the remaining clickthrough analyses conducted throughout the thesis.

3.1 Manual Annotation

A set of 1700 queries from the 45K queries in the training set $A^{(1)}$ was selected for manual annotation, as follows: 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 the set $A^{(1)}$, and ii) the ad click frequency of the query was greater than 10. Each selected query was then manually labeled in different dimensions by three independent annotators.

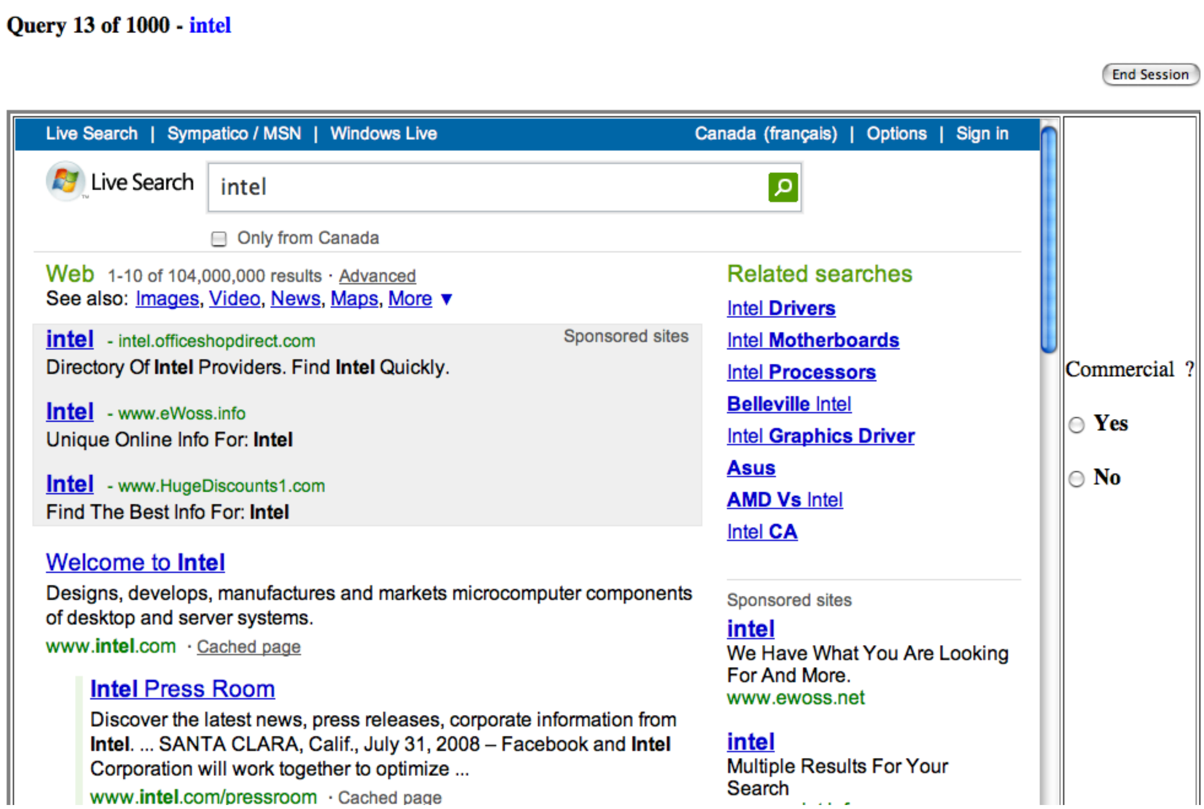


Figure 3.1: A snapshot of the Web application used for the manual annotation

A Web application was developed in PHP for annotation purposes so that each annotator was able to login to the system and view one query at a time along with the corresponding result page returned by a commercial search engine. The Live search engine (currently known as Bing [1]) was chosen in order to keep the labeling consistent with the

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

source of ad click data. A snapshot of the application is depicted in Figure 3.1.

The annotators were responsible for judging the assumed commercial intent of the search queries from the perspective of a general user. If the assumed purpose of submitting a query was to make an immediate or future purchase of a product or service, the query was labeled as “commercial”. Otherwise, if the purpose of the query was assumed to have little to do with commercial activity, it was labeled as “noncommercial”. In a separate round of labeling, the annotators were asked to judge the presumed intent of a general user for the same batch of queries in navigational/ informational dimension and based on the definitions of these queries.

Having a snapshot of each query’s search result page during the annotation process was found helpful for the annotators in order to provide them with a better view of the nature of the query, whether it is commercial or noncommercial and navigational or informational. For instance, if the query is “airline tickets”, the content of the result page that includes words, such as “cheap”, “flight” and “deal” could suggest the commercial nature of this query in general. As another example, if the query is “cheapoair”, the appearance of `www.cheapoair.ca` as the URL of the first result link could suggest the navigational nature of the query in general.

The final label of each query has been assigned based on the majority agreement among the annotators (at least two of them had the same opinion about a query). There was over 80% agreement (i.e. queries for which all annotators assigned the same label) among the annotators in labeling across all the dimensions of query intent. However, the annotators were responsible for judging the presumed intent of the queries from the perspective of a general user, requiring considerable time and effort. To extend this work, the set of labeled queries needed to be increased relatively easily and reliably. For this purpose, Amazon Mechanical Turk (2009) was employed. The obtained set of labeled queries from Mechanical Turk has been further validated and filtered using the above set of manually labeled queries.

3.2 Query Annotation via Crowdsourcing

Crowdsourcing systems (CS) have provided researchers and practitioners with opportunities to solve a wide variety of problems in Web-related contexts. Wikipedia and Linux are among the well-known examples in which a crowd of users explicitly collaborates to build a long lasting artifact that is beneficial to the whole community (Doan *et al.*, 2011). In other examples of crowdsourcing systems, users may implicitly collaborate. For instance, in the ESP game (Von Ahn and Dabbish, 2004), users label images as an implicit effect of playing the game. Another class of crowdsourcing systems, such as Amazon Mechanical Turk (2009), also benefits from users, but the users are coming together for a particular task. The general goal of this group is solving problems, where nothing is long lasting and no community exists. There are various types of crowdsourcing systems that can be classified along many dimensions. The details of these can be found in the work by Doan *et al.* (2011). In particular, Doan et al. define a crowdsourcing system as follows:

> *A system is a CS system if it enlists a crowd of humans to help solve a problem defined by the system owners, and if in doing so, it addresses the following four fundamental challenges: How to recruit and retain users? What contributions can users make? How to combine user contributions to solve the target problem? How to evaluate users and their contributions?*

The remainder of this section and the next aim at addressing these issues for the crowdsourcing process used in this thesis, which has the purpose of annotating queries along various dimensions of query intent.

According to Amazon [2], "*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, transcribing audio recordings*", or in the case of this research work 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.

[2] http://www.amazon.com

In order to obtain a larger set of labeled queries as the ground truth for the training and evaluation purposes, we selected an additional set of 3000 queries from set $A^{(1)}$ to augment the existing set of 1700 manually labeled queries from the previous section. Starting from an arbitrary point in the sorted impression file (approximately $\frac{1}{5}$ of the length of the file from the beginning), 3000 queries were selected, where the query was contained in the set $A^{(1)}$ and was not among the previously labeled 1700 queries. This approach to selection assures that the set of 3000 queries is selected from a continuous period of time in set $A^{(1)}$ (similar to the previous set of 1700 queries). We refer to this set as the *MTurk set*.

In addition, a set of 1000 queries was randomly selected from the manually labeled queries as a *seed set* in order to be used to validate the results obtained from Mechanical Turk. Consequently, a total of 4000 queries were obtained to be labeled by Mechanical Turk and to eventually be used for training and evaluation purposes.

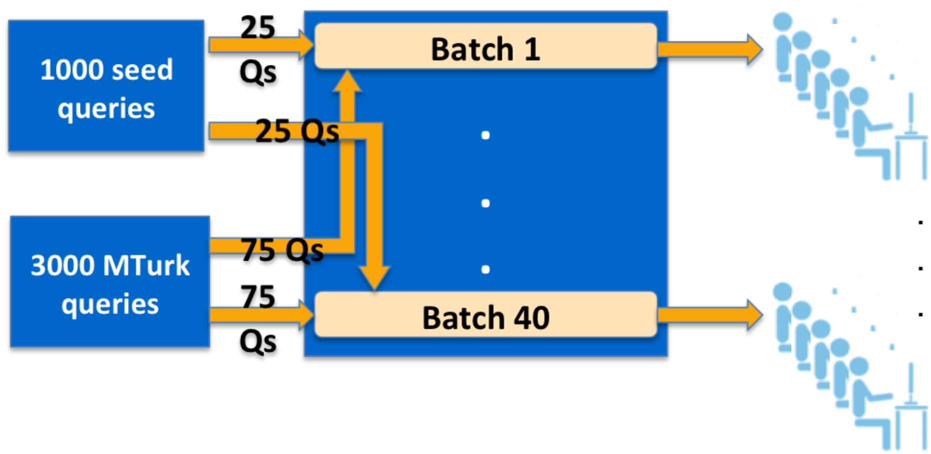


Figure 3.2: The labeling process through the Amazon Mechanical Turk.

The entire set of selected queries (i.e. 4000 queries) was then divided into 40 batches of 100 queries, with each batch containing 25 seed queries and 75 MTurk queries. The labeling process is depicted in Figure 3.2. These batches were submitted to Mechanical Turk, each as a single HIT, in order to be labeled according to the instructions that were provided for the annotators. The annotators 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 an annotator, the labels assigned to the seed queries of the batch were compared against the actual labels of those queries (previously determined by the three local annotators). If the agreement of the annotator with the local annotators was found to be above 60%, the labels assigned by this annotator were accepted. Otherwise, the labels were ignored and the same batch of queries was submitted for an extra round of labeling. If the agreement was found to be above 75%, a bonus was awarded to the annotator.

This process was continued until all batches were successfully labeled by five different annotators. The final label of each query has been assigned based on the majority of the labels obtained for the query. At the end, 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*.

A similar process was repeated in order to obtain labeled queries for the specific sub-categories of the commercial intent. A set of 510 manually labeled commercial queries was considered as the *seed set*, while a set of 1500 queries, which were all labeled as commercial queries from the previous process, was considered as the *MTurk set*. A total of 15 batches (each containing 134 queries) were created with each batch containing 34 seed queries and 100 MTurk queries. These batches were submitted to Mechanical Turk, each as a single HIT, in three rounds. Each round corresponded to one of the three sub-categories of commercial intent: product, brand, and retailer. The annotators were asked to judge the presumed intent of the search queries from the perspective of a general user as follows:

> If the query is related to a specific product, it is labeled as a "specific product", otherwise as a "broad category of products". If the query is related to a specific

retailer, it is labeled as a "specific retailer", otherwise as an "unknown retailer". If the query is related to a specific brand, it is labeled as a "specific brand", otherwise as an "unknown brand".

For instance, the query "Walmart" is considered to represent a broad category of products, with unknown brand, but retailer specific. The query "used car" is considered to be product specific, with unknown brand and unknown retailer. The same strategy for accepting or ignoring the HITs was used, and the final label of each query was assigned based on the majority of the labels obtained for the query in each category.

3.3 Evaluating Annotation Results

For each dimension of query intent, the queries of a HIT are labeled by five different annotators and into two categories. Hence, there are three possible states of agreement level for each query in a HIT: i) all five annotators agree (i.e. 5-0), ii) four agree on one category and one on the other (i.e. 4-1), and iii) three agree on one of the categories and the other two agree on the other category (i.e. 3-2).

Table 3.1: Percentage of HITs falling in different agreement states for different dimensions of intent.

	5-0	4-1	3-2
Commercial /Noncommercial	59.1%	25.8%	15.1%
Navigational/ Informational	57.4%	28.9%	13.7%
Product Specific/ Generic	32.3%	33.7%	34%
Retailer Specific/ Generic	35.4%	38%	26.6%
Brand Specific/ Generic	42.3%	34.3%	23.4%

The result of annotation through Mechanical Turk is reported in Table 3.1 in terms of percentages corresponding to each of these agreement levels for each dimension of query intent. As can be seen in the table, annotators have high agreement on the categories of queries in the two general dimensions (i.e. commercial/ noncommercial and navigational/ informational). There are 59% and 57% of the cases reported for the two general dimensions where there is full agreement among the annotators. Only 15.1% and 13.7% of the queries ended up with low agreement (3-2) among the annotators. For the sub-categories of commercial intent, on the other hand, higher percentage numbers are reported for the moderate (4-1) and low (3-2) agreement levels. This issue will be addressed later in this chapter when training sets for the classification of the sub-categories of commercial intent are adjusted accordingly.

Recall from the previous section that the queries from the seed set were used to qualify the result of a HIT for acceptance. In addition, inter-annotator agreement measures, such as Cohen's Kappa (Cohen, 1960) are employed in order to measure the agreement among the annotators who labeled the submitted queries. Kappa reflects the proportion of agreement corrected for chance, and it is scaled to vary from -1 to +1 so that a negative value indicates a poorer than chance agreement, zero indicates exactly chance agreement, and a positive value indicates better than chance agreement. A value of unity indicates perfect agreement. The use of kappa implicitly assumes that all disagreements are equally serious.

Cohen's kappa measures the agreement between two annotators when each classifies a number of items into a number of mutually exclusive categories. In the case of this work, where multiple annotators label the queries, Fleiss' kappa (Fleiss and Cohen, 1973) is used. Fleiss' kappa works for a number of annotators, each giving categorical labels to the entire set of a fixed number of items. However, in our case the entire set of queries is broken into smaller batches in order to be submitted as reasonable-size HITs to the system. Hence, in each dimension of query intent, we measured Fleiss' Kappa for every submitted HIT which were labeled in that dimension by five independent annotators. The final Kappa value for the entire set with respect to a particular query intent dimension is calculated as the mean of Fleiss' Kappa for the batches in that dimension.

The mean Kappa for each dimension is reported in Table 3.2. The number of batches falling in different margins of Kappa as defined by Landis and Koch (1977) is also presented in Table 3.2. Landis and Koch (1977) give six levels of interpretation for Kappa

Table 3.2: Measure of agreement among the annotators in terms of the number of batches falling in different margins of Kappa along with the mean Fleiss' Kappa for each dimension.

	Poor (< 0)	Slight (0-.20)	Fair (.21-.40)	Moderate (.41-.60)	Substantial (.61-.80)	Perfect (.81-1)	mean Kappa
Commercial/ NonCommercial	0	1	2	13	24	0	0.6028
Navigational/ Informational	0	0	8	8	23	1	0.5948
Product Specific/ Generic	0	4	8	3	0	0	0.2815
Retailer Specific/ Generic	0	1	6	8	0	0	0.3666
Brand Specific/ Generic	0	0	8	7	0	0	0.4139

value as *poor agreement*, *slight agreement*, *fair agreement*, *moderate agreement*, *substantial agreement*, and *perfect agreement*. We put each batch of queries (HIT) into one of these agreement levels according to the value of Fleiss' Kappa calculated for that HIT. The results for 40 HITs of the two general dimensions and 15 HITs of the specific sub-categories are shown in Table 3.2. As can be seen in the table, in the two general dimensions, most of the HITs ended up with substantial agreement. For the three specific sub-categories, most have moderate or fair agreement. This outcome is consistent with the observation obtained from Table 3.1 where the levels of agreement for the queries labeled into the three specific sub-categories were not found to be as strong as the ones in the two major categories. Therefore, for the specific sub-categories in the remaining of the work, we will focus on the queries that have obtained the agreement level of either 5-0 or 4-1 among their annotators.

3.4 Inferring Query Intent Categories

A major goal of this section is to study several aspects of query intent, particularly commercial intent. The contributions in this section are twofold, as follows:

The contribution of query terms and their corresponding ad clicks on commercial queries are studied first. A primary empirical observation indicates that strong commercial terms, such as *sale*, *cheap*, and *store* appear among the top frequent terms in the clicked queries. This observation motivates a hypothesis about contributions of individual terms and their ad clicks towards commercial intent of the queries. A probabilistic model is proposed following this hypothesis.

Later, a methodology is developed for using query specific information and content of search engine result pages (SERPs) to learn and infer intent underlying queries in various dimensions of query categories. The set of labeled queries obtained earlier in the chapter is used to train and evaluate these classifiers.

3.4.1 Term Based Commercial Intent Analysis

In this section, we study the impact of previously recorded ad click information on query intent analysis for sponsored search. In particular, we investigate the relationship between query terms and user click behavior over the advertisement links.

Table 3.3: List of most frequent terms in the data set

group	**most frequent terms**
clicked	free, ebay, games, online, yahoo, home, sale, car, hotels, pages, download, cheap, parts, jobs, used, store, phone, airlines
non-clicked	yahoo, myspace, free, online, school, games, state, home, news, mail, bank, city, pictures, music, university, weather

The empirical study reported in this section is based on the original data set introduced in Section 1.3. We calculated the frequency of terms that appeared in clicked and non-clicked queries of the click logs. As seen in Table 3.3, strong commercial terms, such as *sale*, *cheap*, and *store* appear among the top frequent terms in the clicked queries. Some terms, such as *free*, appear in both. These observations motivate the hypothesis about contributions of individual terms and their ad clicks towards commercial intent of queries. The set of labeled queries obtained in the previous section is used to validate this hypothesis.

If a query receives an ad click, we assume the user intent is more likely to be commercial than if it does not. $P(C = 1|q)$ is defined as the probability that an ad click occurs ($C = 1$) when the query q is entered by a user. Similarly, $P(C = 0|q)$ is defined as the probability that a click does not occur when q is entered. As there are two hypotheses in this model (click or no click), we consider the ratio of the posterior probability of a query under these hypotheses. The relative value of the likelihood-ratio is used for making a decision between the two hypotheses. Given a query q, the log-likelihood-ratio is calculated as follows:

$$\log \frac{P(C=1|q)}{P(C=0|q)} = \log \frac{P(q|C=1)P(C=1)}{P(q|C=0)P(C=0)}$$

Given q consisting of n terms w_1, ..., w_n, where $n > 0$, $P(q|C = 1)$ and $P(q|C = 0)$ are estimated by assuming independence between the query terms. Applying Bayes' Theorem (Büttcher *et al.*, 2010), the log-likelihood-ratio for q can be written as:

$$\log \frac{P(C=1|q)}{P(C=0|q)} = \log \left[\frac{P(C=1)}{P(C=0)} \prod_{1}^{n} \frac{P(w_i|C=1)}{P(w_i|C=0)} \right] \tag{3.1}$$

In addition, the probability of each term contributing or not contributing to a click can be estimated as follows:

$$P(w_i|C=1) = \frac{Q^c_{w_i}}{Q^c_w}$$

$$P(w_i|C=0) = \frac{Q^{\bar{c}}{}_{w_i}}{Q^{\bar{c}}_w}$$

where $Q^c_{w_i}$ and $Q^{\bar{c}}{}_{w_i}$ are the number of times term w_i appears in clicked and non-clicked queries respectively. Q^c_w and $Q^{\bar{c}}{}_w$ are correspondingly the number of all terms in clicked and non-clicked queries. Moreover, $P(C = 1)$ and $P(C = 0)$ may be estimated in terms of the ratio of the number of queries resulting in a click (Q_c) to the total number of queries (Q):

$$P(C=1) = \frac{Q_c}{Q}, and \;\; P(C=0) = 1 - P(C=1)$$

The log-likelihood-ratio value for each of the 4000 queries was calculated according to Equation 3.1. The queries were then sorted by that value, and a cut-off point was determined so that all the queries with the value above the point would be considered as *commercial* and all below the point would be considered as *noncommercial*.

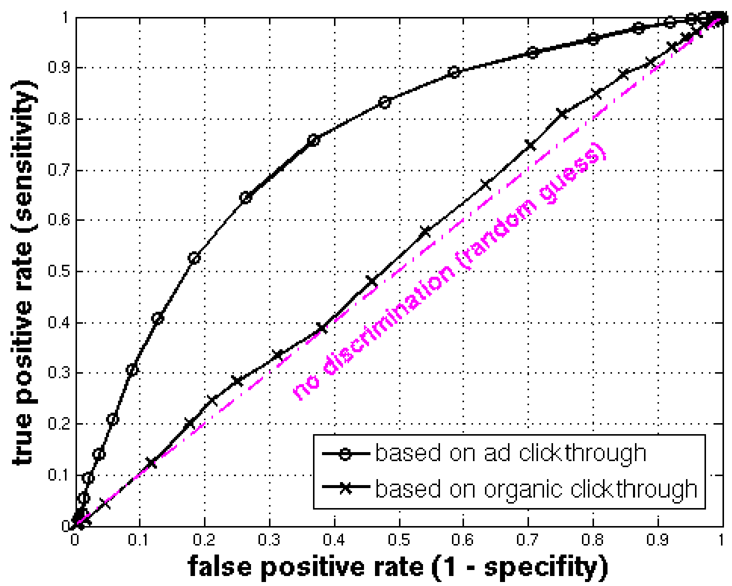


Figure 3.3: ROC curve of prediction from on ad click data versus the prediction from on organic click data.

The predicted label for each query is then compared against the one obtained through the crowdsourcing experiment in order to determine the accuracy of prediction based on a cut-off point. Finally, the true positive rate (sensitivity) is plotted in function of the false positive rate (1-specificity) for different cut-off points as the ROC curve in Figure 3.3. For comparison purposes, the figure also plots a similar result based on an organic click data taken from the same commercial search engine.

We further analyzed the results obtained from ad click data by removing the queries falling within a specific margin above and below the cut-off point. Our reasoning is that queries with a value close to the cut-off point could be considered to have an ambiguous intent (i.e. they could be placed in any of the two categories). We tested this hypothesis

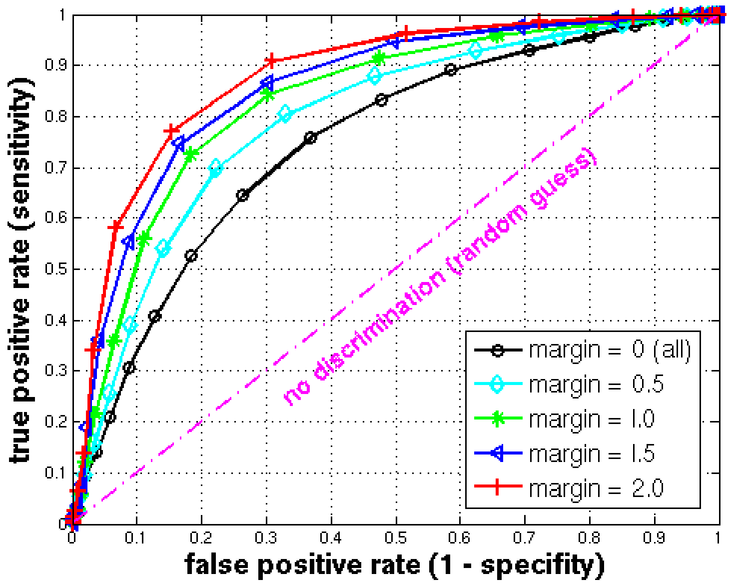


Figure 3.4: ROC curves for varying margin values.

with four values as the margin: 0.5, 1, 1.5, and 2. The resulting ROC curves, based on each margin value, is plotted in Figure 3.4 beside the original ROC curve from the previous figure. As is shown in the figure, the greater the margin, the closer the curve is to the upper left corner.

Each point on the ROC curve represents a sensitivity/ specificity pair corresponding to a particular decision threshold (cut-off point). It is observed that the curve obtained from ad clicks represents a stronger curve in comparison to the one based on organic data (closer to the upper left of the ROC space). In order to test the significance of this difference, the bootstrap test (Hanley *et al.*, 1983) based on AUC (area under the curve) was performed. The test was implemented using pROC package (Robin *et al.*, 2011) drawing 2000 bootstrap replicates with replacement. The results of the test at a significance level of 95% report the p-value to be below $2.2e^{-16}$, suggesting that the difference between the two curves are statically significant. The test reports AUC values of 0.7366 and 0.5287 for the curves based on the ad data and the organic data, respectively.

The main purpose of this study was to confirm that terms with respect to their click history in sponsored search are effective in detecting the commercial intent of queries that include such terms. While the findings of this study provide us with insights about the contribution of query terms and their corresponding ad click information on commercial intent detection, it is more important to note that using click log is costly. Moreover, ad click log information may not work for all dimensions of query intent the same as it does for commercial/ noncommercial categories. Finally, as one of the main objectives of this research work is to study the characteristics of various query intents in terms of their impact on predicting clickthrough and user preferences, relying only on the click log information for intent inference is avoided. In other words, we avoid the possible circularity the use of ad click features would create in this process. Nevertheless, the query string, itself, is among the sources of information to infer the intent of user queries in this work. More details on the set of features used for this purpose will follow in the next section.

3.4.2 Features Set and Inference

As noted by Broder *et al.* (2007), queries are often short, and therefore a query by itself may not reveal much about a user. To determine query intent, the information obtained from a query is enhanced with respect to the search engine result page (SERP) returned for it. For instance, if a query such as “cheap shoes” is entered by the user, the appearance of keywords like “buy”, “free”, and “shipping” in the search result content may indicate the commercial nature of the query.

This approach can be seen similar to the idea used in the previous section, where particular terms appear to be correlated with more ad clicks and therefore with the commercial queries. However, ad click information is not used in this approach as the features are passed to a binary classifier in order to learn the contribution of each term towards the query intent. Particular terms appearing in a SERP can indicate that their corresponding query falls under a particular query intent category. Hence, these SERP features augment features extracted from the query itself (e.g. the number of characters and words in the query string).

Table 3.4: Features set.

Category	Feature	Description
Query Specific	Query length	Number of characters in the query string
	Query segments	Number of words in the query string
	URL-element	Whether the query string has any URL elements, such as .com or .org
	Organic domain	Total number of domains listed among the organic results of which the query string is a substring
Result Content	SERP	Weights vector for the terms extracted from the first result page displayed

This feature set is used to train five binary classifiers to recognize different dimensions of commercial intent as follows: i) commercial/ noncommercial category, ii) navigational/ informational category, iii) product category, iv) brand category, and v) retailer category. The first two dimensions are referred as the *major dimensions of query intent* while the last three are considered as the *sub-categories of commercial intent* throughout the work. Descriptions of features are presented in Table 3.4. Query-specific features have been extracted from the query strings and also from the content of search engine result pages returned for them.

To create the SERP features, each query was submitted to the originating search engine, and the first result page for that query was downloaded. In our prior work (Ashkan and Clarke, 2009; Ashkan *et al.*, 2009a), a number of the most frequent terms were selected from the SERPs to form the feature set. As *commercial intent* with respect to the *brand*, *product*, and *retailer* categories forms a hierarchical structure, the problem of inferring query intent in this work is addressed through the hierarchical classification described for the Web content by Dumais and Chen (2000).

The current setting varies from that of Dumais and Chen (2000) with respect to the content of the Web information used as source of the features. They use a short summary of a Web page created from the page title, the keywords, and either the description tag if it existed or the first 40 words of the body otherwise. Whereas, we consider the caption of the organic results displayed on the first result page as a short summary of the query for which this page (i.e. SERP) is returned by the search engine. Terms from the title and snippet parts of the captions were extracted considering white space and punctuation to separate the terms. These terms were then case-normalized by translating upper case to lower case.

We emphasize that the terms are originally extracted from the organic results only. We removed ads and other commercial content to avoid the obvious bias that ad terms might produce in the classification.

As for the feature selection, a held-out set for each category is considered, where for each term w in the set, the *mutual information*(Cover and Thomas, 2006; Dumais and Chen, 2000; Malik *et al.*, 2011) between w and the category l is computed as follows:

$$MI(w,l) = \sum_{F \in \{w, \bar{w}\}} \sum_{L \in \{l, \bar{l}\}} P(F,L) log \frac{P(F,L)}{P(F)P(L)}$$

where $P(F, L)$ represents the frequency of word w appearing (i.e. $F = w$) or not appearing (i.e. $F = \bar{w}$) inside (i.e. $L = l$) or outside (i.e. $L = \bar{l}$) the category l.

$P(F)$ is the frequency of w appearing (i.e. $F = w$) or not appearing (i.e. $F = \bar{w}$) in the collection, while $P(L)$ represents the popularity of the positive (i.e. $L = l$) or negative (i.e. $L = \bar{l}$) samples of the category in the collection. The above weighting of a word represents its discriminative information across the classes of the target category (Malik *et al.*, 2011).

For each category, the top 7000 features with the largest mutual information with respect to that category are selected as the features used in the classification. Each selected feature w with respect to a SERP s is then weighted using the BM25 term weight (Sparck Jones *et al.*, 2000):

$$\frac{TF(w,s) \times (k_1 + 1)}{k_1 \times [(1-b) + (b \times \frac{|s|}{avgsl})]}$$

where $TF(w, s)$ is the number of occurrences of the term w in s, $|s|$ is the length of the SERP s in terms, and $avgsl$ is the average SERP length; k_1 and b are parameters chosen as $k_1 = 2$ and $b = 0.75$, based on typical values found in previous work (Sparck Jones *et al.*, 2000). Each SERP is then represented as a set of BM25 weights of the terms selected for the corresponding category.

Using the set of 4000 queries labeled through Mechanical Turk, along with the features extracted from queries and SERPs, we trained classifiers for the two main dimensions of query intent: commercial/ noncommercial and navigational/ informational. Two SVM binary classifiers were trained on the extracted features, one based on the commercial/ non-commercial labels and the other based on the navigational/ informational labels. We used the SVM^{light} package (Joachims, 2008) for this purpose, employing 10-fold cross validation to measure the accuracy of each classifier.

For the sub-categories of commercial intent we focused on a total of 2010 queries from the entire set of labeled queries that were labeled as commercial. The presumed intent of these queries is to make an immediate or future purchase of a product or service. As mentioned previously, these queries were further labeled along three dimensions: in terms of their product, brand, and retailer specificity. For instance, the query "United Airlines ticket" is assumed to be product-specific and brand-specific, but with an unknown retailer since a United Airlines ticket can be purchased from many different travel services.

As also indicated previously, the levels of agreement for the queries labeled into the three specific sub-categories were not found to be as strong as the ones in the two major categories. Therefore, in order to have a balance between the accuracy of labeling and the size of the training set, the set of labeled queries with the agreement level of either 5-0 or 4-1 has been chosen as the main training set in order to train the three binary classifiers for the specific sub-categories of commercial intent.

3.5 Discussion of Inference Results

During training the classifiers based on various settings of features, the prediction accuracy was calculated for each classifier using 10-fold cross validation. A report of the prediction accuracy for the classifiers is presented in Table 3.5. These results are reported for two cases in each dimension: i) the classifier is trained based on SERP and query features combined, ii) the classifier is trained in presence of only SERP features. The performance measures for the three classifiers corresponding to sub-categories of commercial intent are also presented towards the end of the table.

Table 3.5: Precision, recall, and accuracy in percentage forms for Commercial/ Noncommercial classifier, Navigational/ Informational classifier, and the three classifiers corresponding to the sub-categories of commercial intent with respect to the SERP and query features combined or to the SERP features only.

Query Intent	**Features**	**Precision**	**Recall**	**Accuracy**
Commercial/ Noncommercial	SERP + Query	88.00	85.99	88.32
	SERP	87.89	84.71	87.76
Navigational/ Informational	SERP + Query	86.80	87.28	85.65
	SERP	83.18	85.48	82.59
Specific Product	SERP + Query	83.05	90.98	84.09
	SERP	82.95	89.67	83.41
Specific Brand	SERP + Query	85.31	90.37	84.54
	SERP	81.39	91.64	82.00
Specific Retailer	SERP + Query	89.29	83.15	87.04
	SERP	87.18	81.18	85.14

Considering query-specific features combined with the SERP features appears to slightly improve the quality of the classifiers as opposed to the case where only SERP features are used. Overall, the agreement obtained among the annotators along the different dimensions of query categories, and the accuracies obtained from the trained query classifiers, confirm that these categories are reasonably distinguishable. Features extracted from the query string, along with the contents of the search engine result pages, are found to be effective in detecting query intent.

While terms with respect to their click history in sponsored search appear to be effective in detecting the commercial intent of queries that include such terms, using click log is costly. As one of the main objectives of this research work is to study the characteristics of various query intents in terms of their impact on predicting clickthrough and user preferences, we avoid the possible circularity the use of ad click features would create in this process.

Classifiers trained from the combination of SERP and query features are used to predict the intent underlying queries in sets $B^{(1)}$, $C^{(1)}$, $A^{(2)}$, and $B^{(2)}$ across the studied dimensions of query intent: i) whether each query is either primarily commercial or primarily noncommercial, ii) whether each query is either mostly navigational or primarily informational, and iii) whether each commercial query targets a specific product, brand and/or retailer. Queries from these four sets, their predicted intents, and the corresponding ad click information form the targets of analyses in the remaining of the thesis.

Chapter 4

Characterizing Search Result Page Context in Sponsored Search

Most existing efforts studying ad clickthrough for sponsored search do so through factors related to ad content, landing pages, and bid terms. In a recent study by Fan and Chang (2010), the authors address the problem of incorporating context into content-based ad placement strategies on blog pages. They argue that ads that conflict with the negative orientation of a blog page are less likely to result in clickthroughs. Hence, even if the ad's content matches with the content of a blog page, it should not be displayed on the page that mostly discusses the corresponding commercial product or service from a negative point of view.

While the focus of this thesis work is on ads placed on search engine result pages, it is hypothesized that the context in which an ad is shown has an impact on the clickthrough rate of the ad. This context can include the positioning of the ad on the result page and the user intent underlying the queries with which the ad appears.

Bringing query intent detection into the context of sponsored search may help advertisers to create more appropriate and relevant ad content, and develop better ranking algorithms by matching the content of ads with the user's query intent, as well as contribute to the general understanding of user intent inference and web search behavior modeling.

The characteristics of search result pages and their corresponding query categories are studied in this chapter with respect to aggregated user click behavior on advertisement links. As discussed in Chapter 2, ad clickthrough rate (CTR) is connected with a notion of ad quality, which can contribute to user satisfaction in sponsored search. CTR is empirically defined as the ratio of the total number of clicks recorded for an ad to the total number of impressions of the ad (AdWords, 2013; Yahoo! Search Marketing, 2013). We adapt this definition to the context of search engine result pages (SERPs), referred to as the *context CTR*, in order to evaluate the performance of SERPs with respect to the various contextual factors studied in this thesis. In other words, the ratio of the total number of clicks recorded for SERPs that have a particular group of contextual factors in common to the total number of appearances of such SERPs is referred to as the context CTR for this group of contextual factors.

The current chapter aims at providing insights about the relationship between ad clickthrough behavior, the characteristics of queries, and the corresponding sponsored results shown on a search result page. In general, the placement of ads appears to have a substantial impact on the number of clicks they receive. This impact is more obvious when the intent underlying the queries for which those ads are displayed is also taken into account. In addition, it is shown that clickthrough behavior is consistent with the query intent identified through classification.

4.1 Impact of Ad Position on Clickthrough

The context CTR for SERPs with particular number of ads can be estimated using the click and SERP information obtained from a search engine's log data. Given the log data $\mathcal{D}$, its SERP information can be sorted according to the number of ads displayed for each page. Let us assume the number of ads on pages varies from 1 to N^{max}, ignoring pages with no ad, for which there can be no ad clicks. Thus, result pages are divided into N^{max} groups, each denoted as set $\mathcal{D}_i$, where i is the number of displayed ads for the pages in that set.

Each SERP in set $\mathcal{D}_i$ is identified by a unique identifier which associates the SERP with its click information stored in the click log. Given the j^{th} SERP in $\mathcal{D}_i$, define c_i^j to represent whether there is an ad click resulting from this page. In other words, $c_i^j = 1$, if there is an ad click associated with this page in the click data, and $c_i^j = 0$ otherwise.

The probability of click for a SERP with $N = i$ ads displayed is denoted by $P(C = 1|N = i)$. This probability can be estimated as the average number of ad clicks per context for SERPs with a particular number of ads i, and it is referred to as the *context CTR*:

$$P(C = 1|N = i) = \frac{\sum_{j=1}^{|\mathcal{D}_i|} c_i^j}{|\mathcal{D}_i|} \qquad 1 \leq i \leq N^{max} \tag{4.1}$$

where $|\mathcal{D}_i|$ indicates the number of SERPs with i ads displayed.

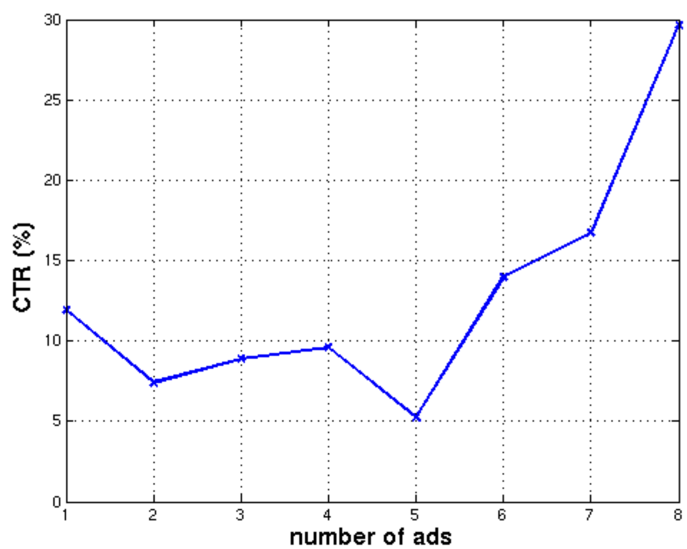


Figure 4.1: Average CTR for SERPs with particular number of ads

These probability values are estimated empirically using the data set described in Section 1.3. We begin by examining the average number of clicks per context for SERPs with a particular number of ads from set $B^{(1)}$. The unique id number for each SERP (the SERP id) is used to determine whether the display of the SERP resulted in an ad click. In this data set, the maximum number of ads displayed on the SERPs is eight. Thus, SERPs are divided into eight groups according to Equation 4.1, i.e. $1 \leq N \leq N^{max}$ and $N^{max} = 8$.

The context CTR for the eight ad-based groups of set $B^{(1)}$ is then calculated, resulting in the plot depicted in Figure 4.1. For clarity of presentation, the points for each particular number of ads are connected. The lines do not imply interpolation. It is noted that these numbers do not represent the estimated CTR values for an individual ad on a result page. They are empirical estimations of the context CTR, representing the likelihood of click for SERPs with particular number of ads. Generally speaking, the more ads displayed on a result page, the more likely an ad will be clicked. This observation could indicate that the number of ads (in part) determines the number of ad clicks on result pages, or that ads are more likely to be displayed on commercially oriented pages.

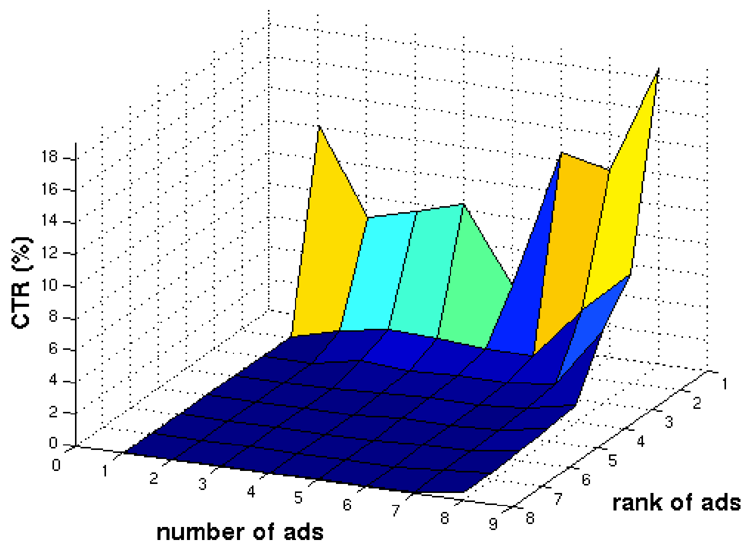


Figure 4.2: Average CTR at specific ranks for SERPs with particular number of ads.

Figure 4.2 shows the same CTR trend when the rank of the clicked ads is also considered:

$$P(C = 1, R = r | N = i) = \frac{\sum_{j=1}^{|\mathcal{D}_i|} c_{i,r}^j}{|\mathcal{D}_i|} \qquad 1 \leq i \leq N^{max} \tag{4.2}$$

where $c_{i,r}^j$ represents whether there is an ad click at rank r for the j^{th} SERP in $\mathcal{D}_i$. The z–axis in Figure 4.2 indicates the context CTR values, $P(C = 1, R = r|N = i)$, for varying rank positions, r, across different numbers of displayed ads, i.

Note that ad clicks mostly occur at the first and the second ranks, and most especially at the first rank. This observation confirms that the chance of ad click decreases as ads are displayed in lower ranks on result pages, possibly as a result of reduced visual attention (Richardson *et al.*, 2007).

4.2 Impact of Ad Location on Clickthrough

In addition to the rank position of ads, their location (top or side of a result page) could affect clickthrough rates. According to Jansen (2007), top-listed ads are often assumed to be more relevant than organic results and side-listed ads. This could affect the frequency of clicks for ads at different locations on result pages. In this regard, it is hypothesized that peaks and valleys in the two plots (Figures 4.1 and 4.2) could arise from the location of different ads for which the clicks are recorded.

In order to study the impact of ad location on its clickthrough, we use an empirical analysis over the logs of sponsored search data explained in Section 1.3. A batch of approximately $43,000$ queries, randomly selected from sets $A^{(1)}$, $B^{(1)}$, and $C^{(1)}$, was submitted to the originating search engine in order to study different possibilities in terms of the number of ads and their locations on the search engine result pages. The results are shown in Table 4.1 in percentage form.

According to this experiment, and consistent with the understanding of the dataset, the maximum number of ads displayed at the right side of a search result page is 5, and the maximum number of ads displayed on top of a result page is 3. It is also assumed the rank of ads are assigned in a way that ads displayed on top (if any) are ranked higher than the ones displayed at the side (Jansen, 2007). 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 is assumed to be: a_{t_1}, a_{t_2}, and a_{s_1}.

The available log data does not record the precise locations of ads, but does supply a

Table 4.1: Percentage of SERPs with particular number of ads on the top and at the side from the results obtained through an experiment on 43K queries.

	$t = 0$	$t = 1$	$t = 2$	$t = 3$
$s = 0$	38.1646	3.0684	0.6890	0.2088
$s = 1$	5.2878	2.5745	0.8734	0.4299
$s = 2$	2.9655	2.0791	0.8719	0.6288
$s = 3$	1.8414	1.6584	0.8612	0.7926
$s = 4$	1.2004	1.1829	0.7515	0.9489
$s = 5$	2.6972	5.2390	3.7437	21.2411

rank for ads, along with the total number of ads displayed on a page. A single ad may appear either on the top or at the side. For instance, if there are three ads, all three may be positioned at the top, or two may be at the top and one at the side, or one may be at the top and two on the side, or all three may appear on the side. Using this information, and the existing constraints on ad placement in the data set, which indicates that no more than three ads may appear on the top and no more than five on the side, we estimate the click probability for different locations of the page. Two probabilistic approaches are proposed for this purpose.

The first approach is based on an assumption indicating that clicking on a top ad is *independent* of the ads displayed at the side of a result page. The second approach, on the other hand, does not require us to make such an assumption; it estimates the click values for different locations by maximizing the entropy of the click probability distribution over the SERPs.

In both approaches, R is a random variable characterizing the distribution of possible ranks of ads at which clicks occur, and N represents the total number of ads displayed on a result page. The main objective here is to estimate the context clickthrough rate at rank r for a varying number of ads at different locations on a SERP.

4.2.1 Clickthrough Estimation for Various Locations – The Primary Attempt

Let $P(C = 1, R = r|N = n)$ denote the average clickthrough rate for a result page on which a total of $N = n$ ads are displayed, and the ad at rank $R = r$ is clicked. Let N_t and N_s be the number of displayed ads on the top and at the side of the result page, so that $N = N_t + N_s$. The probability of appearance of $N_t = t$ ads on the top and $N_s = s$ ads at the side of a result page conditioned on the total number of ads, $N = n$, displayed on the result page is $P(N_t = t, N_s = s|N = n)$.

Note that each cell in Table 4.1 represents the likelihood of SERPs with $N_t = t$ ads on the top and $N_s = s$ ads at the side (i.e. $P(N_t = t, N_s = s)$). Thus, the above conditional probability can be calculated from Table 4.1 as follows:

$$\begin{aligned} P(N_t = t, N_s = s|N = n) &= \frac{P(N_t = t, N_s = s, N = n)}{P(N = n)} \\ &= \frac{P(N_t = t, N_s = s)}{P(N = n)} \end{aligned} \tag{4.3}$$

where $s + t = n$ and $P(N = n)$ can be calculated by the summation of the corresponding probabilities in Table 4.1 (e.g. $P(N = 1) = P(N_t = 0, N_s = 1) + P(N_t = 1, N_s = 0)$).

The current approach is based 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.

The derivations are presented for the first rank, noting that everything can be repeated similarly for the other rank positions. The average CTR at the first rank for varying number of ads can be seen as $P(C = 1, R = 1|N = n)$ for all possible values of n (i.e. $1 \leq n \leq 8$). In order to estimate CTR with respect to the location of ads (top or side) on the page, $P(C = 1, R = 1|N = n)$ values from Figure 4.2 and the estimation of ad's positional distribution on result pages from Table 4.1 are employed.

Given $N = 8$, according to Table 4.1, there is only one possibility where $N_s + N_t = 8$, that is $N_t = 3$ and $N_s = 5$. Hence, $P(N_t = 3, N_s = 5|N = 8) = 1$, and therefore $P(C = 1, R = 1|N = 8)$ is estimated as follows:

$$
\begin{aligned}
& P(C = 1, R = 1|N = 8) \\
= \quad & P(C = 1, R = 1|N_t = 3, N_s = 5)P(N_t = 3, N_s = 5|N = 8) \\
\overset{(a)}{=} \quad & P(C = 1, R = 1|N_t = 3) \times 1 \\
\Rightarrow \quad & P(C = 1, R = 1|N_t = 3) = P(C = 1, R = 1|N = 8) = 0.19
\end{aligned} \tag{4.4}
$$

where the substitution of $P(C = 1, R = 1|N_t = 3, N_s = 5)$ by $P(C = 1, R = 1|N_t = 3)$ in (a) comes from Assumption 1, and $P(C = 1, R = 1|N = 8) = 0.19$ comes from Figure 4.2.

Similarly, for SERPs with $N = 7$ ads displayed, there are two possibilities according to Table 4.1: i) $N_t = 3$, $N_s = 4$ and ii) $N_t = 2$, $N_s = 5$. Hence, according to Equation 4.3, $P(N_t = 3, N_s = 4|N = 7)$ and $P(N_t = 2, N_s = 5, |N = 7)$ can be estimated as 0.2 and 0.8 respectively and used to write $P(C = 1, R = 1|N = 7)$, as follows:

$$
\begin{aligned}
& P(C = 1, R = 1|N = 7) \\
= \quad & P(C = 1, R = 1|N_t = 3, N_s = 4)P(N_t = 3, N_s = 4|N = 7) + \\
& P(C = 1, R = 1|N_t = 2, N_s = 5)P(N_t = 2, N_s = 5|N = 7) \\
= \quad & P(C = 1, R = 1|N_t = 3) \times 0.2 + P(C = 1, R = 1|N_t = 2) \times 0.8 \\
= \quad & (0.19 \times 0.2) + [0.8 \times P(C = 1, R = 1|N_t = 2)] \\
\Rightarrow \quad & P(C = 1, R = 1|N_t = 2) = \frac{1}{0.8} \times (P(C = 1, R = 1|N = 7) - 0.038) \\
\Rightarrow \quad & P(C = 1, R = 1|N_t = 2) = 0.103
\end{aligned} \tag{4.5}
$$

where the value of $P(C = 1, R = 1|N = 7)$ comes from Figure 4.2. Similarly, for $N_t = 1$, the average clickthrough rate can be estimated as:

$$
P(C = 1, R = 1|N_t = 1) = 0.115 \tag{4.6}
$$

Using the estimated values of $P(C = 1, R = 1|N_t = t)$ (for $1 \leq t \leq 3$) from Equations 4.4, 4.5, and 4.6 and also the values of $P(C = 1, R = 1|N = n)$ (for $1 \leq n \leq 5$)

plotted in Figure 4.2, the context CTR at the first rank for cases, where no ads are displayed on the top and at least one ad is displayed at the side, can be estimated. In other words, $P(C=1,R=1|N_t=0,N_s=s)$ is estimated for $1 \leq s \leq 5$, while the total number of ads displayed on a result page is at most 5 ($1 \leq n \leq 5$):

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

The estimated values $P(C=1,R=1|N_t=t)$ (where $1 \leq t \leq 3$) and $P(C=1,R=1|N_t=0,N_s=s)$ (where $1 \leq s \leq 5$) are referred to as the *adjusted values* at the first rank. Due to Assumption 1 which simplifies the calculations, most of the values corresponding to lower rank positions are estimated as negative numbers. For this reason, only the adjusted values at the first rank along with the corresponding values at all ranks are plotted in Figure 4.3 in percentage form. At each point, the adjusted value at all ranks is calculated by the summation of the corresponding values at different ranks (estimated in a similar way as the first rank). The negative values have been treated as zero in calculating the values at all ranks.

Due to the limitations of this approach caused by the simplifying Assumption 1, another approach is proposed to estimate the clickthrough values at different locations of result pages. The following approach presents a probabilistic model for the placement of ads at different locations (top/ side) of a result page where no such assumption is required in order to simplify the analysis. As explained below, the clickthrough values for different locations are estimated by solving an optimization problem.

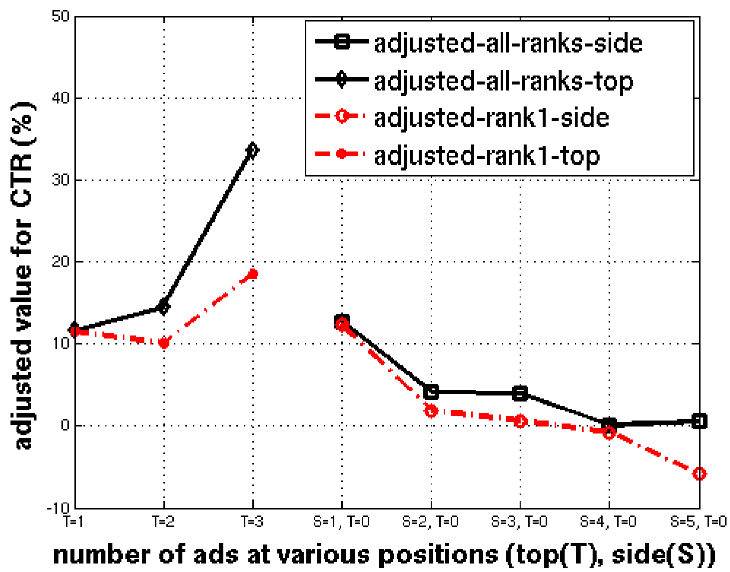


Figure 4.3: Approach 1 – Adjusted plots for average CTR for SERPs with a particular number of ads on top/ side of the page.

4.2.2 Clickthrough Estimation for Various Locations – Maximizing Entropy

For a result page with $N_t = t$ ads displayed on the top and $N_s = s$ ads displayed at the side, summing the likelihood of no click (i.e. $P(C = 0|N_t = t, N_s = s)$) and the likelihood of a click on any ad at rank r ($P(C = 1, R = r|N_t = t, N_s = s)$) where $1 \leq r \leq n$ and $n = t + s$ is obviously 1. In other words, the following equation holds:

$$
\begin{aligned}
&\sum_{c\in\{0,1\}} \sum_{1\leq r\leq s+t} P(C = c, R = r|N_t = t, N_s = s) = 1 \\
\Rightarrow \quad &\sum_{c\in\{0,1\}} \sum_{1\leq r\leq s+t} \frac{P(C = c, R = r, N_t = t, N_s = s)}{P(N_t = t, N_s = s)} = 1 \\
\Rightarrow \quad &\sum_{c\in\{0,1\}} \sum_{1\leq r\leq s+t} P(C = c, R = r, N_t = t, N_s = s) = P(N_t = t, N_s = s)
\end{aligned}
\tag{4.7}
$$

where $P(N_t = t, N_s = s)$ is estimated from the cells in Table 4.1, which represents the likelihood of SERPs with $N_t = t$ ads on the top and $N_s = s$ ads at the side. $P(C, R, N_t, N_s)$ is defined over $C \in \{0, 1\}$ and $R \in \{1, ..., n\}$, $n = t + s$, $0 \leq t \leq 3$, $0 \leq s \leq 5$, and it is described as the likelihood of a SERP resulting in no click or click at rank r when t and s ads are displayed on the top and at the side of the page respectively.

The main objective in this section is to estimate the average clickthrough rate at rank r, for varying number of ads at different locations on a SERP. This value can be viewed as the conditional probability $P(C = 1, R = r|N_t = t, N_s = s)$, which can be derived from $P(C, R, N_t, N_s)$ as follows:

$$P(C = 1, R = r|N_t = t, N_s = s) = \frac{P(C = 1, R = r, N_t = t, N_s = s)}{P(N_t = t, N_s = s)} \tag{4.8}$$

In order to calculate $P(C = 1, R = r|N_t = t, N_s = s)$ in Equation 4.8 for all values of s and t, $P(C, R, N_t, N_s)$ has to be first solved for all possible values of s and t. This distribution is produced from user interactions (i.e. click at $r \geq 1$ or no click) on various context of a result page. In order to cope with various possibilities caused by the dynamic nature of human interaction, a solution with maximum randomness may be reasonable. Entropy (Cover and Thomas, 2006), as a measure of randomness and uncertainty, can be maximized in order to obtain a stable state of the system as an answer for $P(C, R, N_t, N_s)$. Hence, an optimization problem is defined in order to maximize the entropy, $H(P(C, R, N_t, N_s))$, subject to:

1. $\sum_{c\in\{0,1\}} P(C = c, R = r, N = n) = \sum_{c\in\{0,1\}} P(C = c, R = r|N = n)P(N = n)$
2. $\sum_{c\in\{0,1\}} P(C = c, R = r, N = n) = \sum_{c\in\{0,1\}} \sum_{s+t=n} P(C = c, R = r, N_t = t, N_s = s)$
3. $\sum_{c\in\{0,1\}} \sum_{0\leq r\leq s+t} P(C = c, R = r, N_t = t, N_s = s) = P(N_t = t, N_s = s)$
4. $0 \leq P(C = c, R = r, N_t = t, N_s = s) \leq 1$
5. $\sum_{c\in\{0,1\}} \sum_{r,t,s} P(C = c, R = r, N_t = t, N_s = s) = 1$

where $c \in \{0,1\}$, $r \in \{1,...,n\}$, $n = t+s$, $0 \leq t \leq 3$, and $0 \leq s \leq 5$. In the first line of the constraints, $P(N = n)$ can be calculated by the summation of the corresponding values from Table 4.1. As an example, $P(N = 1) = P(N_t = 0, N_s = 1) + P(N_t = 1, N_s = 0)$. Moreover, $P(C = c, R = r|N = n)$ for all possible values of n (i.e. $1 \leq n \leq 8$) can be estimated through Figure 4.2, which represents the average clickthrough rate at different ranks for varying number of ads.

The CVX Optimization Environment for Matlab (Grant and Boyd, 2009) has been used to solve this problem, estimating $P(C = 1, R = r|N_t = t, N_s = s)$ for different values of r, t, and s. The value for $P(C = 1, R = r|N_s = s, N_t = 0)$ may be computed by substituting $t = 0$ in Equation 4.8. Consequently, $P(C = 1, R = r|N_t = t)$ for $t \geq 1$ can be estimated as follows:

$$P(C = 1, R = r|N_t = t) = \sum_{s=0}^{5} P(C = 1, R = r|N_t = t, N_s = s)P(N_s = s|N_t = t) \quad (4.9)$$

where $P(N_s = s|N_t = t)$ is the probability of appearance of $N_s = s$ ads at the side of a result page conditioned on the number of top ads, t, displayed on the result page.

Note that each cell in Table 4.1 represents the likelihood of a SERP with $N_t = t$ ads on the top and $N_s = s$ ads at the side (i.e. $P(N_t = t, N_s = s)$). Also, the summation of the cells in a column represents the likelihood of a SERP with $N_t = t$ ads on the top (i.e. $P(N_t = t)$. Thus, the above conditional probability can be calculated from Table 4.1 as follows:

$$P(N_s = s|N_t = t) \quad = \quad \frac{P(N_t = t, N_s = s)}{P(N_t = t)} \quad (4.10)$$

Similar to the previous section, the estimated values obtained for $P(C = 1, R = r|N_s = s, N_t = 0)$ and $P(C = 1, R = r|N_t = t)$ are referred to as the *adjusted values* at rank r. The values calculated for the queries in set $B^{(1)}$ and corresponding to the clickthrough rate at rank 1, 2, 3, and 4 at different locations are plotted in Figure 4.4 in percentage form. The rates for other ranks (5 to 8) are not presented in the figure, as they are close to zero.

The trend of changes in the adjusted values can be used to explain the dips in Figures 4.1 and 4.2. According to Figure 4.4, the lowest estimated value at each rank is for the case

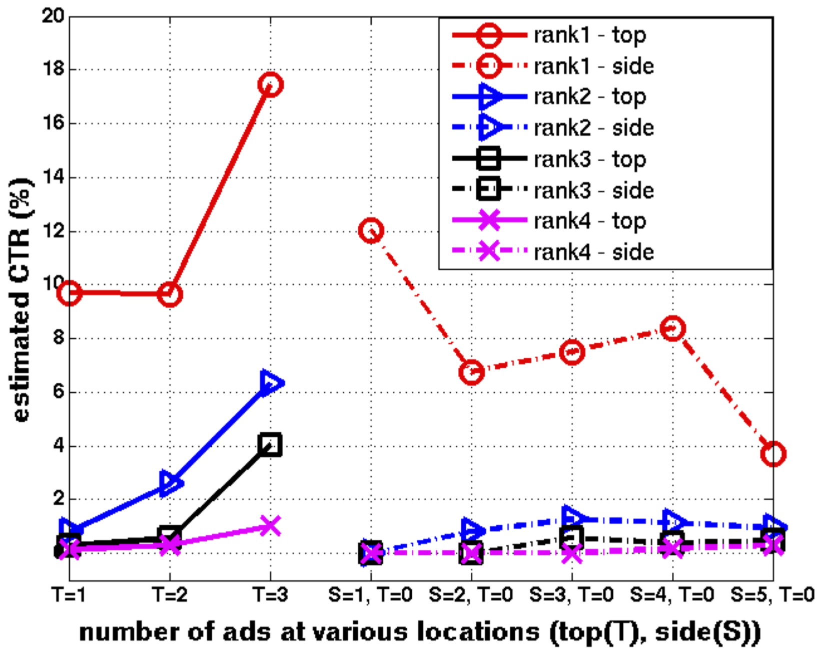


Figure 4.4: Approach 2 – Adjusted plots for average CTR for SERPs with a particular number of ads on top/ side of the page.

where there are 5 ads displayed at the side while no ad appears on the top. This observation can be viewed as a reason for the dip at 5 in Figures 4.1 and 4.2. As can be seen in the figure, the more ads displayed on the top, the more clicks they would receive ($N_t = 1$, $N_t = 2$, $N_t = 3$).

At each rank, comparing the values for top ads (the first three on the corresponding plot) and the ones with no top ad (the last five on the corresponding plot), we observe that ads on the top of a result page are more often the targets of clicks than the ads at the side. This observation will be further investigated when the intent underlying the query displayed on a result page is also taken into consideration.

4.3 Ad Clickthrough Behavior for Different Intents

Having determined the apparent intent underlying each query, a similar approach is taken to that of Equation 4.1, calculating the average clickthrough rate for all the SERPs with a particular number of ads. However, this time, only the result pages are considered for which the associated queries fall into a given class, that is $\mathcal{D}_i^g$ refers to the group of pages in the log that contain i ads and belong to the query class g:

$$P(C = 1|N = i, G = g) = \frac{\sum_{j=1}^{|\mathcal{D}_i^g|} c_i^j}{|\mathcal{D}_i^g|} \qquad 1 \leq i \leq N \tag{4.11}$$

where c_i^j represents whether there is an ad click recorded for the j^{th} SERP in $\mathcal{D}_i^g$. Here, $P(C = 1|N = i, G = g)$ represents the average clickthrough rate for the SERPs that correspond to the query class g and accommodate i ads.

The average clickthrough rate for the four possible combinations of major query classes in pairs (i.e., commercial- navigational, commercial- informational, noncommercial- navigational, and noncommercial- informational) against the number of ads are plotted in Figure 4.5. The plot from Figure 4.1 is also placed in Figure 4.5 in order to provide a basis for comparison. Note that the plots indicate the average clickthrough rate for SERPs with a particular number of ads and associated with particular classes of query intent.

It can be seen in Figure 4.5 that ad clickthrough behavior is distinct for different categories of query intent, and this can indicate that the clickthrough behavior is consistent with the classification results of the general categories of query intent as explained in Chapter 3. Generally speaking, categories that involve commercial intent are the leaders among the others (i.e. plots related to commercial-navigational and commercial-informational categories). This result confirms that the commercial categories of queries receive more ad clicks comparing to the others.

It can also be seen that commercial- navigational queries receive more ad clicks than commercial- informational queries on average. In other words, ads that reflect the intent of commercial-navigational queries seem to be more of a target for clicks than the

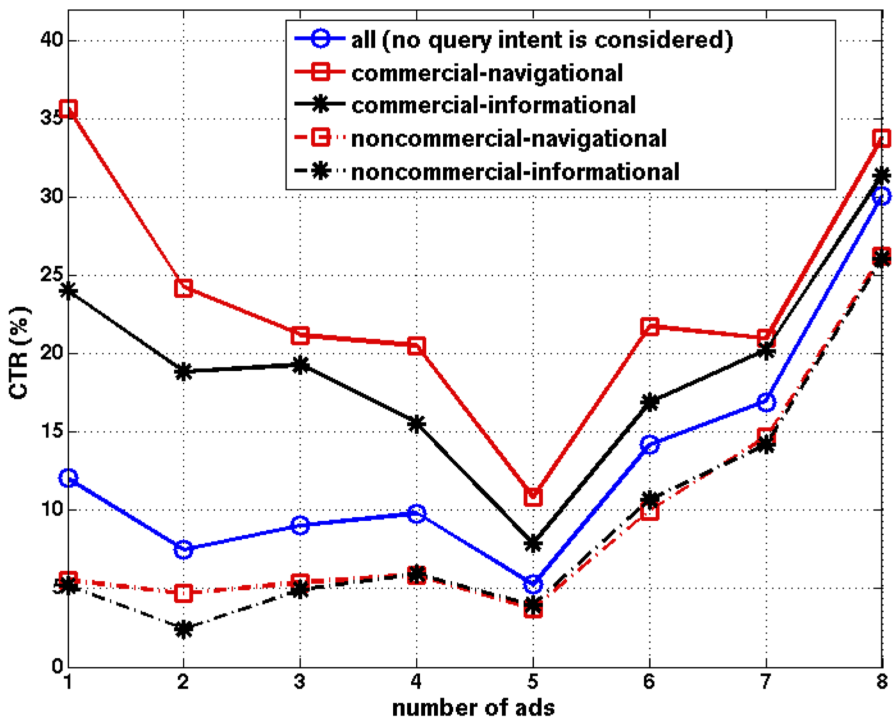
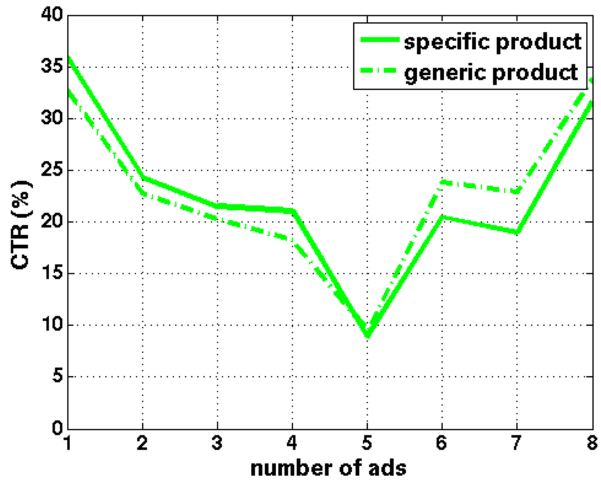


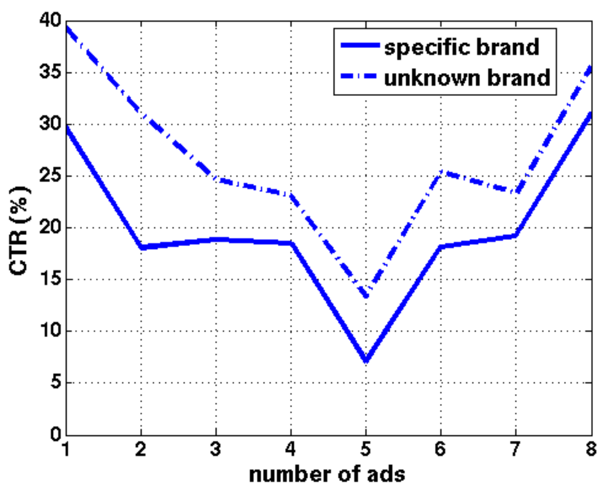
Figure 4.5: CTR for SERPs with particular number of ads and associated with various query types.

ones that reflect the intent of commercial-informational queries. An example illustrating the difference is “Westjet Airlines” as a commercial-navigational query against “airline tickets” as a commercial-informational query. From the retrieval perspective, the chance that a user would find a relevant ad for the former query is greater than the latter, because the former query is restricted by the airline name. From the user perspective, the commercial- navigational intent of the former query may indicate a relatively more focused and goal-directed search (Danaher and Mullarkey, 2003) (the user knows the retailer of the commercial product that they want), resulting in a higher chance of a click or conversion from the user.

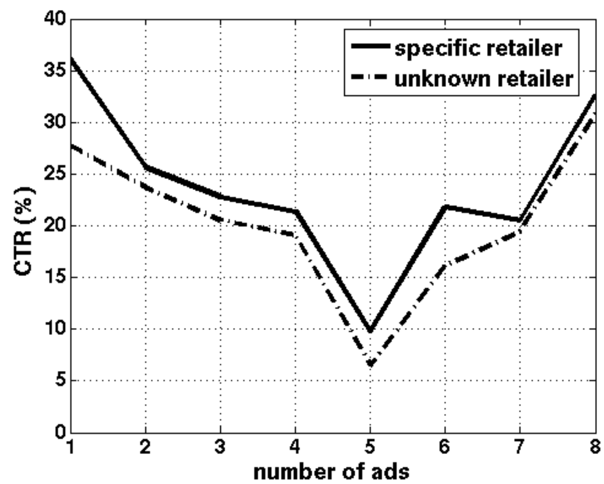
The average clickthrough rate for each sub-category of commercial intent is also depicted in Figure 4.6. It can bee seen that the distinction in clickthrough behavior exists



(a) Product



(b) Brand



(c) Retailer

Figure 4.6: CTR for SERPs with particular number of ads and associated with major query types and associated with sub-categories of commercial intent: Specific Product vs. Generic Product, Specific Retailer vs. Unknown Retailer, and Specific Brand vs. Unknown Brand.

for the sub-categories of commercial intent. Among the three sub-categories, only for the retailer category, when a specific retailer is implied by the query intent, the number of clicks for varying number of ads is always higher compared to the case where the retailer in unknown. In other words, ads are placed in a way that the ones that reflect retailer intent are more of a target of clicks than the others. This finding is consistent with that of Ghose and Yang (2008) for the impact of keyword attributes on consumer search and purchase behavior. Retailer specific queries are also navigational queries where the searcher is the so-called "loyal" customer most likely looking for an information about or a product from a particular retailer. However, searches on specific product or brand names could indicate that the searcher needs a commercial product or service, but doesn't yet know where to buy it, providing competitive search situations.

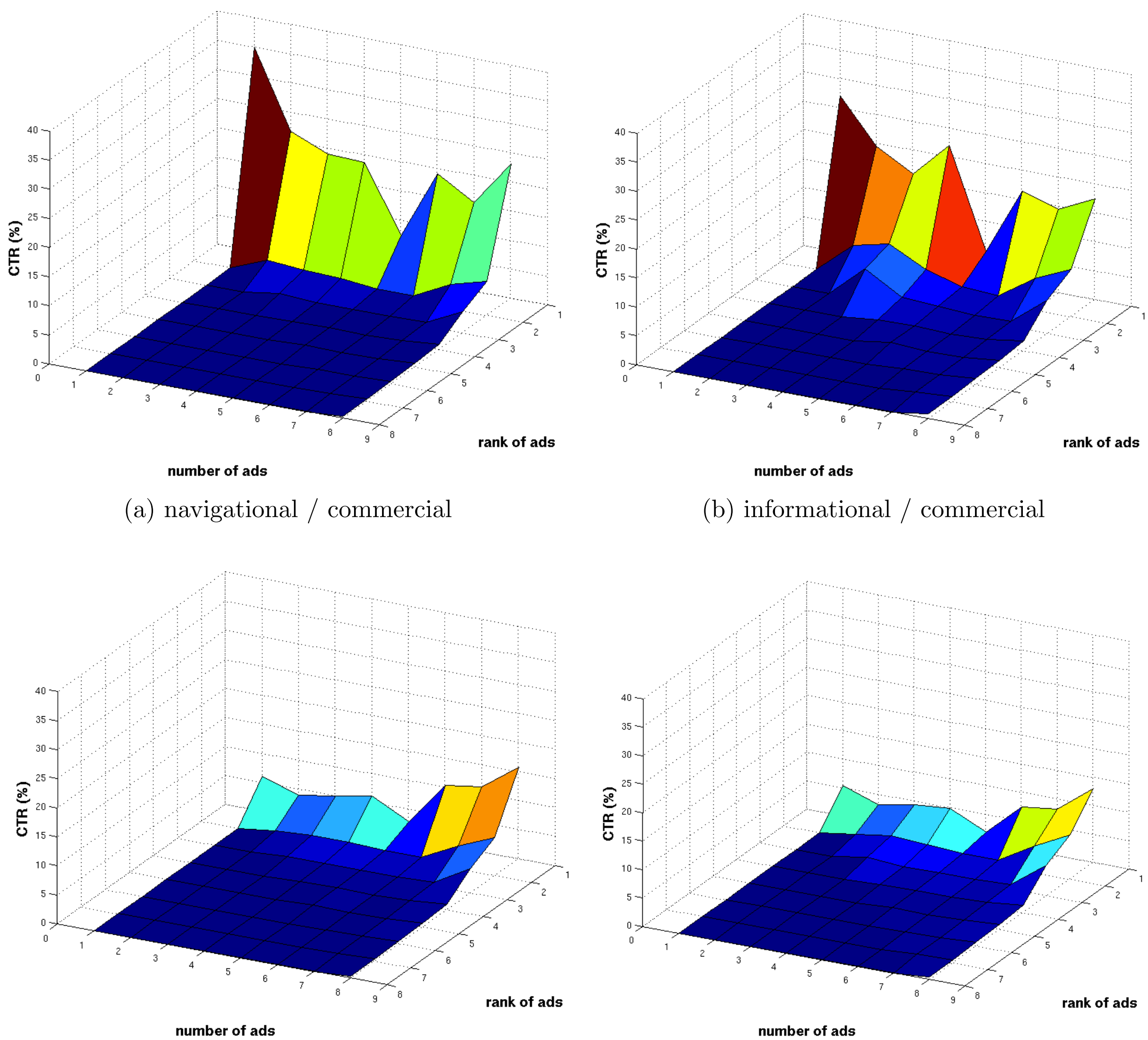
Even a loyal customer could be looking for better deals and providers when they have a specific product intent or brand intent underlying their query, which could be the source of competition. If an ad wins the click and the order, that implies such an ad has taken market share away from a competitor, resulting in an increase in the conversion rate not in the click rate (Ghose and Yang, 2008).

Furthermore, a similar calculation to that of Equation 4.2 can be performed in presence of query intent classes. The average clickthrough rate at rank r for SERPs in $\mathcal{D}_i^g$ (corresponding to query category g and accommodating i ads) can be therefore estimated as follows:

$$P(C=1, R=r|N=i, G=g) = \frac{\sum_{j=1}^{|\mathcal{D}_i^g|} c_{i,r}^j}{|\mathcal{D}_i^g|} \qquad 1 \leq i \leq N \tag{4.12}$$

where $c_{i,r}^j$ represents whether there is an ad click at rank r for the j^{th} SERP in $\mathcal{D}_i^g$. The z–axis in Figures 4.7 and 4.8 represent these CTR values. Figure 4.7 takes into account the impact of the major query categories besides the rank of click and the total number of ads displayed on result pages. Similarly, Figure 4.8 takes the impact of sub-categories of the commercial intent into account, in addition to the rank position and the number of displayed ads.

As can be seen in both figures, variability of clickthrough behavior across different



(a) navigational / commercial

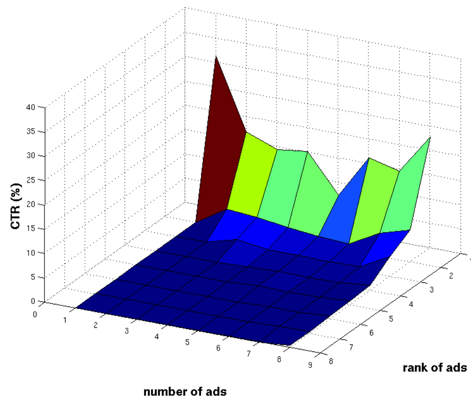
(b) informational / commercial

(c) navigational / non-commercial

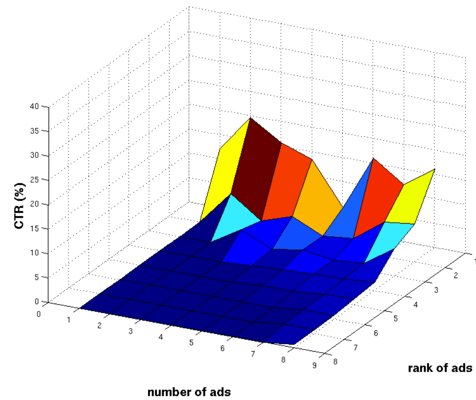
(d) informational / non-commercial

Figure 4.7: Average CTR at specific ranks for SERPs with particular number of ads and corresponding to different categories of query intent.

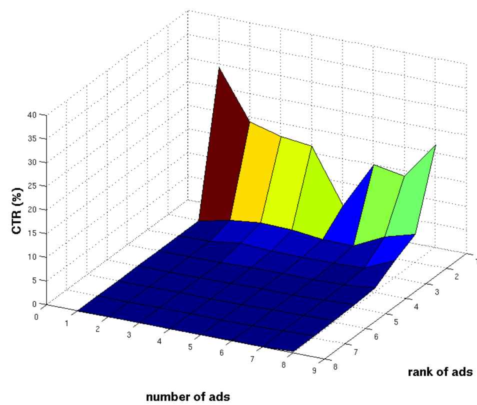
query categories is present. Ad clicks mostly appear at the first and the second ranks for



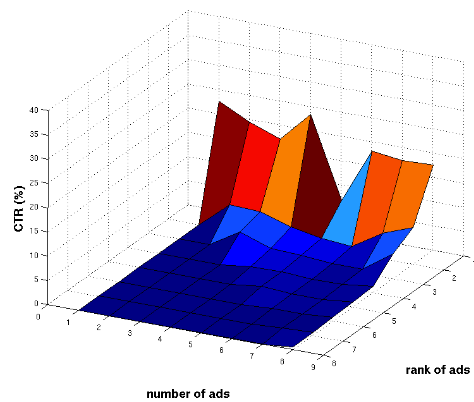
(a) navigational and specific product



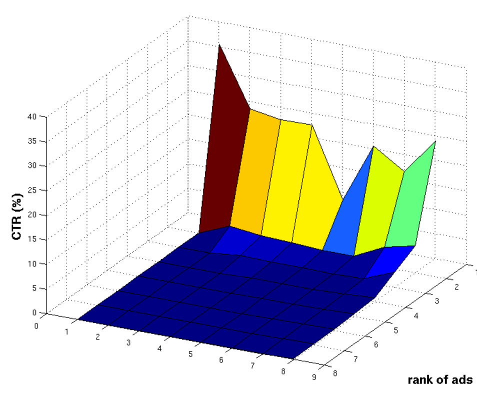
(b) informational and specific product



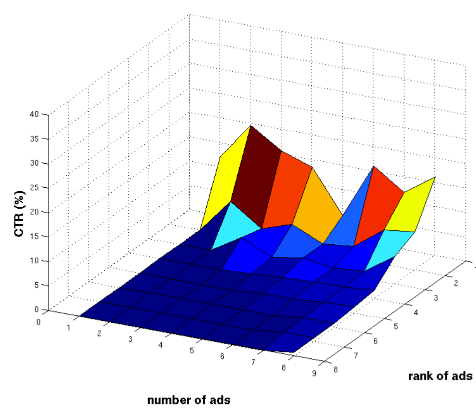
(c) navigational and specific brand



(d) informational and specific brand



(e) navigational and specific retailer



(f) informational and specific retailer

Figure 4.8: Average CTR at specific ranks for SERPs with particular number of ads and corresponding to sub-categories of commercial intent.

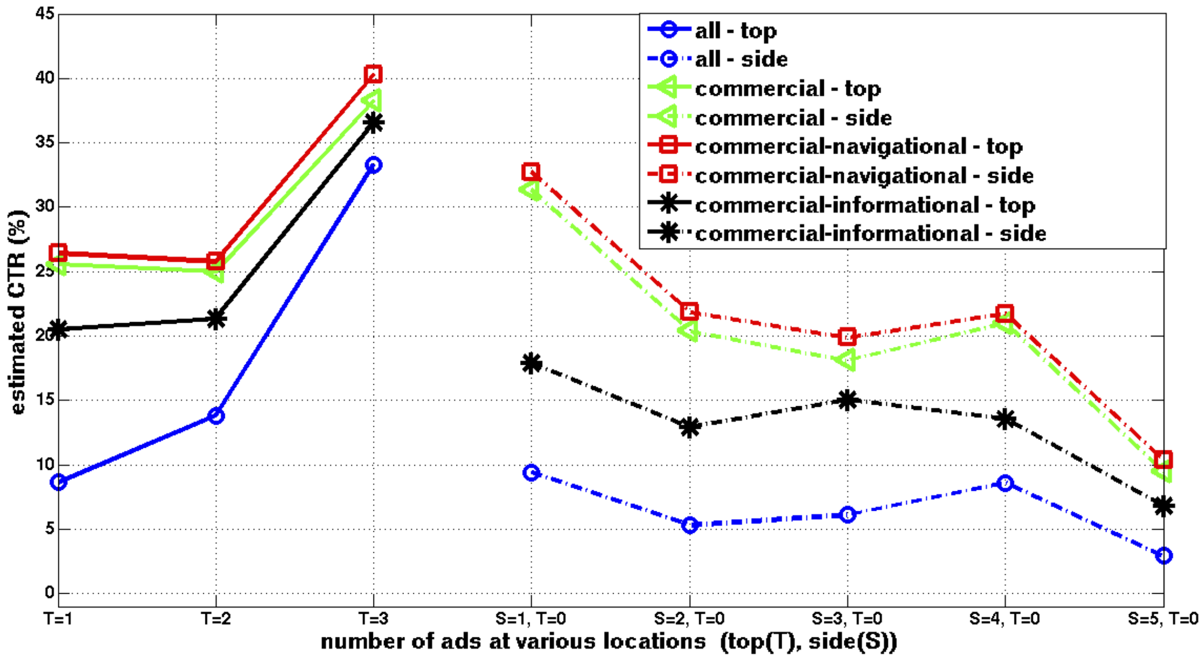


Figure 4.9: Adjusted plots for average CTR for SERPs corresponding to different query intents and with a particular number of ads on top/ side of the page.

all query intent categories, and most especially at the first rank. This observation confirms once again that the clickthrough rate of ads decreases as they are displayed in lower ranks on the page. The values plotted in these two figures will be used in the next section for further ad clickthrough analysis.

In order to study the impact of ad location on clickthrough for commercial intent, the adjusted plots for commercial, commercial- navigational, and commercial- informational categories, along with the case with no query intent are depicted in Figure 4.9. For all the plots, the dip at the point where no ad is displayed on the top and five ads are displayed at the side ($N_t = 0$, $N_s = 5$) indicates that for this specific placement of ads there are few clickthroughs, in turn explaining the dip at five in the corresponding plots of Figure 4.7.

The placement of ads, as depicted in Figure 4.9, appears to have a substantial impact on the number of clicks they receive for different intents. Similar to Figure 4.5, considering the navigational aspect of a commercial query versus the informational aspect of a commercial query appears to have substantial impact on the clickthrough rate. It can be seen that commercial-navigational queries tend to result in relatively higher clickthrough rates on ads placed at different locations on result pages, as compared to commercial-informational queries.

Also note that for all categories, the more ads displayed on the top, the more clicks they would receive ($N_t = 1$, $N_t = 2$, $N_t = 3$), however the difference between the clickthrough rates of top ads and side ads becomes lower when it comes to the leading query categories (i.e. commercial-navigational and commercial). This observation may indicate that when the intent underlying the query is commercial, the effect of the location of ads becomes less significant. However, ads on the top are still the main targets of clicks.

The average clickthrough rate for each sub-category of commercial intent along with the adjusted plots for the impact of location of ads on them is depicted in Figure 4.10. Once again, differences in clickthrough behavior is observable for the sub-categories of commercial intent.

4.4 Context-Based Click Analysis

In the existing methods of clickthrough analysis in sponsored search, it is primarily the content of ads which is used to study the user's click behavior on these links. In the current thesis, factors beyond those extracted from ad content form the targets of study. The intuition behind this idea is that ads positioned at the top of a page may receive more clicks, even if they are less relevant than other ads. Furthermore, a weakly related ad appearing with the results of a commercially oriented query may receive more clicks than a strongly related ad appearing with the results of a less commercially oriented query.

One of the main goals of this thesis is to study whether context based information can reflect a user's click behavior over the advertisement links and derive the ad clickthrough rate. Further efforts could combine this type of information with the ad content factors in order to improve the quality of the existing clickthrough prediction models in spon-

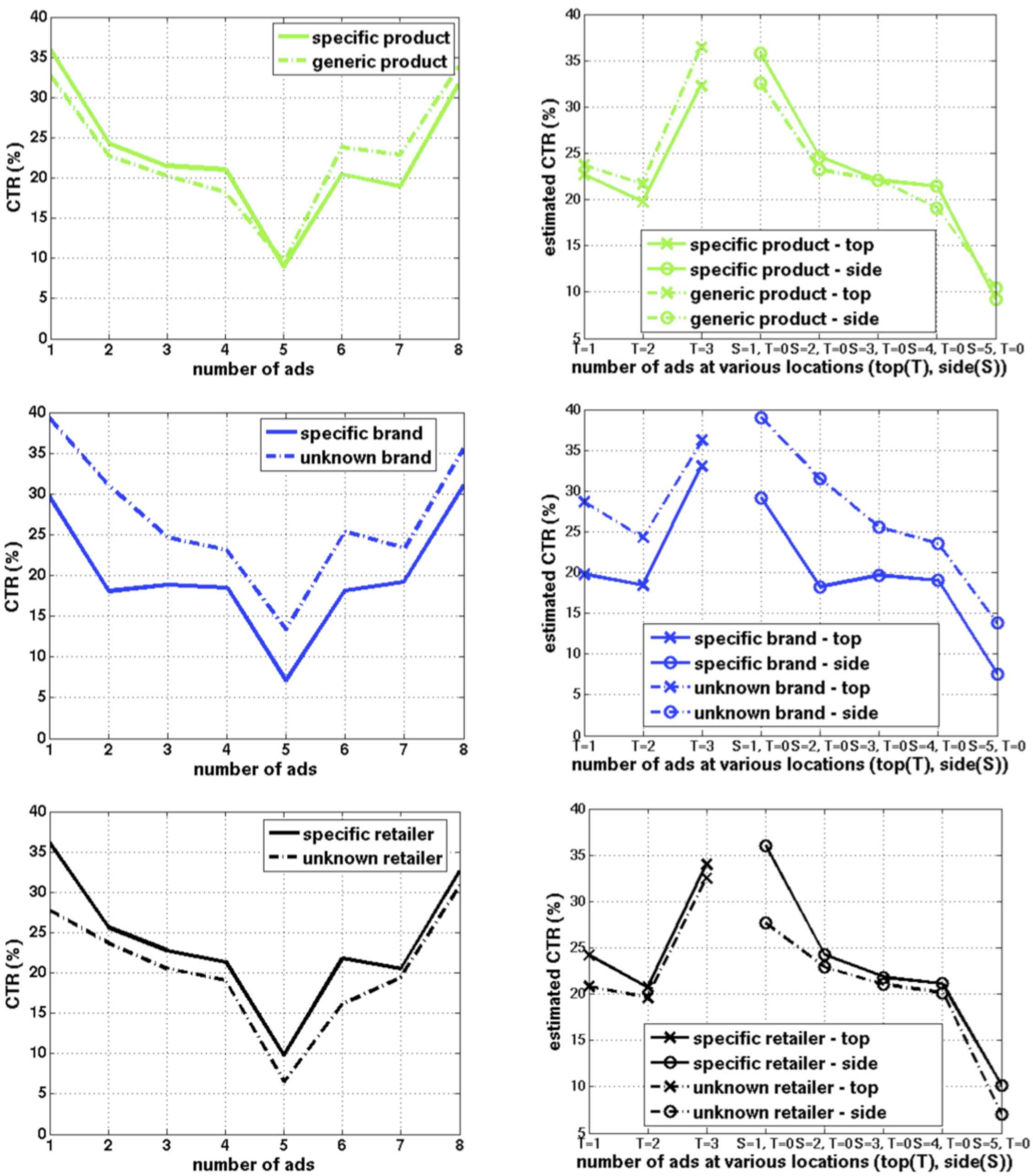


Figure 4.10: Average CTR along with the adjusted plots for average CTR corresponding to a particular number of ads on top/ side of the page for Specific Product vs. Generic Product, Specific Retailer vs. Unknown Retailer, and Specific Brand vs. Unknown Brand.

sored search. This section focuses on context based clickthrough analysis, building on the evidence presented in previous sections.

The findings in the previous sections of this chapter suggest that contextual factors, such as the intent underlying user's queries, the total number of ads displayed on a result page, and the rank positions of ads, influence ad clicks. These contextual factors are therefore assumed to be effective in estimating the clickthrough rate for an ad that appears within a context. In what follows, models are examined that target ads within the context of the SERPs on which they appear. These models estimate the clickthrough rate of an ad as the overall probability of click that it is expected to receive across various contexts in which it is displayed in the history of its appearances.

4.4.1 Baseline Model

As depicted in Figure 4.2, the average CTR varies with respect to the rank position of ads and also to the total number of ads displayed on the page. A *baseline model* is formulated based upon this observation in order to study the quality of ads by estimating their clickthrough rate.

The click rate for a given ad can be estimated with respect to various contexts in which it appears in its history. In the baseline model, the context is defined according to the SERP/ad pair for each appearance of an ad. To be precise, the context is represented by the particular number of ads (e.g. from 1 to 8) that are listed on a search result page (SERP) where the targeted ad appears at a specific rank position in the list.

For a given ad a, the probability of click over a is denoted by $P(C = 1|a)$. This probability value is estimated with respect to the above contextual factors that are obtained from the set of SERPs, $\mathcal{D}_a$, on which ad a appeared in its history of appearances:

$$\begin{aligned} P(C=1|a) &= \sum_{i \in \mathcal{D}_a} P(C=1|a, I=i) P(I=i) \\ &\overset{(1)}{=} \sum_{i \in \mathcal{D}_a} P(C=1|I=i) P(I=i) \end{aligned} \tag{4.13}$$

$$
\begin{aligned}
&\stackrel{(2)}{=} \sum_{i \in \mathcal{D}_a} P(C=1, R=\mathbf{r}_i^a | N=\mathbf{n}_i) P(I=i) \\
&\stackrel{(3)}{=} \frac{1}{|\mathcal{D}_a|} \sum_{i \in \mathcal{D}_a} P(C=1, R=\mathbf{r}_i^a | N=\mathbf{n}_i)
\end{aligned}
$$

where the substitution of $P(C=1|a, I=i)$ by $P(C=1|I=i)$ in (1) is intended to consider the context in which the ad appears rather than its content. The substitution of $P(C=1|I=i)$ by $P(C=1, R=\mathbf{r}_i^a|N=\mathbf{n}_i)$ in (2) accounts for the rank of an ad, $\mathbf{r}_i^a$, as the available contextual factor for ad a on the SERP i. It also accounts for the total number of ads listed on the SERP i, i.e. $\mathbf{n}_i$. Note that the probability of the appearance of each page is assumed to be the same across the set of SERPs, $\mathcal{D}_a$. Hence, $P(I=i)$ is substituted by $\frac{1}{|\mathcal{D}_a|}$ in (3).

The probability $P(C=1, R=\mathbf{r}_i^a|N=\mathbf{n}_i)$ is obtained independent of the content of the ad, with respect to the context in which it appears. This probability distribution has been formulated and empirically calculated as the average context CTR in the previous section (see Equation 4.2). As a result, the baseline model estimates the probability of click for ad a with respect to the aggregated context-based probability of click (i.e. $P(C=1, R=\mathbf{r}_i^a|N=\mathbf{n}_i)$) across the previous appearances of the ad.

4.4.2 Query Intent Model

The impact of the identified query intent as an extra factor in the baseline model is studied under the query intent model. In other words, the probability of a click for an ad displayed on a result page is calculated with respect to: i) the total number of ads displayed on the page, ii) the rank position of the ad on the page, and iii) the intent underlying the query for which the ad is displayed. This model is referred to as the *query intent model*, and it is again based upon the aggregated performance of the context of the ad in the history of its appearances.

If a query appears to be mostly commercial, the effect of the three sub-categories of commercial intent (i.e. product, brand, and retailer) combined with the navigational/ informational intent of the query is taken into account. Otherwise, the combination of non-commercial intent and navigational/ informational intent categories is considered.

Similar to the baseline model, the click probability for a given ad a in the query intent model is estimated, as follows:

$$\begin{aligned} P(C=1|a) &= \sum_{i \in \mathcal{D}_a} P(C=1|I=i)P(I=i) \\ &\stackrel{(1)}{=} \sum_{i \in \mathcal{D}_a} P(C=1, R=\mathbf{r}_i^a | N=\mathbf{n}_i, G=\mathbf{g}_{q_i})P(I=i) \\ &= \frac{1}{|\mathcal{D}_a|} \sum_{i \in \mathcal{D}_a} P(C=1, R=\mathbf{r}_i^a | N=\mathbf{n}_i, G=\mathbf{g}_{q_i}) \end{aligned} \tag{4.14}$$

where the substitution in (1) accounts for the rank of ad ($\mathbf{r}_i^a$), the number of displayed ads on the page ($\mathbf{n}_i$), and the user's intention behind the query (q_i) for SERP i. Here, $\mathbf{g}_{q_i}$ denotes the combination of query types for the query q_i that correspond to the SERP i, and $P(C=1, R=\mathbf{r}_i^a|N=\mathbf{n}_i, G=\mathbf{g}_{q_i})$ represents the context CTR estimated according to Equation 4.12 from the previous section.

$P(C=1, R=\mathbf{r}_i^a|N=\mathbf{n}_i, G=\mathbf{g}_{q_i})$ is estimated based on the average context CTR for the combination of specific sub-categories of commercial intent and the navigational/ informational category given that the corresponding query was detected as commercial earlier (i.e. twelve possible pairs of intents). In this case, $P(C=1, R=\mathbf{r}_i^a|N=\mathbf{n}_i, G=\mathbf{g}_{q_i)}$ in Equation 4.14 is estimated as the average of the context CTR values corresponding to the specific sub-categories of commercial intent at rank r when a total of n ads are displayed on the result page. Otherwise, if the corresponding query is detected to be noncommercial, $P(C=1, R=\mathbf{r}_i^a|N=\mathbf{n}_i, G=\mathbf{g}_{q_i})$ is estimated based on the average CTR for the combination of noncommercial intent and the navigational/ informational category. That is, set G for the specific query model contains two other intent pairs: navigational–noncommercial and informational–noncommercial.

4.4.3 Evaluation Results

The main purpose of this evaluation is to compare the query intent model against the baseline model. The hypothesis here is whether the inclusion of query intent information as a contextual factor for clickthrough analysis provides better correlation and improvement in the estimation process.

For this purpose, we collected some performance metrics that have been used for similar purposes in the literature. Each metric is intended to evaluate the performance of the models in estimating the probability of click for ads in set $C^{(1)}$ (see Chapter 1, Section 1.3.1). In order to avoid train-test contamination, the context CTR values from the previous section (4.4) have been calculated using the set $B^{(1)}$, while in this section they are used for the estimation purposes on the instances in set $C^{(1)}$.

For a given ad a, its actual clickthrough rate across the set of SERPs, $\mathcal{D}_a$, is computed as follows:

$$\mathbf{CTR}_a = \frac{\sum_{i \in \mathcal{D}_a} I_c(a,i)}{|\mathcal{D}_a|} \tag{4.15}$$

where $I_c(a,i)$ is a binary indicator representing whether ad a was clicked once it was displayed on the search result page i.

Kullback-Leibler Divergence (KLD, also known as KL-divergence) measures the lack of fit between a model and the actual values in the data relative to a perfect fit (Eguchi and Copas, 2006; Kullback and Leibler, 1951), which makes it suitable in our case in order to measure the fitness of the estimated values with respect to the actual clickthrough rates.

Given the click probability $P(C = 1|a)$ estimated for ad a from either Equation 4.13 or Equation 4.14, one can formulate KLD with respect to ad a, as follows:

$$KLD_a = [\mathbf{CTR}_a \times log(\frac{\mathbf{CTR}_a}{P(C=1|a)})] + [(1 - \mathbf{CTR}_a) \times log(\frac{1 - \mathbf{CTR}_a}{P(C=0|a)})]$$

where $P(C = 0|a) = 1 - P(C = 1|a)$. The average KLD value is computed across all the available ad instances, where a perfect model would score 0 based on this average KL-divergence metric.

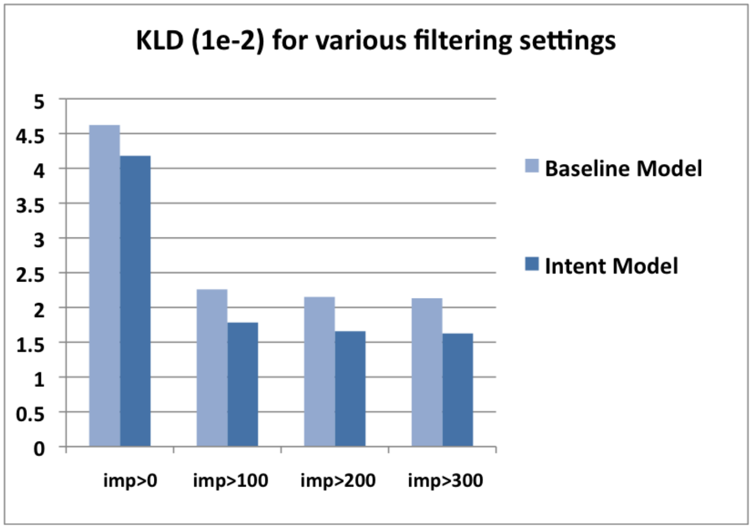
RMSE (root mean square error) is used as another metric of evaluation in order to quantify the accuracy of estimates. The squared error value with respect to ad a is calculated as $SE_a = [\mathbf{CTR}_a - P(C = 1, a)]^2$. The root mean square error is computed as the square root of the average of this value across all ad instances.

The click probability of ads in set $C^{(1)}$ was calculated according to Equation 4.13 for the baseline model. The click probability for this group of ads was separately calculated according to the Equation 4.14 for the query intent model.

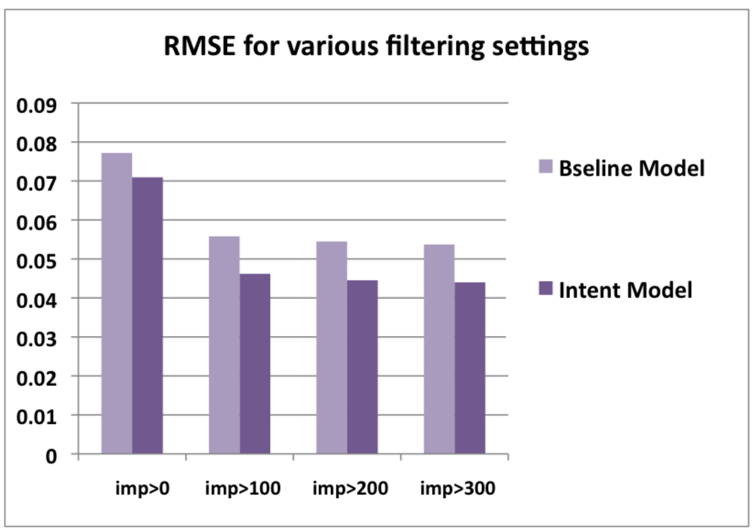
Figure 4.11 reports values of the metrics for the baseline model and the intent model with different filtering settings. There have been various filtering settings reported in literature in order to decrease the amount of noise in the results of the empirical estimation of CTR. Richardson *et al.* (2007) and Debmbsczynski *et al.* (2008) select threshold values of 100 and 200 respectively in order to filter ads in their analyses. The main purpose behind filtering is to avoid any possible noise from ads with too few SERPs. Hence, the results based on various filtering settings are reported in Figure 4.11. The results of the evaluation over the non-filtered instances are also reported to provide a better understanding of the analysis over the entire set.

Apart from the first setting that focuses on all impressions, including the ones with too few appearances, both measures report consistent values across the filtering settings. One can observe that within all settings and with respect to both metrics, the query intent model outperforms the baseline model. This observation suggests that the inclusion of the query intent information to the page contextual factors can better reflect the click behavior of ads displayed on result pages.

In order to further study the difference between the query intent model and the baseline model, we performed paired t-tests and f-tests comparing the set of estimated clickthrough rate values based on the query intent model with the set of estimated clickthrough rate values based on the baseline model. Both tests were carried out at a significance level of 95%, and indicated statistically significant differences in the mean and variance of the two sets of results.



(a) KLD



(b) RMSE

Figure 4.11: Performance measures of the CTR estimation.

4.5 Discussion of Results

This chapter reports an empirical study of ad clickthrough behavior with respect to the search result page context. The main objective of this study is to provide evidence for further analyses of user browsing and click behavior in sponsored search. To sum up, the observation findings of this chapter can be listed as follows:

- Clicks on ads appear to occur mostly at the first and the second ranks, and most especially at the first rank. The clickthrough rate of ads decreases as they are displayed in lower ranks on the page, possibly as a result of reduced visual attention from users (Richardson *et al.*, 2007).
- In addition to the rank position of ads, the location of the page on which they appear, impacts the attention and eventually the clicks that they receive from the user. The top-listed ads tend to receive more clicks compared to the side-listed ads.
- A user's query, as a means of access their search intent, appears to have impact on their click behavior in sponsored search. In particular, commercially oriented queries tend to result in relatively high click rates on ads placed at different locations of the result pages.
- Observing varying clickthrough behavior for different categories of query intent can further confirm the classification results obtained in the previous Chapter. This suggests that there actually exist distinctions in the studied query categories both in terms of the context in which thy appear (i.e. search result page) and of the user's click behavior over the corresponding ads.
- The context-based ad click analysis suggests that the contextual factors such as the intent behind a user's query have correlations with the performance of ads in sponsored search. Comparing a baseline model based upon the position of ads and the total number of displayed ads against a model that also accounts for the query intent suggests that the inclusion of query intent information as a contextual factor provides a better estimation of the ad's quality.

While the findings of this chapter suggest that ad clickthrough prediction techniques could benefit from the query intent information and other contextual factors, there are still questions and limitations that need to be addressed. For instance, what if there are other categories of query intent that should be considered? Is there a better taxonomy of commercial intent than the sub-categories considered in this chapter? In other words, further exploration in this area by identifying a taxonomy of commercial intent represents an important future direction for this work.

Even if an extensive taxonomy is obtained such that a broad range of context is covered, efforts must be carried out to label queries in various dimensions of query intent. In this way, the context model needs to be expanded across various dimensions and training data needs to be collected across various contexts. In other words, further efforts and resources are required to expand the study performed in this chapter in a supervised and offline fashion. Another problem concerns the estimation of context-based click probabilities that compute an aggregated probability across various impressions. This would make more sense if the search engine placed the ad results in random order. The combination of position bias and the varying relevance of ads at different positions demands a model that accounts for the difference in user browsing behavior with respect to co-appearance and the various placements of ads on result pages. Investigating variability of user behavior in Web search systems through the differences that may exist in the interactions for different users and different queries (Carterette *et al.*, 2012; Piwowarski *et al.*, 2009; White and Drucker, 2007) provides a better understanding of the user's interactions and behavioral patterns in these systems.

These all provided motivation to use the findings of this chapter as evidence for modeling a user's browsing and click behavior in a semi-supervised and online fashion in the next chapter. Instead of employing the explicit judgements of the query intent, the contextual factors will be modeled through various query– and page–dependant parameters. These parameters are learned and updated in an online fashion.

Chapter 5

Modeling User Browsing and Click Behavior in Sponsored Search

While the findings of the previous chapter provide insight into ad clickthrough behavior, and how it is influenced by the context in which ads appear, we still must consider how different users browse sponsored links differently. This chapter assumes that the rate at which an ad is viewed and clicked depends both on its own quality and on the quality of the other ads that are displayed above it on the result page. This assumption is similar to that of the cascade model for user behavior (Craswell *et al.*, 2008) in which a document is assumed to be seen only if the user scans over all the ones ranked above it. In other words, the question here is whether one can benefit from the properties of the cascade-based click models in sponsored search in order to model the click on an ad in the context of the other ads displayed on a result page.

In this chapter, we augment our notion of context by defining a group of location and query *biases*, which reflect the variability of the user behavior in search and the ad placement strategy on result pages. Unlike the models from the previous chapter, explicit judgements of query intent are not required, since these contextual factors are modeled through various query– and page–dependant parameters.

A notion of *location bias* is formally modeled in order to account for top-listed and side-listed ads separately. Furthermore, *query biases* are introduced and parameterized in order to account for the probability that the user will initiate browsing of the ad list, and for their persistence (patience) in continuing to browse through the list.

The *initiation* probability with respect to a particular query is defined as the chance that the user who issues this query will eventually initiate browsing the ad list. The *persistence* or *transition* probability with respect to a query is defined as the chance that the corresponding user who examined a particular ad at rank i will continue on to examine the ad at rank $i + 1$. Both the initiation probability and the transition probability are determined separately for different display locations of ads on result pages. These parameters are learned and updated in an online fashion; hence, one benefit of the models studied in this chapter is their *online* nature.

Next, we elaborate on the motivation for this work. This motivation is followed by evidence supporting the main ideas underlying the introduction of location and query biases in sponsored search. These biases are formulated and modeled later in the chapter. The parameters of the model are learned through an expectation-maximization technique. Through the experimental study reported at the end of the chapter, we evaluate the performance of a group of cascade-based click models extended with these biases and applied to the sponsored search domain.

5.1 Motivational Points

As the primary goal of this work, we aim to exploit characteristics of user behavior to improve user browsing models and click prediction. We hope to gain insight not only into the click behavior of users, but into the user's browsing behavior in sponsored search, eventually employing this information to infer an ad's quality. One of these characteristics is a user bias against sponsored search (Jansen and Resnick, 2006), motivating us to study the impact of variability in user persistence in the sponsored search domain. Another characteristic is the user's response to page structure, specifically the locations on the page where ads appear, such as the top and side, and the ordering or ranking of ads at each location.

Building on related work, we develop models that incorporate *location* and *query* biases. We re-visit a group of well-known click models from the organic search domain, adapting and extending them to the sponsored search domain to reflect these biases. As a particular example, in this chapter, we adapt and extend the DBN (Chapelle and Zhang, 2009) click model. Under this approach, we model the set of ads displayed at each location through a separate dynamic Bayesian network, with an extended version of DBN (Chapelle and Zhang, 2009) dedicated to each location. Other models could be extended in a similar fashion, where we generally refer to models that accommodate query and location biases as *location– and query–aware* browsing/click models.

5.2 Location Bias

In contrast to organic links, ad links do not strictly appear one after the other on a result page. In addition to their rank position, ads can be characterized according to the location on the page on which they appear, typically the top (north) or the right (east) side. Exploring the context of search result pages, as in the previous chapter, confirms that top ads are relatively more likely to be clicked compared to the side ads. This observation suggests that users expect top-listed ads to be more relevant than side-listed ads, and therefore they are examined by the user more frequently (Jansen, 2007; Jansen and Resnick, 2006).

In order to provide additional evidence for this suggestion, Figure 5.1 plots the relative click rate for different locations and positions of ads on result pages. These statistics have been collected from set $B^{(2)}$ (see Chapter 1, Section 1.3.2). Relative click rates are computed from the clicks recorded for this set of SERPs with eight ads, such that three of them appear at the top of each page (denoted by the first three positions on the x–axis) and five of them appear at the side of the page (denoted by the last five positions on the x–axis). For the first position of the top-listed ads a click rate of one is assigned in the plot, while for the rest of positions and locations the click rate relative to this position is reported.

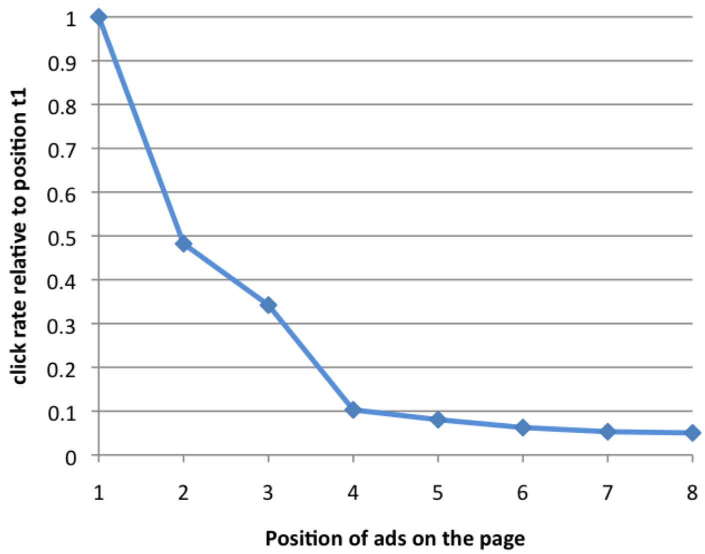


Figure 5.1: Relative click rate for different locations/positions of the result pages.

The impact of the position bias on click rates is obvious. The frequency of clicks decreases towards lower rank positions, possibly due to the reduced visual attention from the user. However, the main intention here is to show the significant drop of click rates starting from position 4, which represents the first rank of the side-listed ads, suggesting that there is relation between the location of ads and user behavior.

Let N_t and N_s be the number of ads displayed on the top and at the side of a result page (as part of a SERP), and define $N = N_t + N_s$. An ad displayed at the rank position i placed on the location l of the page is denoted by $a_{l,i}$. Ads displayed at the top are denoted according to their ranking: $a_{t,1}$,..., a_{t,N_t}. Similarly, side-listed ads are denoted as $a_{s,1}$,..., a_{s,N_s}.

For a result page with a total of N ads displayed for the query q, the following variables are defined to model various characteristics of the user, the query, and the displayed ads:

- U_l is a binary hidden variable representing whether or not the user initiates browsing at location l.

- $E_{l,i}$ is a binary hidden variable indicating whether the user examines the ad at rank i of location l.

- $C_{l,i}$ is a binary variable representing the click observation at rank i of location l.

- $A_{l,i}$ is a binary hidden variable reflecting the user's perceived relevance of the ad displayed at rank i of location l.

- $S_{l,i}$ is a binary hidden variable reflecting the user's satisfaction (post-click relevance) once they click on the ad displayed at rank i of location l.

For all these variables, l represents the location of the ad displayed on the page which may be t (for the top-listed ads) and s (for the side-listed ads). Also note that U_l and $E_{l,i}$ are properties of the query (thus, properties of the user who issues the query) and of the location, which means they are shared across result pages with the same query. The sequence of click observations obtained from the page with respect to its ads represent $C_{l,i}$ values. Finally $A_{l,i}$ and $S_{l,i}$ are considered to be properties of ads with respect to the query for which they are displayed, so they are defined over ad-query pairs.

5.3 Query Bias

The query itself represents an important aspect of ad context, which can significantly impact the expected click rate. Even though terms in the query act as triggers for ad selection, the nature of the query, and the user intent underlying the query, still plays a major role.

As shown in Chapter 3, information about user intent can, to some extent, be inferred from the organic results appearing on the search result page. As shown in Chapter 4, if a query is commercially oriented (i.e. the user may be intending to purchase a product or service), the user may be more likely to click on an ad. A weakly related ad appearing with the results of a commercially oriented query may receive more clicks than a strongly related ad appearing with the results of a less commercially oriented query. On the other hand, if there is no commercial intent underlying the query, the user may not consider the ads at all.

As also discussed in Chapter 4, within the category of commercial queries, finer levels of query intent may also be considered. For instance, consider the query "running shoe' versus the query "Nike shoe". Both are commercial queries, making their result pages appropriate for sponsored links. The first query targets a specific commercial product, i.e., a shoe, without regard for the brand, while the second targets a special brand of this product, i.e., a Nike shoe.

One could argue the user who enters the first query might be more engaged in the browsing process, since any brand of shoe might do. On the other hand, the user who issues the second query may be a relatively loyal user who is looking for their favorite brand among the top results. If they do not find such results among the top-listed sponsored links, they may either abandon the search or move on to the organic results (Jansen and Resnick, 2006). In other words, a user who issues a query for a specific product, with no specific brand in mind, may have a greater tendency to scan through the entire sponsored links as opposed to a user who is looking for a specific brand.

As the query intent tends to vary across multiple dimensions, parametric biases can be formulated that depend on user queries. Two types of *query bias* are introduced in this chapter, as factors involved in browsing behavior. The first query bias deals with the initiation of browsing over advertisement links, while the second one reflects user persistence in browsing advertisement links.

5.3.1 Initiation

The trigger point to start examining an ad list should matter more for sponsored results as opposed to organic results. The reason goes back to the user's bias against sponsored links, as compared to organic links (Jansen and Resnick, 2006), where users appear to have a greater tendency to examine organic links rather than the sponsored links. Hence, we formulate the following hypothesis for the way a user targets a list of ads at particular location of a result page:

It is assumed that if a user initially has any motivation to consider ads at a particular location of the page then they start examining ads. In other words, $E_1^l = 1$, if and only if the user has any intention to consider ads placed at location l of the page:

$$U_l = 1 \Leftrightarrow E_{l,1} = 1 \tag{5.1}$$

Therefore, the initiation probability u_l^q is defined at the query-level and the location-level, where q represents the query and l represents the location. The initiation probability u_l^q represents the chance that the user will eventually initiate browsing ads listed at location l of the page, i.e., $P(U_l = 1) = u_l^q$. Using this definition and Equation 5.1, given query q, the probability of examining the first ad at location l can be calculated as follows:

$$P(E_{l,1} = 1) = P(U_l = 1) = u_l^q \tag{5.2}$$

Once a user starts browsing ads listed at a particular location, their browsing persistence can be addressed using a variation of the cascade model, as explained next.

5.3.2 Persistence

Newer versions of the cascade model (Chapelle and Zhang, 2009; Guo *et al.*, 2009a,b; Zhu *et al.*, 2010) share a notion of the user's *patience* or *persistence* (λ) as they move from link to link. In order to model the relationship between user persistence and the associated query, this work introduces variability into the persistence parameter λ. We assume that different users have different levels of patience in browsing through an ad list. As a result, the persistence parameter for the DBN model in Equation 2.5 can be revised to λ_l^q in order to take into account the user's query (q) and the ad's location on the page (l).

The proposed user browsing and click model for ads is depicted in Figure 5.2. Note that this model accounts for both the location bias and the query bias with respect to both browsing initiation and browsing persistence. Given query q, if the user initiates browsing (see Eq. 5.2), they begin examining the list of ads displayed at location l. If the user examines ad a displayed at the rank position i at this location, two scenarios are possible:

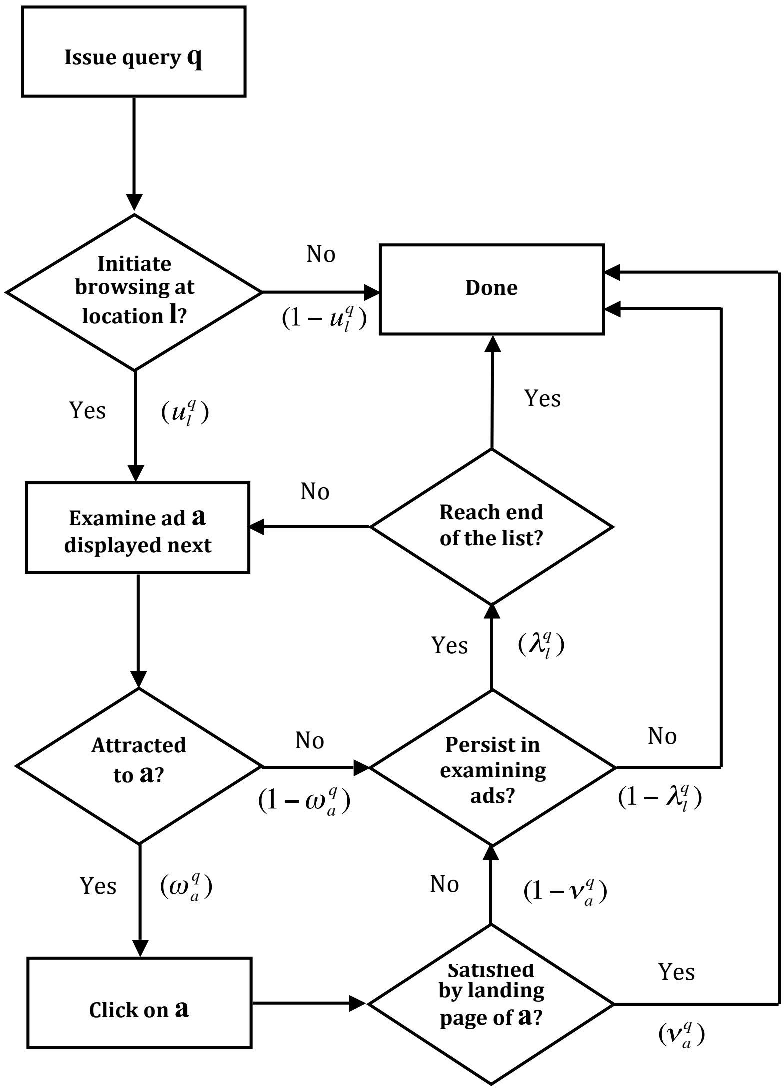


Figure 5.2: The location- and query- aware browsing model.

click or no click. If they do not click on the ad (i.e. $C_{l,i} = 0$), they either move on to the next ad with probability λ_l^q, or they abandon their search with probability $1 - \lambda_l^q$:

$$P(E_{l,i+1} = 1 | E_{l,i} = 1, C_{l,i} = 0) = \lambda_l^q \tag{5.3}$$

On the other hand, the user may examine the ad (i.e. $E_{l,i} = 1$) and perceive it to be relevant to their commercial need (i.e. $A_{l,i} = 1$) with the attraction probability of ω_a^q and click on it (i.e. $C_{l,i} = 1$). The perceived relevance may depend on the presentation quality of the ad, such as the ad title or its creative, and of course on the user's understanding of the relevance of the ad to their query. In other words, if a user examines and ad and finds it relevant to their commercial need, they click on it:

$$E_{l,i} = 1, A_{l,i} = 1 \Leftrightarrow C_{l,i} = 1 \tag{5.4}$$

According to the cascade assumption, when there is no examination, there is clearly not going to be a click. The click probability at rank i given the examination state can therefore be formulated as:

$$\begin{aligned} &P(C_{l,i} = 1|E_{l,i} = 1) = P(A_{l,i} = 1) = \omega_a^q \\ &P(C_{l,i} = 1|E_{l,i} = 0) = 0 \end{aligned} \tag{5.5}$$

Three scenarios are likely upon a click:

1. With a probability of ν_a^q, the user may be satisfied by the content of the landing page and stop looking at ads:

$$\begin{aligned} &P(S_{l,i} = 1|C_{l,i} = 1) = \nu_a^q \\ &S_{l,i} = 1 \Rightarrow E_{l,i+1} = 0 \\ &P(E_{l,i+1} = 1|S_{l,i} = 1) = 0 \end{aligned} \tag{5.6}$$

2. With a probability of $(1 - \nu_a^q)\lambda_l^q$ they may be unsatisfied by the landing page and move on to the next ad:

$$\begin{aligned} &P(S_{l,i} = 0|C_{l,i} = 1) = 1 - \nu_a^q \\ &P(E_{l,i+1} = 1|E_{l,i} = 1, S_{l,i} = 0) = \lambda_l^q \end{aligned} \tag{5.7}$$

3. With a probability of $(1 - \nu_a^q)(1 - \lambda_l^q)$ they may abandon their search.

$$P(E_{l,i+1} = 0|E_{l,i} = 1, S_{l,i} = 0) = 1 - \lambda_l^q$$

Here, ν_a^q is the *satisfaction probability* representing the probability that the user, who clicked on ad a and viewed its landing page, finds the ad satisfactory with respect to their query q. Note that if the user does not click on ad a ($C_{l,i} = 0$), they never get the chance to see the content of the landing page, so $S_{l,i} = 0$, which can be formalized as follows:

$$\begin{aligned} &C_{l,i} = 0 \Rightarrow S_{l,i} = 0 \\ &P(S_{l,i} = 0 | C_{l,i} = 0) = 1 \end{aligned} \tag{5.8}$$

Also note that the introduction of the parameter ω_a^q (representing the attraction probability) allows the model to implicitly account for the content-based quality of the ads while the ad title and ad creative are not known from the available log data.

5.4 Parameter Inference

In this section, we examine the inference procedure for the extended version of the DBN model, introduced earlier. The algorithm will be essentially the same for any other cascade-based click model, once extensions are applied to accommodate the location bias and the two types of query bias.

In the case that all random variables are known throughout the training samples (i.e. *complete* data case), a maximum likelihood approach would solve for the parameters of the model that maximize the log-likelihood of the observed data. However, in cases similar to our model, the introduced hidden variables result in missing data, suggesting a maximum likelihood estimate of the parameters over the *incomplete* data. The expectation maximization (EM) (Dempster *et al.*, 1977; Do and Batzoglou, 2008) technique is a natural generalization of the maximum likelihood estimation to the incomplete data case. In particular, EM attempts to find parameters of the model that maximize the log-likelihood of the observed data, which consists of the click signals in the case of our problem.

Hence, the proposed inference technique finds the maximum likelihood estimates of the parameters set $\theta = (u_t^q, u_s^q, \lambda_t^q, \lambda_s^q, \omega_a^q, \nu_a^q)$ corresponding to the hidden variables of the model. There is no known way to analytically solve for the model, however, $Pr(C|\theta)$ can be locally maximized using the standard iterative approach for the expectation maximization (Dempster *et al.*, 1977). Each iteration consists of the calculation of the expected complete-data log-likelihood with respect to the posterior distribution of the hidden variables (E-step) followed by updating and improving the value of each parameter through maximizing the expected complete-data log-likelihood function (M-step). We adopt an initial estimate of 0.5 for all parameters.

In what follows, we explain the inference procedure for the model's parameters. Note that the superscript j added to the variables, originally introduced in Section 5.2, indicates the SERP j to which these variables belong.

Initiation Probability: As for u_t^q and u_s^q, the posterior distribution of their corresponding hidden variables is calculated in the E-step of each iteration. We explain the analysis for u_l^q in the general form which can be expanded to the top-listed ads and the side-listed ads by substituting l with t and s, correspondingly.

Define $\phi(U_l^j)$ as the posterior distribution of the variable U_l^j given the click sequence C_l^j observed on the location l of the SERP j and the current value of the parameter u_l^q:

$$\phi(U_l^j = 1) = P(U_l^j = 1 | C_l^j, u_l^q) = \begin{cases} u_l^q & \text{if there is no click} \\ 1 & \text{otherwise} \end{cases} \tag{5.9}$$

For a given SERP of query q, if no ad click is observed over the ads listed at location l of the page, there is a probability of u_l^q that the user started examining these ads, and a probability of $1 - u_l^q$ that the user skipped the list in the first place. On the other hand, if any click is recorded for this location, the cascade assumption implies that the user considered the ad list and started examining from the first ad; hence, the posterior probability becomes 1 (See Eq. 5.1).

Given the posterior distribution $\phi(U_l^j)$ for the j^{th} SERP of q ($1 \leq j \leq M$), the expected complete-data log-likelihood function $Q(u_l^q, u_l^{q(k)})$ (Dempster *et al.*, 1977) at the iteration k of the inference procedure can be computed as follows:

$$
\begin{aligned}
Q(u_l^q, u_l^{q(k)}) &= \sum_{j=1}^{M} P(U_l^j = 0|C_l^j, u_l^{q(k)}) \log(1-u_l^q) + P(U_l^j = 1|C_l^j, u_l^{q(k)}) \log(u_l^q) \\
&= \sum_{j=1}^{M} \phi(U_l^j = 0) \log(1-u_l^q) + \phi(U_l^j = 1) \log(u_l^q) \qquad (5.10)
\end{aligned}
$$

where $u_l^{q(k)}$ denotes the current value of the parameter u_l^q, which is used for computing $\phi(U_l^j)$ according to Equation 5.9.

The expected complete-data log-likelihood function can then be locally maximized by solving for the partial derivative of the function with respect to the parameter u_l^q to be 0 at this iteration:

$$
\begin{aligned}
& \frac{\partial Q(u_l^q, u_l^{q(k)})}{\partial u_l^q} = 0 \qquad (5.11) \\
\Rightarrow\ & \sum_{j=1}^{M} \phi(U_l^j = 0) \frac{-1}{1-u_l^q} + \phi(U_l^j = 1) \frac{1}{u_l^q} = 0 \\
\Rightarrow\ & \sum_{j=1}^{M} (1-\phi(U_l^j = 1)) \frac{-1}{1-u_l^q} + \phi(U_l^j = 1) \frac{1}{u_l^q} = 0 \\
\Rightarrow\ & \sum_{j=1}^{M} \frac{\phi(U_l^j = 1) - u_l^q}{u_l^q(1-u_l^q)} = 0 \\
\Rightarrow\ & u_l^q = \frac{\sum_{j=1}^{M} \phi(U_l^j = 1)}{M}
\end{aligned}
$$

Therefore, after the M-step of the algorithm at the iteration k, the value of the parameter u_l^q will be updated to the above value, which locally maximizes the Q function:

$$
u_l^{q(k+1)} = \operatorname*{argmax}_{u_l^q} Q(u_l^q, u_l^{q(k)}) = \frac{\sum_{j=1}^{M} \phi(U_l^j = 1)}{M} \qquad (5.12)
$$

Persistence Probability: One of the major challenges in the inference procedure relates to the variability of the persistence probability. As stated before, previous models incorporate a fixed parameter that is usually obtained from domain knowledge. However, we assume persistence varies as follows:

The persistence in transitioning from ad to ad depends on the user's query and on the location of the ad list. Hence, two parameters λ_t^q and λ_s^q are defined to reflect the user's query q and the location l on the page (i.e. either t or s representing the top and the side locations respectively).

The analysis is explained for the general form of parameter λ_l^q, where $\phi(E_{l,i+1}^j)$ is defined as the posterior distribution of variable $E_{l,i+1}^j$ for the SERP j, given the observed click sequence C_l^j at the location l, the previous examination state at rank i, and the current value of λ_l^q:

$$\begin{aligned}
\phi(E_{l,i+1}^j = 1) &= P(E_{l,i+1}^j = 1 | C_l^j, E_{l,i}^j = 1, \lambda_l^q) \\
&= \frac{P(E_{l,i+1}^j = 1, E_{l,i}^j = 1, C_l^j, \lambda_l^q)}{P(E_{l,i}^j = 1, C_l^j, \lambda_l^q)} \\
&= \frac{P(E_{l,i+1}^j = 1, E_{l,i}^j = 1 | C_l^j, \lambda_l^q) P(C_l^j | \lambda_l^q)}{P(E_{l,i}^j = 1 | C_l^j, \lambda_l^q) P(C_l^j | \lambda_l^q)} \\
&= \frac{P(E_{l,i+1}^j = 1, E_{l,i}^j = 1 | C_l^j, \lambda_l^q)}{P(E_{l,i}^j = 1 | C_l^j, \lambda_l^q)}
\end{aligned} \tag{5.13}$$

We note that the posterior distribution for the transition variable $E_{l,i+1}^j$ is computed at the presence of the previous examination state that occurs at rank i (i.e. $E_{l,i}^j$). This is due to the cascade assumption which implies there is no continuation in transition at the rank position $i+1$ unless $E_{l,i}^j = 1$.

The numerator and denominator of Equation 5.13 are computed according to Equation A.6 from Appendix A using the forward-backward algorithm described in the Appendix.

In each iteration, the posterior distribution is computed for the possible transitions across the existing rank positions of the location l of each SERP j. The value of the persistence parameter λ_l^q at the k^{th} iteration can then be updated by maximizing the expected complete-data log-likelihood function:

$$\begin{aligned}
\lambda_l^{q(k+1)} &= \operatorname*{argmax}_{\lambda_l^q} Q(\lambda_l^q, \lambda_l^{q(k)}) \qquad (5.14)\\
&= \operatorname*{argmax}_{\lambda_l^q} \sum_{j=1}^{M} \sum_{i=1}^{N_l^j-1} [P(E_{l,i+1}^j = 0 | C_l^j, E_i^j = 1, \lambda_l^{q(k)})) \log(1-\lambda_l^q) \\
&\qquad + P(E_{l,i+1}^j = 1 | C_l^j, E_i^j = 1, \lambda_l^{q(k)}) \log(\lambda_l^q)] \\
&= \operatorname*{argmax}_{\lambda_l^q} \sum_{j=1}^{M} \sum_{i=1}^{N_l^j-1} \phi(E_{l,i+1}^j = 0) \log(1-\lambda_l^q) + \phi(E_{l,i+1}^j = 1) \log(\lambda_l^q) \\
&= \frac{\sum_{j=1}^{M} \sum_{i=1}^{N_l^j-1} \phi(E_{l,i+1}^j = 1)}{\sum_{j=1}^{M} N_l^j - 1}
\end{aligned}$$

where N_l^j is the number of ads appearing at the location l of the SERP j, resulting in N_l^j-1 possible transitions between ads at this location. The argmax function, similar to the case for the initiation parameters, finds the local maximum by taking the partial derivative of the expected complete-data log-likelihood function with respect to the λ_l^q parameter.

Relevance parameters: Parameters ω_a^q and ν_a^q can be estimated in a similar fashion as Chapelle and Zhang (2009) in organic search. We provide details of the inference here, taking different locations of the page into account.

Define $\phi(A_{l,i}^j)$ as the posterior distribution of the variable $A_{l,i}^j$ given the click sequence C_l^j observed on the location l of SERP j, ad a displayed at rank i on this location, and the current value of the parameter ω_a^q. In order to compute this posterior probability, two cases need to be considered:

Case 1: If ad a is not clicked, i.e. $C_{l,i}^j = 0$:

$$\begin{aligned}
\phi(A_{l,i}^j = 1) &= P(A_{l,i}^j = 1 | C_l^j, \omega_a^q) \\
&= \frac{P(A_{l,i}^j = 1, C_l^j, |\omega_a^q)}{P(C_l^j)} \\
&= \frac{\sum_{e\in\{0,1\}} P(A_{l,i}^j = 1, C_l^j, E_{l,i}^j = e | \omega_a^q)}{P(C_l^j)} \qquad (5.15)
\end{aligned}$$

Based on the relation of examination, perceived relevance, and click as shown in Equation 5.4, it is clear that $A^j_{l,i} = 1, C^j_{l,i} = 0 \Rightarrow E^j_{l,i} = 0$. In other words, in the posterior case that ad a is not clicked but it could be found relevant, the only possible case is that the user did not get a chance to examine it, i.e. $E^j_{l,i} = 0$. Thus, $P(A^j_{l,i} = 1, C^j_l, E^j_{l,i} = 1|\omega^q_a) = 0$ in Equation 5.15 above, and the posterior probability can be further simplified, as follows:

$$\begin{aligned}
\phi(A^j_{l,i} = 1) &= \frac{P(A^j_{l,i} = 1, C^j_l, E^j_{l,i} = 0|\omega^q_a)}{P(C^j_l)} \\
&= P(A^j_{l,i} = 1, E^j_{l,i} = 0|C^j_l, \omega^q_a) \\
&= P(A^j_{l,i} = 1)P(E^j_{l,i} = 0|C^j_l, \omega^q_a) \\
&= \omega^q_a P(E^j_{l,i} = 0|C^j_l, \omega^q_a)
\end{aligned}$$

where $P(E^j_{l,i} = 0|C^j_l, \omega^q_a)$ can be estimated according to Equation A.3 from Appendix A

Case 2: If ad a is clicked, i.e. $C^j_{l,i} = 1$, according to Equation 5.4, we can conclude that the user perceived the ad to be relevant. Thus, the posterior probability becomes 1, i.e. $\phi(A^j_{l,i} = 1) = 1$ in this case.

Overall, the posterior probability for the variable $A^j_{l,i}$ can be stated as follows:

$$\begin{aligned}
\phi(A^j_{l,i} = 1) &= P(A^j_{l,i} = 1|C^j_l, \omega^q_a) \\
&= \begin{cases} \omega^q_a P(E^j_{l,i} = 0|C^j_l, \omega^q_a) & \text{if there is no click on ad } a \text{ at rank } i \\ 1 & \text{otherwise} \end{cases}
\end{aligned} \tag{5.16}$$

The value of the attraction parameter ω^q_a on the k^{th} iteration can then be updated by maximizing the expected complete-data log-likelihood function:

$$\begin{aligned}
\omega^{q(k+1)}_a &= \operatorname*{argmax}_{\omega^q_a} Q(\omega^q_a, \omega^{q(k)}_a) \\
&= \operatorname*{argmax}_{\omega^q_a} \sum_{j=1}^{M} \sum_{l\in\{t,s\}} \sum_{i=1}^{N^j_l} [P(A^j_{l,i} = 0|C^j_l, \omega^{q(k)}_a) \log(1 - \omega^q_a) \\
&\qquad + P(A^j_{l,i} = 1|C^j_l, \omega^{q(k)}_a) \log(\omega^q_a)]
\end{aligned} \tag{5.17}$$

$$= \underset{\omega_a^q}{\operatorname{argmax}} \sum_{j=1}^{M} \sum_{l \in \{t,s\}} \sum_{i=1}^{N_l^j} I(\mathbf{a}_{l,i}^j = a)[\phi(A_{l,i}^j = 0) \log(1 - \omega_a^q) + \phi(A_{l,i}^j = 1) \log(\omega_a^q)]$$

$$= \frac{\sum_{j=1}^{M} \sum_{l \in \{t,s\}} \sum_{i=1}^{N_l^j} I(\mathbf{a}_{l,i}^j = a)\phi(A_{l,i}^j = 1)}{\sum_{j=1}^{M} \sum_{l \in \{t,s\}} \sum_{i=1}^{N_l^j} I(\mathbf{a}_{l,i}^j = a)}$$

where $I(.)$ is a binary indicator function, such that $I(\mathbf{a}_{l,i}^j = a) = 1$ if the i^{th} ad placed at location l of the j^{th} SERP is a, and is 0 otherwise. This function is introduced into the equation above, since instances of parameter ω_a^q are only available in the SERPs of query q with ad a displayed.

A similar inference procedure is applied for the estimation of the posterior probability for the satisfaction variable $S_{l,i}^j$. Details of this procedure are explained in Section A.2.2 of Appendix A. As a result, the posterior probability for the variable $S_{l,i}^j$ can be stated, as follows:

$$\begin{aligned} \phi(S_{l,i}^j = 1) &= P(S_{l,i}^j = 1 | C_l^j, \nu_a^q) \\ &= \begin{cases} 0 & \text{if there is no click on ad } a \text{ at rank } i \\ \frac{\nu_a^q}{(1-\lambda_l^q + \nu_a^q \lambda_l^q)} P(E_{l,i+1}^j = 0 | C_l^j, \nu_a^q) & \text{otherwise} \end{cases} \end{aligned} \tag{5.18}$$

where $P(E_{l,i+1}^j = 0 | C_l^j, \nu_a^q)$ can be estimated according to Equation A.3 from Appendix A.

Finally, the value of satisfaction parameter ω_a^q on iteration k^{th} can be updated by maximizing the expected complete-data log-likelihood function in a similar fashion as was done for the attraction parameter. The update formula for this parameter in the k^{th} iteration is as follows:

$$\nu_a^{q(k+1)} = \frac{\sum_{j=1}^{M} \sum_{l \in \{t,s\}} \sum_{i=1}^{N_l^j} I(\mathbf{a}_{l,i}^j = a, C_{l,i}^j = 1)\phi(S_{l,i}^j = 1)}{\sum_{j=1}^{M} \sum_{l \in \{t,s\}} \sum_{i=1}^{N_l^j} I(\mathbf{a}_{l,i}^j = a, C_{l,i}^j = 1)}$$

where $I(\mathbf{a}_{l,i}^j = a, C_{l,i}^j = 1) = 1$ if the i^{th} ad placed at location l of the j^{th} SERP is a, and it was clicked. Otherwise, if ad a is not clicked on SERP j, the indicator function returns 0.

5.5 Evaluation Results

Experiments conducted in this section is based on the sponsored search data described in Chapter 1. Two subsets of the data (i.e. $A^{(2)}$ and $B^{(2)}$) from setting #2 described in Section 1.3.2 are targeted in this chapter, as the purpose of the evaluation is twofold:

1. We evaluate the impact of the query biases, user initiation, and persistence, on user browsing and click behavior.
2. We evaluate the impact of the location bias and the result page structure, on user browsing and click behavior, along with the query biases.

In each round of experiments, using either set $A^{(2)}$ or set $B^{(2)}$, one pass is made over the SERPs in order to do online learning and testing. In each case, before taking the current SERP into account for training purposes, we make a prediction of the click probability of its ads using the values of the parameters obtained from the previously observed SERPs. Once the prediction is finished for the current SERP, its actual click signals are added to the rest of the training samples to update the posterior distributions, and to locally maximize the expected complete-data log-likelihood. One of the benefits of this learning process is its ability to perform online learning and testing. Therefore, the contextual factors for any pair of query and ad are continuously updated as more instances of the query and ad appear in the log.

5.5.1 Click Models

A group of click models from the organic search domain are borrowed as baselines for the experiments conducted in this chapter. The cascade model (Craswell *et al.*, 2008) has been selected as the primary baseline of the analysis. Under this model, the user is assumed to be infinitely persistent, continuing their examination unless they make a click (see Eq. 2.4). In addition, the Dynamic Bayesian Network (DBN) model (Chapelle and Zhang, 2009) has been selected as it is one of the state-of-the-art click models in which both the perceived relevance and the post-click relevance, and also the user's persistence, are modeled. However, the persistence probability is assumed to be constant across all sessions in DBN (see Eq. 2.5).

Finally, the Dependent Click Model (DCM) (Guo *et al.*, 2009b) has been selected as one of the variations of the cascade model in which a notion of varying persistence probability has been considered. However, the assumption of an infinitely persistent user is still partially included in this model, as the user is assumed to examine a document with a probability of one given they have not clicked on the previously examined document. There is a position-dependent parameter defined in this model, representing the probability that the user would be willing to see more results after a click (See Eq. 2.6). However, this transition probability is assumed to be the same across all queries.

None of these click models require extra information about the client side, ad content, or bid terms, which enables us to reproduce them on our log data. For each click model, we evaluate the performance of the model in the presence of the biases that were introduced in this work. Our goal is to understand whether any of these models and settings are able to better predict user behavior on advertisement links. The first group of settings reflect the query biases (initiation probability and persistence). The second group of settings reflects the query biases and the location biases together.

5.5.2 Click Prediction

As discussed previously, an ad is assumed to be clicked if and only if the user gets the chance to view the ad within the context that it is shown and finds it relevant to their commercial need. Therefore, given a new search result page $\mathbf{j}$ for the query $\mathbf{q}$, the probability of click for a given ad a, displayed at rank $\mathbf{i}$ on the location $\mathbf{l}$ of the SERP, can be predicted as follows:

$$\begin{aligned} P(C^{\mathbf{j}}_{\mathbf{l,i}} = 1) &= P(E^{\mathbf{j}}_{\mathbf{l,i}} = 1) P(C^{\mathbf{j}}_{\mathbf{l,i}} = 1 | E^{\mathbf{j}}_{\mathbf{l,i}} = 1) \\ &= \omega^{\mathbf{q}}_{\mathbf{a}} \, P(E^{\mathbf{j}}_{\mathbf{l,i}} = 1) \end{aligned} \tag{5.19}$$

where the attraction probability $\omega_{\mathbf{a}}^{\mathbf{q}}$ is substituted according to Equation 5.5, and $P(E_{\mathbf{l,i}}^{\mathbf{j}} = 1)$ is estimated according to Equation A.4 from Appendix A using the forward-backward algorithm (Rabiner, 1989) and with respect to the parameters of the model learned throughout the inference procedure. Incorporating $P(E_{\mathbf{l,i}}^{\mathbf{j}} = 1)$ into the click probability model assures that the quality of the ad is evaluated in the context of the preceding ads; satisfying one of the major goals of this chapter.

The above equation belongs to the model augmented with the location and query biases in the previous section. For the original models (i.e. CAS, DCM, and DBN) and the models within the other settings, a simpler form of this equation can be created in the same way (for instance, the parameter **l** should be removed from the equation for the models that do not implement the location bias).

5.5.3 Evaluation Metrics

We evaluate the performance of the models within different settings based on two standard metrics: *average log-likelihood* and *average perplexity.* For each model, the average log-likelihood (LL) is calculated with respect to the predicted click probability for each ad and the actual click signals observed in the log of M SERPs. For instance, the average log-likelihood for a click model augmented with the location bias is formulated as:

$$LL = \frac{\sum_{j=1}^{M} \sum_{l \in \{t,s\}} \sum_{i=1}^{N_l^j} \mathbf{c}_{l,i}^j \log P(C_{l,i}^j = 1) + (1 - \mathbf{c}_{l,i}^j) \log(1 - P(C_{l,i}^j = 1))}{\sum_{j=1}^{M} N_t^j + N_s^j} \tag{5.20}$$

where $\mathbf{c}_{\mathbf{l,i}}^{\mathbf{j}}$ represents the actual click signal recorded for the ad displayed at the rank position i of the location l of the SERP j, and $P(C_{l,i}^j = 1)$ is the probability of click that the model predicts for this ad (Eq. 5.19). N_l^j indicates the number of ads displayed at location l of the SERP j. A larger value of LL indicates a better model fit, where the ideal value is zero.

To further study the impact of the biases at different rank positions, the perplexity measure is used as the log-likelihood powers computed independently at each rank position (Zhang *et al.*, 2011). For instance, the average perplexity for the rank position i of location l can be computed across the M SERPs, as follows:

$$\rho_{l,i} = exp(\frac{-1}{M_{l,i}} \sum_{j=1}^{M_{l,i}} \mathbf{c}_{l,i}^j \log P(C_{l,i}^j = 1) + (1 - \mathbf{c}_{l,i}^j) \log(1 - P(C_{l,i}^j = 1))) \tag{5.21}$$

where $M_{l,i}$ denotes the number of pages in the log that displayed an ad at the rank position i of the location l. A lower value of the perplexity indicates a better fit between the model and the actual data. The details of the experiments based on each group of settings, along with their results, are presented next.

5.5.4 Evaluating the Impact of Query Biases

Settings that reflect different combinations of the two query biases are categorized and labeled in order to be referenced easily throughout the experiments. Their description can be found in Table 5.1. The formulation and inference of all the extended models are similar to those explained in the previous section. These settings do not reflect the location bias since they are evaluated on the larger sample of the data, set $A^{(2)}$ (see 1.3.2), which includes various number of ads on the search result pages with no location information recorded.

Table 5.1: The settings related to the query biases for the models under the experiment.

settings	**click models**		
original setting of the model	CAS	DCM	DBN (0.9, 0.5, 0.1, 0.001)
varying persistence	CAS+VP	-	DBN+VP
varying persistence and initiation	CAS+VPI	DCM+VI	DBN+VPI

Note that only for DCM, the variability of persistence factor is not accommodated in the model since DCM itself has a notion of variability in user persistence (as described before). Hence, the browsing initiation probability is the only factor added to DCM in this round of experiments.

In addition, as the persistence factor (λ) is fixed in the original DBN model (see Eq. 2.5), we train DBN using various values of this parameter. Thus, under the first setting of DBN in Table 5.1, four values for λ are provided, each resulting in a separate run for DBN. Among these, the run with $\lambda = 0.9$ represents the original DBN model with relatively patient users while the rest, particularly the one with $\lambda = 10^{-3}$, represents relatively impatient users.

Table 5.2: Runs across different settings of the query biases for the cascade model, DCM model, and DBN model.

model	**Cascade**			**DCM**		**DBN**					
setting	orig.	VP	VPI	orig.	VI	λ=0.9 (orig.)	λ=0.5	λ=0.1	λ=10^{-3}	VP	VPI
run#	1	2	3	4	5	6	7	8	9	10	11

Consequently, a total of 11 different runs are generated for evaluation purposes. Table 5.2 depicts these runs across different settings that are numbered from 1 to 11. The average log-likelihood for the runs are computed according to Equation 5.20 and plotted in Figure 5.3, where the run ids match those shown in Table 5.2. Each group of models are plotted in a different pattern. It can be seen that the overall performance of the DBN runs is relatively better than the other two models, as expected.

A variation of user persistence is already included in the original DCM model, whereas the inclusion of varying user persistence provides improvements over the original settings of the other two models. As for the DBN model, the inclusion of a varying persistence probability (i.e. DBN+VP in run 10) shows improvement over the original DBN (i.e. DBN with $\lambda = 0.9$ in run 6) and over DBN with $\lambda = 0.5$ in run 7. However, DBN+VP shows similar performance to that of the two settings of DBN with impatient users (i.e. DBN with $\lambda = 0.1$ in run 8 and DBN with $\lambda = 0.001$ in run 9). This may be due to the fact that the inclusion of the varying persistence probability is intended to reflect the lower persistence of the users in browsing through the advertisement links which is essentially the same as assuming the user to be impatient. Thus, runs 8, 9, and 10 are relatively comparable. One could argue that DBN+VP is able to reflect the impatience of the users in browsing

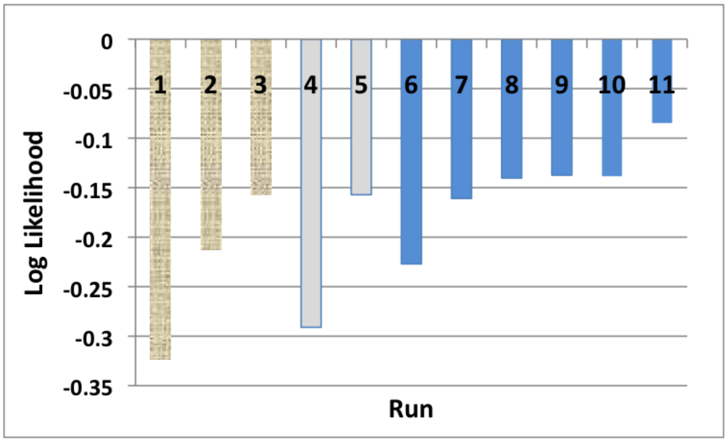


Figure 5.3: The average log-likelihood for the 11 runs.

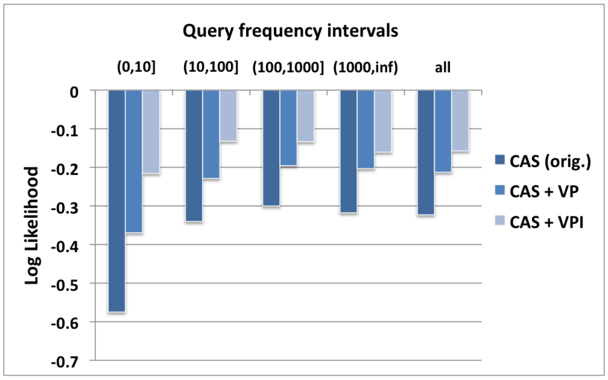
through the advertisement links without assuming a fixed persistence probability for all users.

Runs 3, 5, and 11 substantially outperform their peers in cascade group, DCM group, and DBN group respectively, suggesting that modeling user initiation and persistence better reflects user browsing and click behavior in sponsored search. In order to quantify these relative improvements, we note that a larger value of LL indicates a better model fit, where the ideal value is zero (see Equation 5.20). Given two models M_1 and M_2 with LL values ll_1 and ll_2 respectively, the improvement of the later model over the former one can be computed as (Guo *et al.*, 2009a; Zhu *et al.*, 2010):

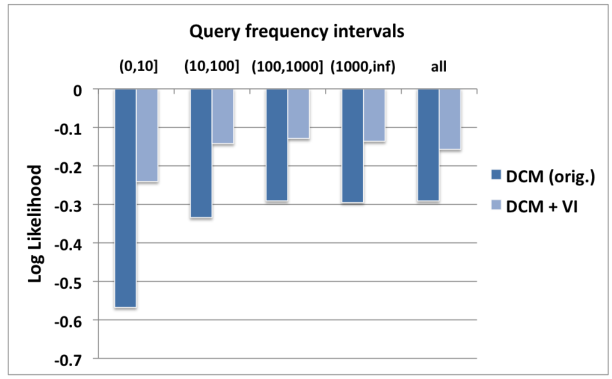
$$imp_{M_1}^{M_2} = \frac{exp(ll_2) - exp(ll_1)}{exp(ll_1)} = exp(ll_2 - ll_1) - 1 \tag{5.22}$$

Using this equation, the improvement of runs 3, 5, and 11 over their corresponding original setting (i.e., runs 1, 4, and 10) is computed as 17.9%, 14.3%, and 15.2%, respectively. In other words, the inclusion of both user initiation probability and varying persistence in examination improves the performance of all the original models.

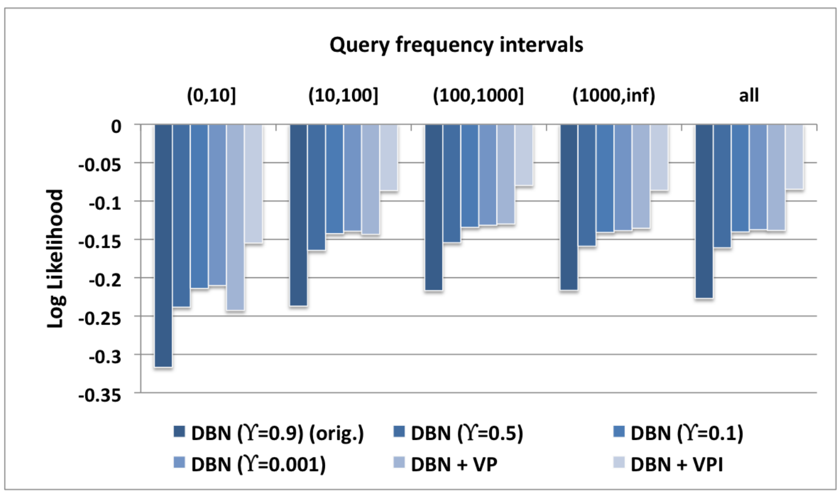
In order to compare the performance of the models at different stages of online training, we considered the queries in the sequence of their SERPs as they appeared in the log and therefore in the learning process. Five intervals of frequency are depicted across the x–axis



(a) Cascade



(b) DCM



(c) DBN

Figure 5.4: The average log-likelihood of studied click models with various settings of query biases and across different query frequencies.

in Figure 5.4, each indicating the number of times that a query has been seen before the current prediction performance is calculated and averaged across all the queries that fall in this query interval. Note that the last interval contains all the queries in set $A^{(2)}$. Except for the early stage of the learning process (i.e. the first interval), the performance of the models across the rest is pretty consistent. As can be seen in these plots, for all the three models, the inclusion of both query biases (i.e. CAS+VPI, DCM+VI, and DBN+VPI runs) results in substantial improvement in the performance of the corresponding models across all the frequency groups.

5.5.5 Evaluating the Impact of the Location Bias in Addition to Query Biases

In this section, we evaluate the performance of click models under the location bias, along with the query biases. For this purpose, set $B^{(2)}$ is used as it contains only those SERPs from set $A^{(2)}$ for which eight ads are displayed. Therefore, the location of these ads is known: the top three appear at the top and the bottom five appear at the side of the corresponding search result page.

Table 5.3: The additional setting related to the location bias for the models under the experiment.

setting	**click models**		
varying persistence and initial motivation for different locations	CAS+VPIL	DCM+VIL	DBN+VPIL

In addition to the settings listed under Table 5.1, a new setting is considered for this round of experiments, in order to study the impact of location bias on the click models. Table 5.3 depicts and labels this setting in order to be referenced along with the ones introduced in Table 5.1. We note again that the variability of the persistence probability is not incorporated into DCM under this setting, since DCM itself has a notion of variability in persistence.

By adding the 3 runs labeled in Table 5.3 to the 11 runs labeled in Table 5.1, a total of 14 runs are generated for this round of the experiments. These runs are numbered from 1 to 14 as depicted in Table 5.4.

To summarize the performance of the 14 runs reported in Table 5.4, the average log-likelihood (LL) across all the rank positions and locations are plotted in Figure 5.5. The

Table 5.4: Runs across different settings of the query biases and the location bias for the cascade model, DCM Model, and DBN model.

model	**Cascade**				**DCM**		
setting	orig.	VP	VPI	VPIL	orig.	VI	VIL
run#	1	2	3	4	5	6	7
model	**DBN**						
setting	$\lambda = 0.9$ (orig.)	$\lambda = 0.5$	$\lambda = 0.1$	$\lambda = 10^{-3}$	VP	VPI	VPIL
run#	8	9	10	11	12	13	14

run ids in the figure match those shown in the table. Each group of models are presented in a different pattern.

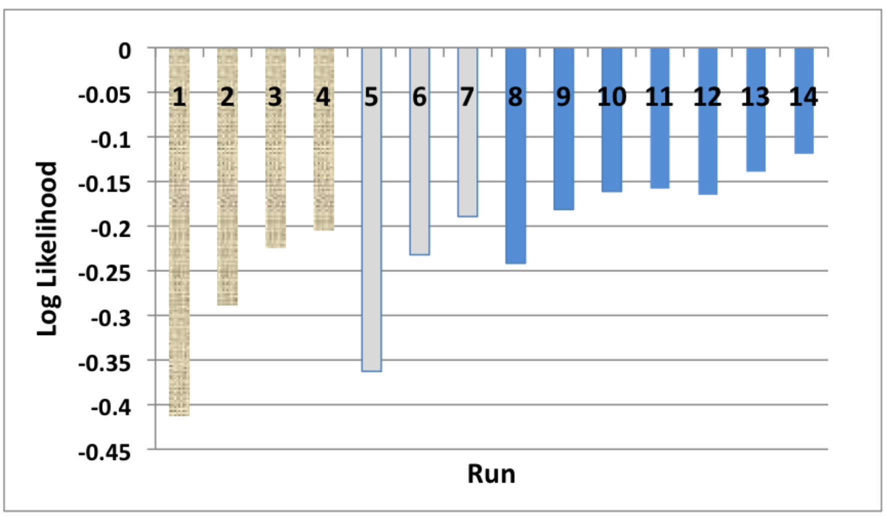


Figure 5.5: The average log-likelihood for the 14 runs.

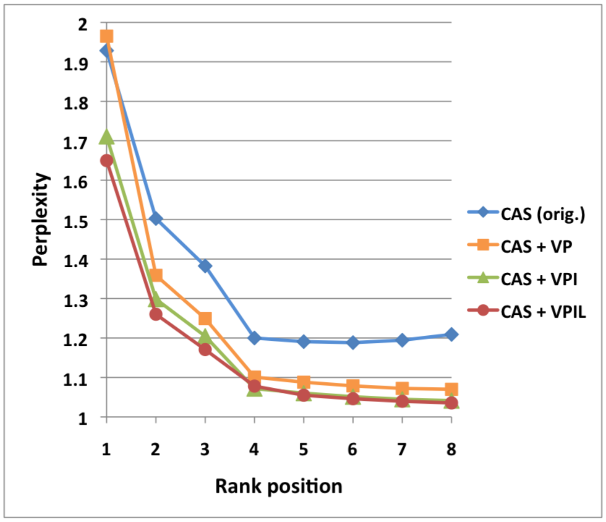
Similar to the experiments performed on set $A^{(2)}$ and depicted in Figure 5.3, the overall performance of the runs over set $B^{(2)}$ report relative superiority of the DBN runs over the other two models in Figure 5.5. This result could be due to the more realistic way of interpreting user behavior under the DBN model, e.g. the consideration of the perceived and post-click relevance, which appears to reflect user behavior in sponsored search as well.

The numbers reported in Figure 5.5 are, generally speaking, lower than the ones reported in Figure 5.3 possibly due to the limitations of the set $B^{(2)}$. Remember that set $B^{(2)}$ has been chosen as a subset of set $A^{(2)}$ in order to be able to identify the precise location of ads on the SERPs.

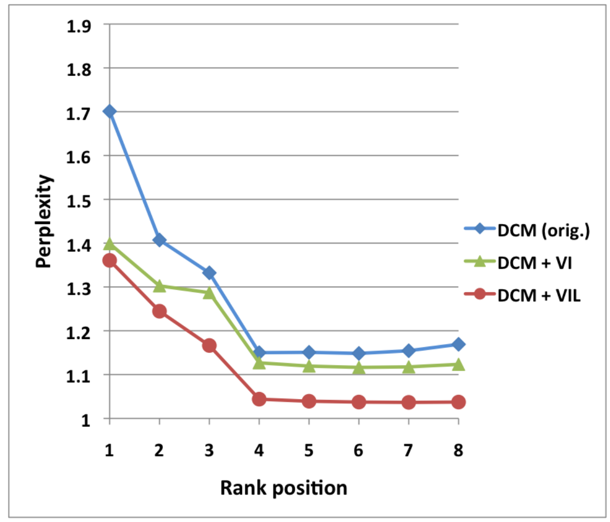
Using Equation 5.22, the improvement of runs 4 (i.e., CAS+VPIL), 7 (i.e., DCM+VIL), and 14 (i.e., DBN+VPIL) over their corresponding original setting (i.e., runs 1, 5, and 8) is computed. The relative improvements are computed as 23%, 19%, and 13% respectively, suggesting that the inclusion of the location bias as well as user initiation probability and varying persistence in examination improves the performance of all the original models.

To further study the impact of introduced biases, especially the location bias, the perplexity measure is used as the log-likelihood powers computed independently at each rank position. A lower value of perplexity indicates a better fit between the model and the actual data. The plots in Figure 5.6 depict these results for each click model under the various settings of the biases. The lines do not imply interpolation. The performance of these models in their original settings is also plotted. We note that the first three rank positions on the x–axis represent the possible positions for the top-listed ads, while the last five (from 4 to 8) represent the side-listed ads.

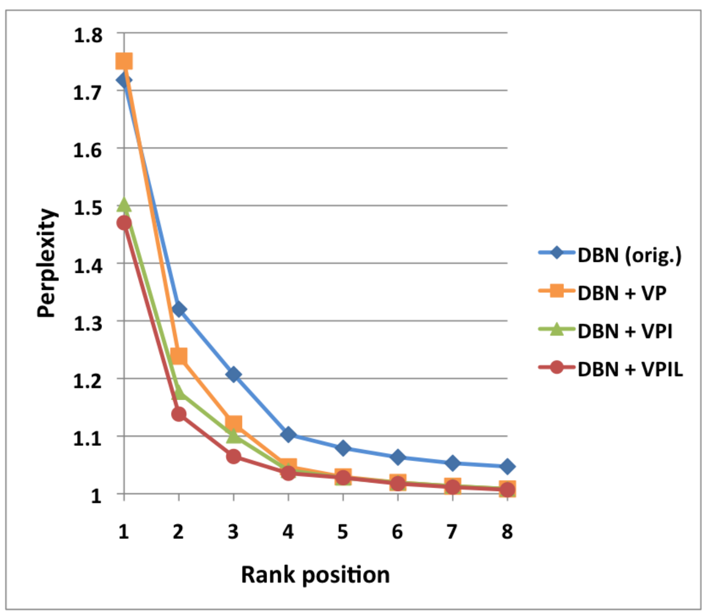
As can be observed in Figure 5.6, for all models, the inclusion of various combinations of the biases results in better performance comparing to the original setting of the models. Among the various settings, the ones that accommodate the location bias as well as the query biases have superior performance in all plots, suggesting that the introduction of the location bias brings further improvement. The superiority of the performance appears stronger for the top-listed ads. This could be due to the fact that the top-listed ads receive the major attention and clicks from the users whereas the click chance over the side ads is much lower resulting in a more sparse samples for the side location.



(a) Cascade model under various settings



(b) DCM model under various settings



(c) DBN model under various settings

Figure 5.6: Impact of various biases on the studied click models.

5.6 Summary

This chapter studies the impact of a group of contextual factors on modeling user behavior in sponsored search, allowing clickthrough analysis in a semi-supervised and online fashion. These contextual factors include the probability that the user will initiate browsing advertisement links at different locations on the page and their persistence in continuing to browse these links. User initiation and persistence are modeled as query biases, while ad placement is modeled as a location bias. A group of existing probabilistic click models are adapted and extended to incorporate these contextual factors. The newly introduced parameters can be learned from click signals recorded in the logs of a commercial search engine.

The evaluation results indicate that significant improvements can be achieved in click prediction once the overall quality of ads shown on a result page, along with location bias and query biases, are taken into consideration. Further investigation in this direction over other well-known click models are among the future directions for this work. Comparing the effectiveness of these factors in sponsored search versus organic search is also among the directions for future work.

Findings of this chapter are obtained without any extra effort carried out to collect ground truth information, such as query intent categories. The main source of training and inference is the log data obtained from the user's search experience on a commercial search engine. Using the trained click models, the objective of the next chapter is to explore whether the contextual parameters that were learned for these models reflect any distinctions in user search intent and behavior. This information may provide a better understanding of user behavior with respect to their query which may be used as helpful signals to better target context-based ad clickthrough prediction in the sponsored search domain.

Chapter 6

Patterns Found in User Behavior

A major purpose of this chapter is to determine if query and location biases can better reflect user browsing and click behavior. To a large extent, the findings of Chapter 5 confirm the superior performance of click models when these biases are incorporated. The current chapter aims to explore whether these biases reflect clear distinctions in user behavior, indicating that they can be helpful signals in future click prediction models for sponsored search.

6.1 User Intent and Query Biases

In order to study the relation between the query bias parameters (initiation probability and persistence probability) and search intent, the set of 4000 queries from Section 3.2 is used to represent the commercial/ non-commercial intents of user queries. Moreover, the set of 2010 commercial queries labeled along the three dimensions of commercial intent (i.e. retailer, product, and brand) are used in this section.

Remember that labels for each query along each dimension (i.e., commercial/non-commercial, brand, product, and retailer) have been obtained from the majority agreements among the annotators. These labels are employed to determine whether parameters of the

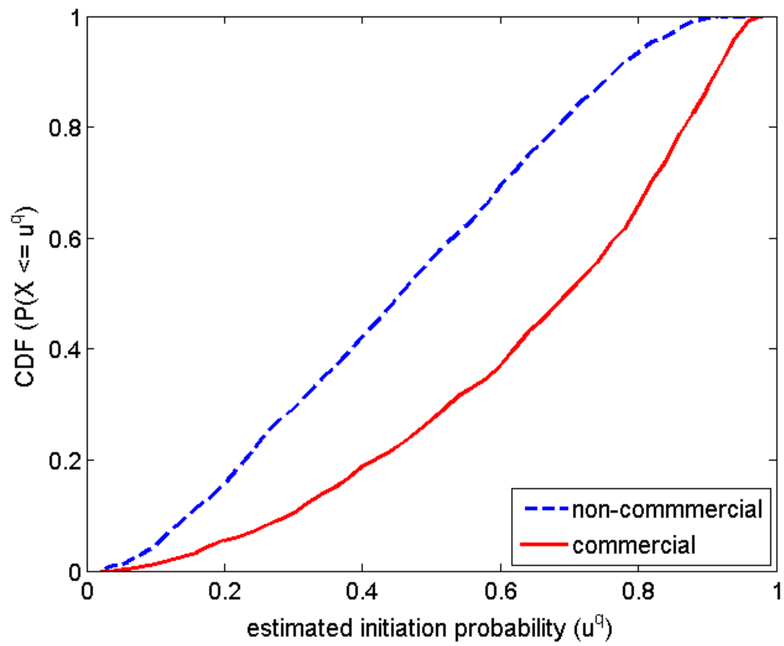


Figure 6.1: Difference in the initiation probability for the commercial/non-commercial dimension.

models that were defined over user queries and learned through expectation maximization actually reflect varying behavioral patterns across different query intent categories.

We specifically focus on the result of DBN+VPI in order to study user intent with respect to the browsing initiation probability (u^q) and persistence probability (λ^q).

For the commercial/non-commercial dimension, the cumulative density function (CDF) is calculated with respect to the values of u^q separately for commercial queries and non-commercial queries. The two CDF curves are depicted in Figure 6.1. Note that each point (u^q, ρ) on a CDF curve indicates that the probability of having an initiation probability value less than u^q is ρ for the corresponding query category. Also note that for each pair of CDFs reported in this chapter, a two-sample KS-test (Papoulis *et al.*, 2002) (Kolmogorov-Smirnov test) was performed at a significance level of 95%, where the difference between the CDFs for each pair is found to be significant.

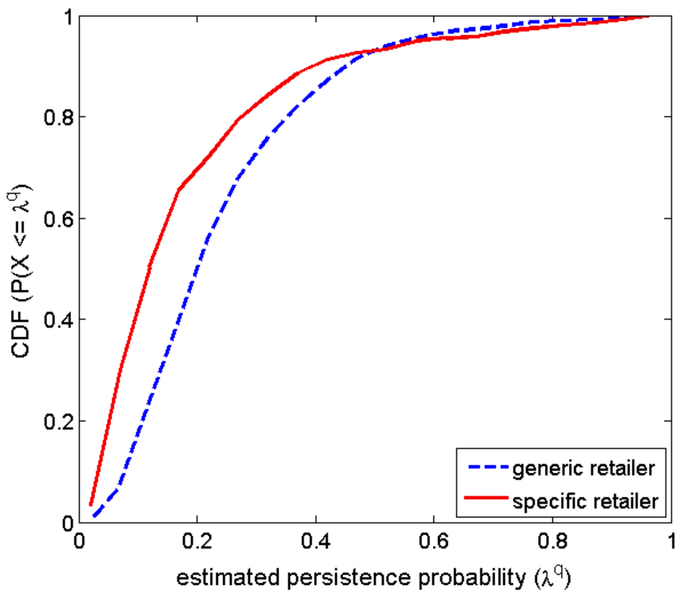
By comparing the trend of the curves in Figure 6.1, one can observe that the commercial category on average has higher probabilities for larger values of u^q compared to non-commercial queries. This observation can be intuitively justified, since a user with commercial intent is more likely to initiate browsing an ad list compared to a user with non-commercial intent.

In addition to the initiation probability, the persistence probability could vary across different query intents. If the commercial intent behind a user's query results in a higher probability to initiate browsing, different aspects of their commercial intent could also result in varying browsing behavior. To illustrate this variability, we calculated the CDF with respect to the λ^q parameter for sub-categories of commercial intent (i.e., retailer, brand, and product).

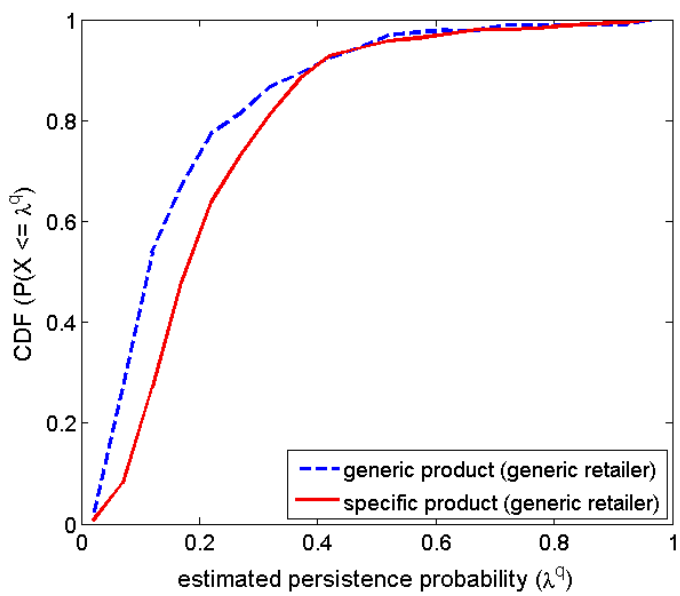
As observed in Figure 6.2a, retailer-specific queries exhibit lower persistence compared to non-retailer-specific queries. This observation is consistent with that of Ghose and Yang (2008) for the impact of keyword attributes on consumer search and purchase behavior. Retailer-specific queries are usually navigational queries, for which the user may be a loyal customer, perhaps looking for information about a particular retailer, and expecting to find this information towards the top of the list.

Figure 6.2b, on the other hand, depicts CDF curves for product-specific queries against the generic ones, such that none of the categories include any particular retailer name in their queries, in order to avoid the impact of the retailer name. It can be seen that users are relatively more persistent in browsing through the ad list if their query names a specific product. This observation could indicate that the user may need a commercial product, but does not yet know where to buy it, providing competitive search situations where the user is more persistent in browsing.

We further study persistence in the product dimension with respect to the presence and absence of the brand information. Figure 6.3 shows CDF curves for the four combinations of product and brand categories. Consistent with earlier observation, queries that reflect specific product names have a relatively higher persistence probability. This observation can be confirmed by comparing the CDF for specific products against that of generic products in the presence of a brand name (specific brand), and also by comparing the



(a) Difference in persistence probability for the retailer dimension.



(b) Difference in persistence probability for the product dimension in absence of the retailer information.

Figure 6.2: Distinctions found in user behavioral parameters across the commercial intent sub-categories.

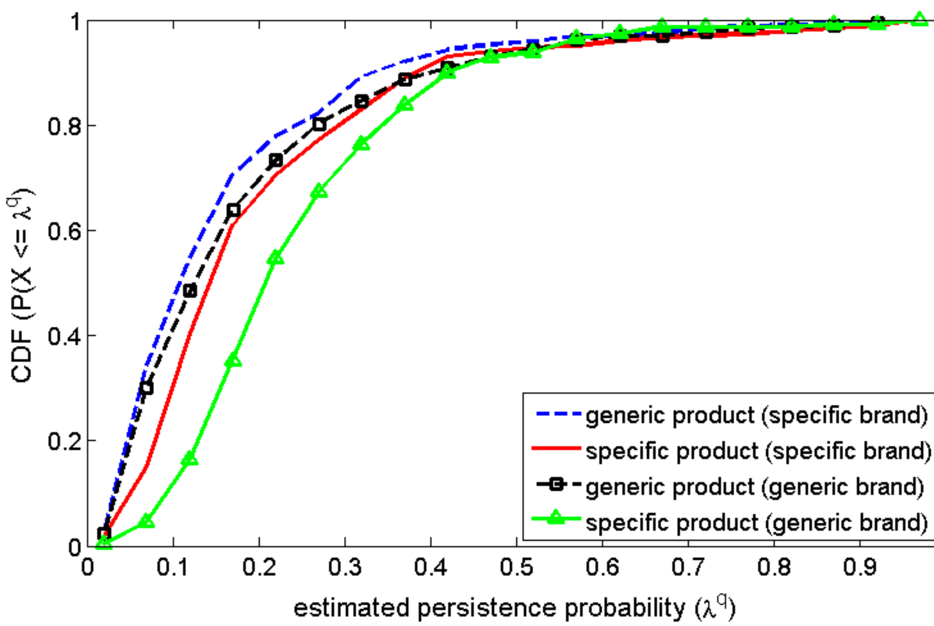


Figure 6.3: Difference in the persistence probability for the product dimension with respect to the brand information.

CDF for specific products against that of generic products in the absence of a brand name (generic brand).

We draw attention to the effect of a brand name on product-specific queries. By comparing the CDFs of the product-specific queries in absence of a brand name (generic brand) against the product-specific queries with a specific brand name, one can observe that once the brand name is included in the query users are less persistent in browsing. The former query type targets a specific commercial product (e.g. running shoes), while the latter targets a specific brand of a specific product (e.g. Nike shoes). One could argue the user who enters the first query would be more engaged in the browsing process, since any brand might do. On the other hand, the user who issues the second query may be a relatively loyal user who is looking for their favorite brand among the top results. If they do not find the brand among the top results, they may either abandon the search or move on to the organic results. This behavior could suggest that showing few but highly targeted ads for a specific brand of a product query is better than showing many ads that reflect various competing brands. Whereas, presenting various ads for competing brands of a product could be more effective for brand-generic product queries.

These sorts of observations are interesting and at the same time may be helpful in the sense that they are obtained from independent experiments. One group of experiments (from Chapter 5) empirically calculates the bias parameters with respect to the click signals recorded in the search engine log. The second round of experiments (in the current Chapter) matches these values against query types that have been obtained independently, and finds distinctive patterns of user behavior based on them. These patterns not only shed light on user behavior, they may also suggest the development of user dependent properties to be used as signals for ad click analysis in sponsored search.

6.2 User Behavior and Location Bias

With respect to the location bias, we study whether the behavioral parameters that are defined for different locations and are learned through the expectation maximization can reflect distinctions in user behavior at different locations on result pages. We focus on the result of DBN+VPIL on set $B^{(2)}$ to address the location bias parameters: u_s^q, u_t^q, λ_s^q, and λ_t^q.

Figure 6.4 depicts two types of comparison results for our purposes: one depicts the sorted values of $u_t^q - u_s^q$ for the corresponding queries, and the other contains the sorted values of $\lambda_t^q - \lambda_s^q$ across the queries. Note that t stands for the top location and s stands for the side location. The idea is to examine whether users behave differently for top-listed ads and side-listed ads in terms of: i) their initiation probability to begin examining ads listed at different locations, and ii) their persistence probability in continuing to examine ads at different locations.

As it is seen in both plots of Figure 6.4, the parameters reflecting user behavior over the top ads appear to be larger than the corresponding values for the side ads for most of the queries. In more detail, the initiation probability for the top location is found to be

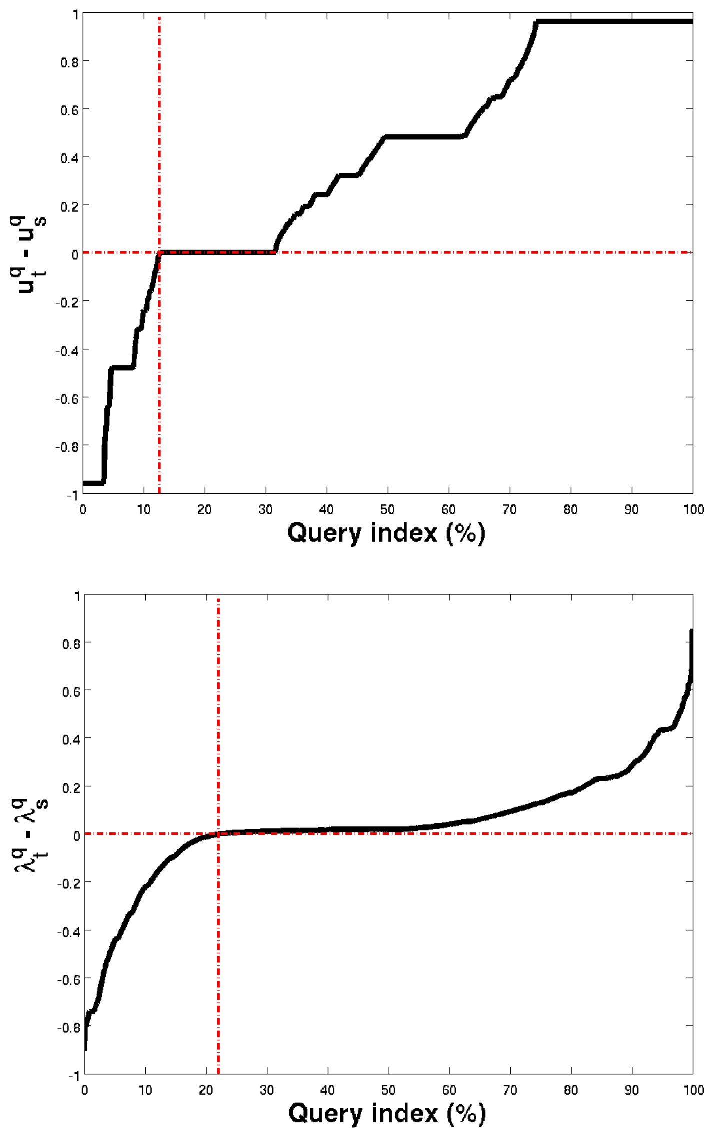


Figure 6.4: Difference in the initiation probability and the persistence probability for various locations of the result pages (t stands for top and s stands for side) and across the sample of queries from set $B^{(2)}$.

higher than or equal to the initiation probability at the side for about 88% of the queries (note the vertical line that shows the threshold of 12%). This number appears to be about 78% with respect to the persistence probability at different locations. In other words, users are found to be more likely to initiate browsing of the top ads compared to the side ads. They are also found to be more persistent in browsing through the top ads as opposed to the side ads.

The hidden parameters u_t^q, u_s^q, λ_t^q, and λ_s^q are learned through the expectation maximization technique, as explained previously, independent of any assumption about the superiority of the top ads over the side ads. However, the results of learning indicate such a superiority in Figure 6.4 for most cases. These observations confirm that users generally pay more attention to the top ads as opposed to the side ads; a signal that can be accommodated by the click models for better understanding of ad clickthrough prediction in sponsored search.

Chapter 7

Conclusions and Future Directions

Understanding the user needs and behaviors from implicit feedback can help search providers improve search personalization and ideally increase user satisfaction. This thesis addresses this problem by modeling the context of search in terms of the query, the nature of the organic results, the list of ads displayed on the result page, and the positioning of the ads on the page, and studies the impact of these contextual factors on user behavior in sponsored search. The context is further augmented by considering the list of ads displayed along with a particular ad on a result page, suggesting the use of cascade model of user behavior. In order to account for variability of user behavior and to account for their bias against the sponsored search domain, a group of biases are introduced to these cascade-based models. The main goal of these models is to gain insight into user browsing and click behavior in sponsored search, which in turn improves one's ability to infer the probability of clicks on the advertisement links. Finally, the extent to which the proposed biases reflect varying behavioral patterns is explored, confirming that correlations exist between the biases and user search behavior.

7.1 Contributions

Overall, this thesis proposes models for ad click analysis and conducts empirical studies based on the behavior of users in the domain of sponsored search. The contributions of the thesis can be summarized from three general perspectives:

- User intent analyses
- Context-based click analyses
- User behavioral analyses

These contributions and findings of the thesis with respect to each contribution are further discussed next.

7.1.1 User Intent Analyses

The initial part of the thesis focuses on modeling and inferring user intent for better understanding of the user click behavior in sponsored search. The newer dimension of query intent, commercial intent, is of particular interest in this work due to the commercial nature of sponsored search. If a query is commercially oriented (i.e. the user may be intending to purchase a product or service), the user may be more likely to click on an ad. On the other hand, if there is no commercial intent underlying the query, the user may not consider the ads at all.

Another aspect of this initial part addresses the gap between this newer dimension of query categories and the traditional categories resulting in the following combinations: commercial-navigational, commercial-information, noncommercial-navigational, and noncommercial-informational. Commercial queries are studied at a finer level by considering whether they reflect a particular brand, product, or retailer, introducing sub-categories of commercial intent.

The primary empirical study conducted in Chapter 3 indicates that strong commercial terms, such as “sale”, “cheap”, and “store” appear among the most frequent terms in the clicked queries. A probabilistic model is proposed that confirms the contributions of individual terms and their ad clicks towards commercial intent of the queries. Nevertheless, using click information for intent analysis does not appear to be an effective option for three reasons: i) obtaining click log information is costly, so an approach based on click information can not be easily replicated and used as a general framework, ii) although click information from the sponsored search domain may work for commercial intent detection, it does not appear to be as effective for the traditional categories of query intent, and iii) one of the main objectives of this thesis is to study the characteristics of various query intents in terms of their impact on predicting clickthrough and user preferences, hence relying on the click log information may create circularity in this process.

Therefore, relying on the query string itself appears to be among the promising sources of information to infer the intent of user queries. Since the query is short and it may not reveal much about the user, additional sources of information are used to enrich the query. A methodology is developed for using query specific information and the content of search engine result pages (SERPs) to identify the intent underlying user queries in various dimensions. There are two settings addressed in the thesis: i) query features are combined with SERP features, and ii) SERP features are used alone.

Samples of labeled queries are needed as ground truth in order to train a binary classifier in each dimension of query intent and evaluate the inference results. Therefore, a semi-automatic approach is proposed for labeling a batch of training queries in various dimensions of query intent by using crowdsourcing. This approach addresses two challenges encountered in the annotation process via crowdsourcing: i) combining annotators' contributions automatically and effectively, and ii) evaluating the annotation results.

Consistency of annotation results and accuracy of inference results discussed in Chapter 3 and the distinction in click behavior (resulted from various query types) discussed in Chapter 4 validate the query intent studies conducted in the initial part of the thesis. Overall, findings of this part of the thesis can be summarized as follows:

- Variability in the user's search intent is reflected through their queries, and it can be used as a means to model the variability in user's behavior over the advertisement links. We confirm that terms with respect to their click history in sponsored search are effective in detecting the commercial intent of queries that include such terms.

- The agreement obtained among the annotators along the different dimensions of query categories, and the accuracies obtained from the trained query classifiers, confirm that the studied query categories are reasonably distinguishable.

- Features extracted from the query string, along with the contents of the search engine result pages, are found to be effective in detecting query intent.

7.1.2 Context-Based Click Analyses

The content of ads is the primary source of information used in the literature to study click behavior over advertisement links. In the current thesis, factors beyond those extracted from ad content form the targets of study. Variability in click behavior over the advertisement links is therefore studied with respect to the context of search result pages that accommodate these links. This context can include the number of ads displayed on the result page, the positioning of ads on the page, and the user intent underlying the queries with which these ads appear. The intuition behind this idea is that ads positioned at the top of a page may receive more clicks, even if they are less relevant than other ads. Furthermore, a weakly related ad appearing with results of a commercially oriented query may receive more clicks than a strongly related ad appearing with the results of a less commercially oriented query.

Given a group of contextual factors, a notion of *context CTR* is proposed with respect to the search engine result pages (SERPs) that have these factors in common. The context CTR is defined as the ratio of the total number of clicks recorded for SERPs that have a particular group of contextual factors in common to the total number of appearances of such SERPs. This empirical metric is used for two purposes in the thesis:

- It is first intended to evaluate the performance of SERPs with respect to various contextual factors, and

- When aggregated over the historical impressions of a particular ad, it is intended to infer the expected quality of the ad across these impressions.

The query intent categories have been obtained from the results of the initial part of the work. The available log data supply the rank for ads, along with the total number of ads displayed on a page, but it does not record the precise locations of ads. Using statistical analysis, the click probability for different locations of a page is estimated from the log data. There are two attempts proposed for this purpose in Chapter 4. In the main attempt, a probability distribution is defined for the average clickthrough rate with respect to the studied contextual factors, including the location of ads. Since a click on an ad is the result of user interactions with a result page, various possibilities of this distribution can be seen as results of the dynamic nature of human interactions. Hence, a solution with maximum randomness is assumed to be reasonable. Thus, the entropy of the distribution, as a measure of randomness and uncertainty, is maximized in order to obtain a stable state of the system and an answer for the distribution.

Using the estimate of click probability for different locations, and the empirical estimate of click probability with respect to the other contextual factors, the findings of the work according to the first objective can be summarized as follows:

- The number of ads displayed on result pages appears to show correlation with the number of ad clicks recorded for these pages.

- In general, the placement of ads appears to have a substantial impact on the number of clicks they receive. In particular, ad clicks appear to mostly occur at the first and the second ranks, and most especially at the first rank.

- User click behavior on ads is found to be distinct for different categories of query intent, and this can indicate that the click behavior is consistent with the classification results of the general categories of query intent as explained in Chapter 3. Generally speaking, categories that involve commercial intent are the leaders among the others. This result confirms that the commercial categories of queries receive more ad clicks comparing to the others.

- Certain click behavior of different users can be justified according to their query intent. For example, we show that SERPs associated with commercial- navigational queries attract more ad clicks comparing to the SERPs associated with commercial-informational queries. A query with commercial- navigational intent may indicate a relatively more focused and goal-directed search (Danaher and Mullarkey, 2003) (the user knows the retailer of the commercial product that they want), resulting in a higher chance of a click or conversion from the user.

- Among the three sub-categories, only for the retailer category, when a specific retailer is implied by the query intent, the number of clicks for varying number of ads is always higher compared to the case where the retailer is unknown. In other words, ads are placed in a way that the ones that reflect retailer intent are more of a target of clicks than the others.

- Further investigation on the placement of ads confirms that ads displayed on top of result pages are more often the targets of clicks than the ads displayed at the side.

- The difference between the clickthrough rates of top ads and side ads becomes lower when it comes to the leading query categories (i.e. commercial-navigational and commercial). This observation may indicate that when the intent underlying the query is commercial, the effect of the location of ads becomes less significant. However, ads at the top are still the main targets of clicks.

Overall, the above findings suggest that contextual factors, such as the intent underlying user's queries, the total number of ads displayed on a result page, and the rank positions of ads result in varying click behavior for the associated result pages. These contextual factors are therefore assumed to be effective in estimating the clickthrough rate for an ad that appears within a context. Hence, two models are presented that target ads within the context of the SERPs on which they appear. These models estimate the clickthrough rate of an ad as the overall probability of click that it is expected to receive across various contexts in which it is displayed in the history of its appearances.

In the first model, referred to as the *baseline model*, the context is defined according to the SERP/ad pair for each appearance of an ad (i.e., impression), where the context is represented by the particular number of ads that are listed on the page and by the rank position of the ad on this page. In the second model, referred to as the *query intent model*, we study the impact of the identified query intent as an extra factor in the baseline model. Comparing the performance of the baseline model against the intent model suggests that the inclusion of query intent information as a contextual factor provides a better estimation of the ad's quality.

Overall, the findings of context-based ad click analysis suggest that ad clickthrough prediction techniques could benefit from the query intent information and other contextual factors. However, there still remains questions and limitations that need to be addressed. For instance, what if there are other categories of query intent that should be considered? What if there exists a better taxonomy of commercial intent than the sub-categories introduced earlier? Even if an extensive taxonomy is obtained such that a broad range of context is covered, efforts must be carried out to label queries in various dimensions of query intent. In this way, the context model needs to be expanded across various dimensions and training data needs to be collected across various contexts.

These all provided motivation to use the earlier findings as evidence for modeling user browsing and click behavior in a semi-supervised and online fashion, which is the focus of the last part of the thesis. Instead of employing the explicit judgements of the query intent, the contextual factors will be modeled through various query– and page–dependant parameters. These parameters are learned and updated in an online fashion.

7.1.3 User Behavioral Analyses

The main assumption here is based on the cascade model of user behavior (Craswell *et al.*, 2008) in which a document is assumed to be seen only if the user scans over all the ones above. This concept can be extended into the domain of sponsored search such that the rate at which an ad is viewed and clicked is assumed to depend both on its own quality and on the quality of the other ads that are displayed above it on the result page.

Other aspects of user browsing behavior that is addressed in this part of the thesis and studied in Chapter 5 include: i) variability in the behavior of different users, ii) differences that may exist between how users behave in the sponsored search domain as opposed to the organic search, and iii) ad placement strategy on the result page.

The first criterion suggests that browsing behavior varies across different users in sponsored search. One approach would be to deal with various pairs of user/query differently. Since this approach brings too many possibilities, a simpler approach would be to assume that users who enter the same query have similar behavior. The latter case is addressed in this thesis. The second criterion suggests the consideration of the user's bias against sponsored search (Jansen and Resnick, 2006). For this purpose, query biases are introduced into the cascade model. The first query bias deals with the initiation of browsing over advertisement links, while the second one reflects persistence in browsing advertisement links. Finally, the third criterion accounts for the user's response to the page structure, specifically the locations on the page where ads appear, such as the top and side, and the ordering or ranking of ads at each location. Location bias has been introduced to address these issues.

The location and query biases are modeled and formulated through various parameters that all depend on the user's query in order to reflect variability of user behavior. In particular, we assume that if a user initially has any motivation to consider ads on a particular location of the page then they start examining ads. This initial motivation may vary across different users. Moreover, we assume that different users have different levels of patience (persistence) in browsing through an ad list placed at a particular location of the result page.

These parameters are learned through an expectation-maximization technique that maximizes the log-likelihood of the click signals observed from the logs of a commercial search engine. A group of cascade-based click models are extended with respect to the proposed biases and applied in the sponsored search domain. The main goal here is to understand whether any of these models and settings are able to better predict user behavior on advertisement links. The first group of settings reflect the query biases (initiation probability and persistence). The second group of settings reflects query and location biases together.

The final study aims to explore whether these biases reflect any distinction in user behavior. The bias parameters learned from the click signals earlier are matched against query types that have been obtained independently. In particular, the relation between query bias parameters (initiation probability and persistence probability) and search intent are studied. Furthermore, the location parameters are used to study distinctions in user behavior at different locations on result page.

Overall, the findings of the last part of the thesis can be summarized as follows:

- Empirical studies confirm that the inclusion of various combinations of query and locations biases results in better performance in click prediction comparing to the original settings of the cascade-based models. Among the various settings, the ones that accommodate the location bias as well as the query biases have superior performance.

- The superiority of the performance of the extended models appears stronger for top-listed ads, which is explained due to the fact that the top-listed ads receive the major attention and clicks from the users whereas the click chance over the side ads is much lower resulting in a more sparse samples for the side location.

- Separate empirical studies from Chapter 6 confirm that the learned bias parameters reflect distinctions in user behavior, suggesting that they can be used as helpful signals in future click prediction models for sponsored search.

- Some of the interesting patterns found in user behavior can be summarized as follows: A user with commercial intent is more likely to initiate browsing an ad list compared to a user with non-commercial intent. Different aspects of the user's commercial intent can result in varying browsing behavior. For instance, retailer-specific queries exhibit lower persistence compared to non-retailer-specific queries, suggesting that the user may be a loyal customer, perhaps looking for information about a particular retailer, and expecting to find this information towards the top of the list. It is also found that users are relatively more persistent in browsing through the ad list if their query names a specific product, suggesting that the user may need a commercial product, but does not yet know where to buy it, providing competitive search situations where the user is more persistent in browsing. Finally, studying the effect of brand name on product-specific queries indicates that once the brand name is included in the query users are less persistent in browsing, suggesting that showing few but highly targeted ads for a specific brand of a product query is better than showing many ads that reflect various competing brands. Whereas, presenting various ads for competing brands of a product could be more effective for brand-generic product queries.

Using click signals to model user browsing behavior and to learn parameters of the model appear effective to improve understanding of the click behavior of users in sponsored search. The patterns found in user browsing behavior not only shed light on the way user targets sponsored search, they may also suggest the development of user dependent properties to be used as signals for ad click analysis in this domain.

7.2 Future Directions

In this thesis, query intent is considered in terms of the popularity of user needs. In other words, query intent is defined as the most popular understanding of the information need of a typical user in Web search. For this reason, throughout the labeling process, the annotators were asked to judge the assumed commercial intent of the search queries from the perspective of a general user. However, different users may issue the same query and have varying information needs, suggesting that ambiguity exists in query intent identification. Novelty and diversity (Agrawal *et al.*, 2009; Clarke *et al.*, 2008) behind user's queries is another aspect that could be targeted in sponsored search. It is unrealistic to assume that the relevance of an ad to the commercial need of a user is independent of the neighbor ads displayed at the same time. As a result of this assumption, redundant ads, in terms of their topic coverage, may be shown for a given query. For example, the ads for the query "Banff Alberta" should have diversity, covering different topics about Banff, such as "Hotels in Banf", Vacation Packages in Banff, "Hot Springs in Banff", and "Banff Helicopter Tour" rather than dedicating several ads to pages specifically on hotel packages for Banff.

As Web queries have different meanings for different users, the results shown for queries should reflect the diversity in various query topics. Abandonment of ads appears to be more likely than the regular Web results due to the users bias against ads or to the lack of coverage of the topic of interest. Returning similar ads of a strong individual relevancy to a given query may produce a high score on a standard evaluation measure (Aslam *et al.*, 2005; Sakai, 2009), but would certainly be viewed unfavorably by a user who might be interested in a possibly rarer but rather different aspect of the same query, and therefore it may not score hight on intent-aware evaluation measures (Ashkan and Clarke, 2011; Clarke *et al.*, 2011). Possible future work in this direction could evaluate ads in a context that takes the diversity of displayed results into account, as well as their relevancy.

The query intent analysis in this work is limited to the commercial category of query intent, its sub-categories in terms of brand/retailer/product information, and the traditional categories of query intent, navigational and informational. This study can be seen as an initial step towards a long-term goal of extending the traditional categories of Web queries by developing and evaluating the seeds of a taxonomy for commercial search. Expanding this taxonomy, studying and comparing the clickthrough behavior for different dimensions of commercial intent, represents a future direction for this work.

The empirical evaluations of this study is limited to a data set from a single commercial search engine. An interesting experiment would be to replicate this work over other sources of data and compare the results. The available data consists of a sample of SERPs from a larger pool of data, but in the thesis it is assumed that the sample is independent and identically distributed (i.i.d.). Once the SERPs are sorted based on their time, they are treated as a sequence of result pages that can be targeted in an online setting. The complete pool of data with the sorted SERPs would provide a better sample of the reality.

Furthermore, the limitations in the available data set brings some limitations to the experimental studies performed in the thesis. For instance, the location of the ads is not recorded in the data, which creates ambiguity in location-based analysis. For this reason, in the last round of experiments in Chapter 5, a sample of search result pages with eight ads displayed are used such that the location of ads can be certainly identified.

Another point that can be addressed in future work is to explore the extent that contextual factors would be helpful in clickthrough analysis. An effective way of evaluating the performance of contextual factors would be to study the performance of a general click prediction model that works based on the content of the ads once in presence of context-based information and once in absence of this information, and compare them against each other. This could not be studied in the thesis due to the lack of information about the content of ads in the data.

The evaluation results from Chapter 5 indicate that significant improvements can be achieved in click prediction once the overall quality of ads shown on a result page, along with location bias and query biases, are taken into consideration. Further investigation in this direction over other well-known click models are among the future directions for this work. Comparing the effectiveness of these factors in sponsored search versus organic search is also among the directions for future work.

With respect to the user behavior modeling studies conducted towards the end of the thesis, the simplifying assumptions regarding a user's approach to browsing an ad list introduces limitations into the work. Instead of linearly browsing through the list, a user may randomly view an ad at a particular rank position or location, or they may move up and down in the list during their browsing session. However, the cascade assumption of linearly browsing enables us to represent the behavior of the majority of users, which may be considered as a reasonable starting point to better understand user behavior in sponsored search. It also enables us to model ads in the context of the preceding ads. More complex models are required in the future in order to address random viewing and skipping over different positions in the ad list.

Last, but not least, variability of user behavior is modeled through the parameters defined over queries by assuming that users issuing the same query have generally similar behavior. As indicated by Carterette *et al.* (2012) no two users interact with a system in exactly the same way. While a query-based representation of users appears to be effective in the domain of sponsored search (Yan *et al.*, 2009), a more realistic assumption would define the parameters over user/query pairs.

Appendix

Appendix A

Details of Parameter Estimation for the Location– and Query–Aware Model

Details on the inference algorithm used throughout Chapter 5 are provided in this appendix. Note that the variables across this section are mostly used in their general form. By adding a superscript j to the variables, the same formulations can be used for a particular SERP j from the search log.

A.1 Forward-Backward Variables

Following the forward-backward algorithm (Rabiner, 1989) and the details provided in (Chapelle and Zhang, 2009), the forward variable $\alpha_{l,i}$ and the backward variable $\beta_{l,i}$ can be defined as follows:

$$\begin{aligned}\alpha_{l,i}(e) &= P(C_{l,1}, ..., C_{l,i-1}, E_{l,i} = e|\theta) \\ \beta_{l,i}(e) &= P(C_{l,i}, ..., C_{l,N_l}|E_{l,i} = e, \theta)\end{aligned} \tag{A.1}$$

where $\alpha_{l,i}(e)$ is the probability of the partial click observations sequence $C_{l,1}, ..., C_{l,i-1}$ and the examination state e for the ad a listed at rank position i and location l, given the parameters of the model, $\theta = (u_t^q, u_s^q, \lambda_t^q, \lambda_s^q, \omega_a^q, \nu_a^q)$. Similarly, $\beta_{l,i}(e)$ is the probability of the partial click observations sequence $C_{l,i}, ..., C_{l,N_l}$ given the user's examination state e for ad a and θ. The recursion formula for these variables in our problem setting can be stated, as follows:

$$\alpha_{l,i+1}(e) = \sum_{e' \in \{0,1\}} \alpha_{l,i}(e') P(E_{l,i+1} = e, C_{l,i} | E_{l,i} = e')$$
$$\beta_{l,i-1}(e) = \sum_{e' \in \{0,1\}} \beta_{l,i}(e') P(E_{l,i} = e', C_{l,i-1} | E_{l,i-1} = e)$$

where the conditional probability $P(E_{l,i+1}, C_{l,i}|E_{l,i})$ can be computed based on the DBN model (Chapelle and Zhang, 2009) and adopted to our setting by using Equations 5.5 to 5.8, as follows:

$$P(E_{l,i+1}, C_{l,i}|E_{l,i}) = \sum_{s \in \{0,1\}} P(E_{l,i+1}|S_{l,i} = s, E_{l,i}) P(S_{l,i} = s|C_{l,i}) P(C_{l,i}|E_{l,i}) \qquad \text{(A.2)}$$

Table A.1 depicts the values of this conditional probability computed according to the above Equation and based on the possible values for the variables.

Table A.1: The probability distribution for $P(E_{l,i+1}, C_{l,i}|E_{l,i})$.

$E_{l,i+1}$	$C_{l,i}$	$E_{l,i}$	$P(E_{l,i+1}, C_{l,i}\|E_{l,i})$
0	0	0	1
0	0	1	$(1 - \lambda_l^q)(1 - \omega_a^q)$
0	1	0	0
0	1	1	$\omega_a^q \ (1 - \lambda_l^q + \nu_a^q \lambda_l^q)$
1	0	0	0
1	0	1	$\lambda_l^q \ (1 - \omega_a^q)$
1	1	0	0
1	1	1	$\lambda_l^q \omega_a^q (1 - \nu_a^q)$

Finally, the base cases for the forward and backward variables can be obtained with respect to user initiation probability stated in Equation 5.2, as follows:

$$\alpha_{l,1}(0) = 1 - u_l^q, \qquad \alpha_{l,1}(1) = u_l^q$$

$$\beta_{l,N_l+1}(0) = 1, \qquad \beta_{l,N_l+1}(1) = 1$$

where N_l is the number of ads appearing at the location l of the page.

A.2 Estimating Posterior Probabilities

We use forward and backward variables to compute the posterior probability of the transition state ($E_{l,i}$) and the satisfaction variable ($S_{l,i}$) that are among hidden variables of the model.

A.2.1 Transition Probability

Given the click sequence C_l observed for the ads listed on the location l of a result page, the posterior probability of the user examining or not examining an ad listed at the rank position i can be formulated using Bayes' rule as:

$$P(E_{l,i} = e|C_l, \theta) = \frac{P(C_l|E_{l,i} = e, \theta)P(E_{l,i} = e|\theta)}{P(C_l|\theta)}$$

According to the Markov property, given a state, past observations are independent of the future observations. In our case, given the user's decision about examining the ad at rank i, the previous click events are independent of the future ones. This allows us to express the above equation as follows:

$$\begin{aligned} & P(E_{l,i} = e|C_l, \theta) \\ = \quad & \frac{P(C_{l,1}, ..., C_{l,i-1}|E_{l,i} = e, \theta)P(C_{l,i}, ..., C_{l,N_l}|E_{l,i} = e, \theta)P(E_{l,i} = e|\theta)}{P(C_l|\theta)} \\ = \quad & \frac{P(C_{l,1}, ..., C_{l,i-1}, E_{l,i} = e|\theta)P(C_{l,i}, ..., C_{l,N_l}|E_{l,i} = e, \theta)}{P(C_l|\theta)} \end{aligned}$$

where the numerator consists of the forward and backward variables, and the denominator is a normalization factor to make $P(E_{l,i} = e|C_l, \theta)$ a probability measure such that

$\sum_{e\in\{0,1\}} P(E_{l,i} = e|C_l, \theta) = 1$. As a result, the conditional probability can be estimated using the forward-backward algorithm.

$$P(E_{l,i} = e|C_l, \theta) = \frac{\alpha_{l,i}(e)\beta_{l,i}(e)}{\sum_{e''\in\{0,1\}} \alpha_{l,i}(e'')\beta_{l,i}(e'')} \tag{A.3}$$

The probability of examination given the model can be similarly estimated, as follows:

$$\begin{aligned}
& P(E_{l,i} = e|\theta) && \text{(A.4)}\\
= & \sum_{C_l} P(C_l, E_{l,i} = e|\theta) \\
= & \sum_{C_l} P(C_l|E_{l,i} = e, \theta)P(E_{l,i} = e|\theta) \\
= & \sum_{C_l} P(C_{l,1}, ..., C_{l,i-1}|E_{l,i} = e, \theta)P(C_{l,i}, ..., C_{l,N_l}|E_{l,i} = e, \theta)P(E_{l,i}|\theta) \\
= & \sum_{C_l} P(C_{l,1}, ..., C_{l,i-1}, E_{l,i} = e|\theta)P(C_{l,i}, ..., C_{l,N_l}|E_{l,i} = e, \theta) \\
= & \sum_{C_l} \alpha_{l,i}(e)\beta_{l,i}(e)
\end{aligned}$$

Finally, the probability of user being in the examination state e at rank i and in state e' at rank $i+1$, given the model and the click observations sequence C_l, can be expressed as:

$$\begin{aligned}
& P(E_{l,i} = e, E_{l,i+1} = e'|C_l, \theta) && \text{(A.5)}\\
= & \frac{P(E_{l,i} = e, E_{l,i+1} = e', C_l|\theta)}{P(C_l|\theta)} \\
= & \frac{1}{P(C_l|\theta)} \times P(C_{l,1}, ..., C_{l,i-1}, E_{l,i} = e|\theta) \;\times\; P(E_{l,i+1} = e', C_{l,i}|E_{l,i} = e, \theta) \\
& \quad \times\; P(C_{l,i+1}, ..., C_{l,N_l}|E_{l,i+1} = e'|\theta) \\
= & \frac{\alpha_{l,i}(e)\; P(E_{l,i+1} = e', C_{l,i}|E_{l,i} = e)\; \beta_{l,i+1}(e')}{\sum_{e''\in\{0,1\}} \alpha_{l,i}(e'')\beta_{l,i}(e'')}
\end{aligned}$$

As a result, $P(E_{l,i+1} = 1, E_{l,i} = 1|C_l)/P(E_{l,i} = 1|C_l)$ can be computed by dividing Eq. A.5 by Eq. A.3:

$$\begin{aligned} & \frac{P(E_{l,i+1} = 1, E_{l,i} = 1|C_l)}{P(E_{l,i} = 1|C_l)} \qquad \text{(A.6)} \\ = & \frac{\frac{1}{P(C_l|\theta)}\ \alpha_{l,i}(1)\ P(E_{l,i+1} = 1, C_{l,i}|E_{l,i} = 1)\ \beta_{l,i+1}(1)}{\frac{1}{P(C_l|\theta)}\ \alpha_{l,i}(1)\ \beta_{l,i}(1)} \\ = & \frac{\beta_{l,i+1}(1)}{\beta_{l,i}(1)}\ P(E_{l,i+1} = 1, C_{l,i}|E_{l,i} = 1) \end{aligned}$$

where e and e' from Equations A.3 and A.5 are substituted by 1, and $P(E_{l,i+1} = 1, C_{l,i}|E_{l,i} = 1)$ can be estimated according to Equation A.2.

A.2.2 Satisfaction Probability

Define $\phi(S_{l,i})$ as the posterior distribution of the variable $S_{l,i}$ given the click sequence C_l observed on the location l of a SERP where ad a displayed at rank i on this location. In order to compute this posterior probability, two cases need to be considered:

Case 1: If ad a is not clicked, i.e. $C_{l,i} = 0$, then according to Equation 5.8, $S_{l,i} = 0$. Thus, the posterior probability $\phi(S_{l,i} = 1) = 0$ in case of no click at rank i.

Case 2: If ad a is clicked, i.e. $C_{l,i} = 1$:

$$\begin{aligned} \phi(S_{l,i} = 1) &= P(S_{l,i} = 1|C_l, \theta) \\ &= \sum_{e \in \{0,1\}} P(S_{l,i} = 1|E_{l,i+1} = e, C_l, \theta)P(E_{l,i+1} = e|C_l, \theta) \\ &= P(S_{l,i} = 1|E_{l,i+1} = 0, C_l, \theta)P(E_{l,i+1} = 0|C_l, \theta) \end{aligned}$$

where the last substitution above is permitted because of Equation 5.6, which indicates that $S_{l,i} = 1 \Rightarrow E_{l,i+1} = 0$, and therefore, $P(S_{l,i} = 1|E_{l,i+1} = 1, C_l, \theta) = 0$. Once the effect of the examination state at rank i is taken into account, the above equation can be further simplified as follows:

$$
\begin{aligned}
&\phi(S_{l,i}=1) \\
= \quad & \frac{P(S_{l,i}=1, E_{l,i+1}=0, C_l, \theta)}{P(E_{l,i+1}=0, C_l, \theta)} P(E_{l,i+1}=0|C_l, \theta) \\
= \quad & \frac{\sum_{e\in\{0,1\}} P(S_{l,i}=1, E_{l,i+1}=0, E_{l,i}=e, C_l, \theta)}{\sum_{e\in\{0,1\}} P(E_{l,i+1}=0, E_{l,i}=e, C_l, \theta)} P(E_{l,i+1}=0|C_l, \theta) && \text{apply } C_{l,i}=1 \Rightarrow E_{l,i}=1 \\
= \quad & \frac{P(S_{l,i}=1, E_{l,i+1}=0, E_{l,i}=1, C_l, \theta)}{P(E_{l,i+1}=0, E_{l,i}=1, C_l, \theta)} P(E_{l,i+1}=0|C_l, \theta) && \text{apply } S_{l,i}=1 \Rightarrow E_{l,i+1}=0 \\
= \quad & \frac{P(S_{l,i}=1, C_l, E_{l,i}=1, \theta)}{P(E_{l,i+1}=0, C_l, E_{l,i}=1, \theta)} P(E_{l,i+1}=0|C_l, \theta) \\
= \quad & \frac{P(S_{l,i}=1, C_{l,1}, ..., C_{l,i}=1, ..., C_{l,N_l}, E_{l,i}=1, \theta)}{P(E_{l,i+1}=0, C_{l,1}, ..., C_{l,i}=1, ..., C_{l,N_l}, E_{l,i}=1, \theta)} P(E_{l,i+1}=0|C_l, \theta) && \text{independence assumption} \\
= \quad & \frac{P(S_{l,i}=1, C_{l,i}=1, E_{l,i}=1, \theta)}{P(E_{l,i+1}=0, C_{l,i}=1, E_{l,i}=1, \theta)} P(E_{l,i+1}=0|C_l, \theta) && \text{apply chain rule} \\
= \quad & \frac{P(S_{l,i}=1|C_{l,i}=1, E_{l,i}=1, \theta) P(C_{l,i}=1|E_{l,i}=1, \theta) P(E_{l,i}=1|\theta)}{P(E_{l,i+1}=0, C_{l,i}=1|E_{l,i}=1, \theta) P(E_{l,i}=1|\theta)} P(E_{l,i+1}=0|C_l, \theta) \\
= \quad & \frac{P(S_{l,i}=1|C_{l,i}=1, E_{l,i}=1, \theta) P(C_{l,i}=1|E_{l,i}=1, \theta)}{P(E_{l,i+1}=0, C_{l,i}=1|E_{l,i}=1, \theta)} P(E_{l,i+1}=0|C_l, \theta) && \text{apply } C_{l,i}=1 \Rightarrow E_{l,i}=1 \\
= \quad & \frac{P(S_{l,i}=1|C_{l,i}=1, \theta) P(C_{l,i}=1|E_{l,i}=1, \theta)}{P(E_{l,i+1}=0, C_{l,i}=1|E_{l,i}=1, \theta)} P(E_{l,i+1}=0|C_l, \theta) && \text{apply Eq. 5.5, 5.6, and A.2} \\
= \quad & \frac{\nu_a^q \omega_a^q}{\omega_a^q\ (1-\lambda_l^q+\nu_a^q\lambda_l^q)} P(E_{l,i+1}=0|C_l, \theta) \\
= \quad & \frac{\nu_a^q}{(1-\lambda_l^q+\nu_a^q\lambda_l^q)} P(E_{l,i+1}=0|C_l, \theta) && \text{(A.7)}
\end{aligned}
$$

where $P(E_{l,i+1} = 0|C_l, \theta)$ can be estimated according to Equation A.3. Overall, the posterior probability for the variable $S_{l,i}$ can be summarized, as follows:

$$\begin{aligned}
\phi(S_{l,i} = 1) &= P(S_{l,i} = 1|C_l, \theta) \\
&= \begin{cases} 0 & \text{if there is no click on ad } a \text{ at rank } i \\ \frac{\nu_a^q}{(1-\lambda_l^q+\nu_a^q\lambda_l^q)} P(E_{l,i+1} = 0|C_l, \theta) & \text{otherwise} \end{cases}
\end{aligned} \tag{A.8}$$

Bibliography

AdWords (2013). Glossary: Basic AdWords Terms. `http://support.google.com/adwords/topic/15464?hl=en&ref_topic=1710534`.

Aggarwal, G., Feldman, J., Muthukrishnan, S., and Pál, M. (2008). Sponsored search auctions with Markovian users. *Internet and Network Economics*, pages 621–628.

Agichtein, E., Brill, E., Dumais, S., and Ragno, R. (2006). Learning user interaction models for predicting Web search result preferences. In *Proceedings of the* 29^{th} *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–10.

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

Amazon Mechanical Turk (2009). `http://www.mturk.com`.

Ashkan, A. and Clarke, C. L. A. (2009). Characterizing commercial intent. In *Proceedings of the* 18^{th} *ACM Conference on Information and Knowledge Management*, pages 78–87.

Ashkan, A. and Clarke, C. L. A. (2011). On the informativeness of cascade and intent-aware effectiveness measures. In *Proceedings of the* 20^{th} *International Conference on World Wide Web*, pages 407–416.

Ashkan, A. and Clarke, C. L. A. (2012). Modeling browsing behavior for click analysis in sponsored search. In *Proceedings of the* 21^{st} *ACM International Conference on Information and Knowledge Management*, pages 2015–2019.

Ashkan, A. and Clarke, C. L. A. (2013). Impact of query intent and search context on clickthrough behavior in sponsored search. *Knowledge and Information Systems Journal*, **34**(2), 425–452.

Ashkan, A., Clarke, C. L. A., Agichtein, E., and Guo, Q. (2008). Characterizing query intent from sponsored search clickthrough data. In *Proceedings of the SIGIR Workshop on Informational Retrieval for Advertising*, pages 15–22.

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

Ashkan, A., Clarke, C. L. A., Agichtein, E., and Guo, Q. (2009b). Estimating ad clickthrough rate through query intent analysis. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 222–229.

Aslam, J., Yilmaz, E., and Pavlu, V. (2005). The maximum entropy method for analyzing retrieval measures. In *Proceedings of the* 28^{th} *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34.

Baeza-Yates, R., Calderán-Benavides, L., and González-Caro, C. (2006). The intention behind Web queries. In *Proceedings of the* 13^{th} *International Symposium on String Processing and Information Retrieval*, pages 98–109.

Becker, H., Meek, C., and Chickering, D. (2007). Modeling contextual factors of click rates. In *Proceedings of the* 22^{nd} *AAAI Conference on Artificial Intelligence*, pages 1310–1315.

Beitzel, S., Jensen, E., Frieder, O., Grossman, D., Lewis, D., Chowdhury, A., and Kolcz, A. (2005). Automatic Web query classification using labeled and unlabeled training data. In *Proceedings of the* 28^{th} *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 581–582.

Belkin, N. (1993). Interaction with texts: Information retrieval as information-seeking behavior. *Information retrieval*, **93**, 55–66.

Benczúr, A., Bíró, I., Csalogány, K., and Sarlós, T. (2007). Web spam detection via commercial intent analysis. In *Proceedings of the* 3^{rd} *International Workshop on Adversarial Information Retrieval on the Web*, pages 89–92.

Bendersky, M., Gabrilovich, E., Josifovski, V., and Metzler, D. (2010). The anatomy of an ad: Structured indexing and retrieval for sponsored search. In *Proceedings of the* 19^{th} *International Conference on World Wide Web*, pages 101–110.

Beyond Search Data (2007). Beyond search semantic computing and internet economics program. `http://research.microsoft.com/en-us/um/redmond/about/collaboration/awards/beyondsearchawards.aspx`.

Bodoff, D. (2006). Relevance for browsing, relevance for searching. *Journal of the American Society for Information Science and Technology*, **57**(1), 69–86.

Brenes, D. J., Avello, D. G., and González, K. P. (2009). Survey and evaluation of query intent detection methods. In *Proceedings of the Workshop on Web Search Click Data*, pages 1–7.

Broder, A. (2002). A taxonomy of Web search. *SIGIR Forum*, **36**(2), 3–10.

Broder, A., Fontoura, M., Gabrilovich, E., Joshi, A., Josifovski, V., and Zhang, T. (2007). Robust classification of rare queries using Web knowledge. In *Proceedings of the* 30^{th} *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 231–238.

Büttcher, S., Clarke, C. L. A., and Cormack, G. (2010). *Information retrieval: Implementing and evaluating search engines*. The MIT Press.

Carmel, E., Crawford, S., and Chen, H. (1992). Browsing in hypertext: A cognitive study. *IEEE Transactions on Systems, Man and Cybernetics*, **22**(5), 865–884.

Carrasco, J., Joseph, J., Daniel, C., Fain, C., Lang, K., and Zhukov, L. (2003). Clustering of bipartite advertiser-keyword graph.

Carterette, B., Kanoulas, E., and Yilmaz, E. (2012). Incorporating variability in user behavior into systems based evaluation. In *Proceedings of the* 21^{st} *ACM International Conference on Information and Knowledge Management*, pages 135–144.

Chapelle, O. and Zhang, Y. (2009). A dynamic Bayesian network click model for Web search ranking. In *Proceedings of the* 18^{th} *World Wide Web Conference*, pages 1–10.

Chen, W., Ji, Z., Shen, S., and Yang, Q. (2011). A whole page click model to better interpret search engine click data. In *Proceedings of the* 21^{st} *Conference on Artificial Intelligence*.

Choo, C., Detlor, B., and Turnbull, D. (1998). A behavioral model of information seeking on the Web: preliminary results of a study of how managers and IT specialists use the Web. In *Proceedings of the* 61^{st} *Annual Meeting of the American Society for Information Science*, volume 35, pages 290–302.

Clarke, C. L. A., Agichtein, E., Dumais, S., and White, R. (2007). The influence of caption features on clickthrough patterns in Web search. In *Proceedings of the* 30^{th} *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 135–142.

Clarke, C. L. A., Kolla, M., Cormack, G., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of the* 31^{st} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 659–666.

Clarke, C. L. A., Craswell, N., Soboroff, I., and Ashkan, A. (2011). A comparative analysis of cascade measures for novelty and diversity. In *ACM International Conference on Web Search and Data Mining*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, **20**(1), 37.

Cover, T. and Thomas, J. (2006). *Elements of Information Theory*. Wiley-Interscience.

Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. (2008). An experimental comparison of click position-bias models. In *Proceedings of the First International Conference on Web search and Web Data Mining*, pages 87–94.

Dai, H., Zhao, L., Nie, Z., Wen, J., Wang, L., and Li, Y. (2006). Detecting online commercial intention (OCI). In *Proceedings of the* 15^{th} *International World Wide Web Conference*, pages 829–837.

Dai, N., Qi, X., and Davison, B. (2011). Bridging link and query intent to enhance Web search. In *Proceedings of the* 22^{nd} *ACM conference on Hypertext and Hypermedia*, pages 17–26.

Danaher, P. and Mullarkey, G. (2003). Factors affecting online advertising recall: A study of students. *Journal of Advertising Research*, **43**(03), 252–267.

Debmbsczynski, K., Kotlowski, W., and Weiss, D. (2008). Predicting ads clickthrough rate with decision rules. In *Proceedings of the WWW Workshop on Target and Ranking for Online Advertising*.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, **39**(1), 1–38.

Do, C. B. and Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature biotechnology*, **26**(8), 897–899.

Doan, A., Ramakrishnan, R., and Halevy, A. (2011). Crowdsourcing systems on the World–Wide Web. *Communications of the ACM*, **54**(4), 86–96.

Dumais, S. and Chen, H. (2000). Hierarchical classification of Web content. In *Proceedings of the* 23^{rd} *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 256–263.

Dupret, G. and Piwowarski, B. (2008). A user browsing model to predict search engine click data from past observations. In *Proceedings of the* 31^{st} *Annual International ACM*

SIGIR Conference on Research and development in Information Retrieval, pages 331–338.

Edelman, B., Ostrovsky, M., and Schwarz, M. (2007). Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords. *American Economic Review*, **97**(1), 242–259.

Eguchi, S. and Copas, J. (2006). Interpreting kullback–leibler divergence with the neyman–pearson lemma. *Journal of Multivariate Analysis*, **97**(9), 2034–2040.

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

Fan, T. and Chang, C. (2010). Sentiment-oriented contextual advertising. *Knowledge and Information Systems*, **23**(3), 321–344.

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

Fourney, A., Mann, R., and Terry, M. (2011). Characterizing the usability of interactive applications through query log analysis. In *Proceedings of the* 29^{th} *International Conference on Human Factors in Computing Systems*.

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

Ghosh, A. and Mahdian, M. (2008). Externalities in online advertising. In *Proceeding of the* 17^{th} *International Conference on World Wide Web*, pages 161–168.

Grant, M. and Boyd, S. (2009). CVX: Matlab software for disciplined convex programming. `http://stanford.edu/~boyd/cvx`.

Guo, F., Liu, C., Kannan, A., Minka, T., Taylor, M., Wang, Y., and Faloutsos, C. (2009a). Click chain model in Web search. In *Proceedings of* 18^{th} *World Wide Web Conference*, pages 11–20.

Guo, F., Liu, C., and Wang, Y. (2009b). Efficient multiple-click models in Web search. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 124–131. ACM.

Hanley, J. A., McNeil, B. J., *et al.* (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, **148**(3), 839–843.

Hillard, D., Schroedl, S., Manavoglu, E., Raghavan, H., and Leggetter, C. (2010). Improving ad relevance in sponsored search. In *Proceedings of the Third ACM International Conference on Web search and Data Mining*, pages 361–370.

Hotchkiss, G., Alston, S., and Edwards, G. (2005). Eye tracking study. *Research white paper, Enquiro Search Solutions Inc*.

Hu, B., Zhang, Y., Chen, W., Wang, G., and Yang, Q. (2011). Characterizing search intent diversity into click models. In *Proceedings of the* 20^{th} *World Wide Web Conference*, pages 17–26.

Hu, D., Yang, Q., and Li, Y. (2008). An algorithm for analyzing personalized online commercial intention. In *Proceedings of the* 2^{nd} *International Workshop on Data Mining and Audience Intelligence for Advertising*, pages 27–36.

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

Jansen, B. and Resnick, M. (2005). Examining searcher perceptions of and interactions with sponsored results. In *Workshop on Sponsored Search Auctions*.

Jansen, B. and Resnick, M. (2006). An examination of searcher's perceptions of nonsponsored and sponsored links during ecommerce Web searching. *Journal of the American Society for Information Science and Technology*, **57**(14), 1949–1961.

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

Jansen, B., Booth, D., and Spink, A. (2008). Determining the informational, navigational, and transactional intent of Web queries. *Information Processing and Management*, **44**(3), 1251–1266.

Joachims, T. (2008). SVMlight support vector machine. `http://svmlight.joachims.org`.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the* 28^{th} *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 154–161.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., and Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems*, **25**(2).

Kang, I. (2005). Transactional query identification in Web search. *Information Retrieval Technology*, pages 221–232.

Kang, I. and Kim, G. (2003). Query type classification for Web document retrieval. In *Proceedings of the* 6^{th} *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 64–71.

Kelly, D. and Teevan, J. (2003). Implicit feedback for inferring user preference: A bibliography. *ACM SIGIR Forum*, **37**(2), 18–28.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22**(1), 79–86.

Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

Lee, U., Liu, Z., and Cho, J. (2005). Automatic identification of user goals in Web search. In *Proceedings of the* 14^{th} *International World Wide Web Conference*, pages 391–400.

Li, X., Wang, Y., and Acero, A. (2008). Learning query intent from regularized click graphs. In *Proceedings of the* 31^{st} *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 339–346.

Li, Y., Krishnamurthy, R., Vaithyanathan, S., and Jagadish, H. V. (2006). Getting work done on the Web: supporting transactional queries. In *Proceedings of the* 29^{th} *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 557–564.

Liu, C., Guo, F., and Faloutsos, C. (2009). BBM: Bayesian browsing model from petabyte-scale data. In *Proceedings of the* 15^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 537–546.

Liu, Y., Zhang, M., Ru, L., and Ma, S. (2006). Automatic query type identification based on click through information. *Information Retrieval Technology*, pages 593–600.

Malik, H., Fradkin, D., and Moerchen, F. (2011). Single pass text classification by direct feature weighting. *Knowledge and Information Systems*, **28**(1), 79–98.

Morrison, J., Pirolli, P., and Card, S. (2001). A taxonomic analysis of what World Wide Web activities significantly impact people's decisions and actions. In *Proceedings of the CHI'01 Extended Abstracts on Human Factors in Computing Systems*, pages 163–164.

Navarro-Prieto, R., Scaife, M., and Rogers, Y. (1999). Cognitive strategies in Web searching. In *Proceedings of the* 5^{th} *Conference on Human Factors and the Web*, pages 19–34.

Nettleton, D., Calderán-Benavides, L., and Baeza-Yates, R. (2007). Analysis of Web search engine query session and clicked documents. In *Proceedings of the* 8^{th} *International Workshop on Knowledge Discovery on the Web*, pages 207–226.

Oard, D. and Kim, J. (2001). Modeling information content using observable behavior. In *Proceedings of the* 64^{th} *Annual Conference of the American Society for Information Science and Technology*, pages 481–488.

Oard, D., Kim, J., *et al.* (1998). Implicit feedback for recommender systems. In *Proceedings of the AAAI Workshop on Recommender Systems*, pages 81–83.

O'Day, V. and Jeffries, R. (1993). Orienteering in an information landscape: how information seekers get from here to there. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*, pages 438–445.

Papoulis, A., Pillai, S., and Unnikrishna, S. (2002). *Probability, random variables, and stochastic processes*. McGraw-Hill New York.

Piwowarski, B., Dupret, G., and Jones, R. (2009). Mining user Web search activity with layered Bayesian networks or how to capture a click in its context. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 162–171.

Poblete, B. and Baeza-Yates, R. (2008). Query-sets: using implicit feedback and query patterns to organize Web documents. In *Proceedings of the* 17^{th} *International World Wide Web Conference*, pages 41–50.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *speech recognition*, **77**(2), 257–286.

Radlinski, F., Bennett, P., Carterette, B., and Joachims, T. (2009). Redundancy, diversity and interdependent document relevance. *ACM SIGIR Forum*, **43**(2), 46–52.

Radlinski, F., Szummer, M., and Craswell, N. (2010). Inferring query intent from reformulations and clicks. In *Proceedings of the* 19^{th} *International Conference on World Wide Web*, pages 1171–1172.

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

Richardson, M. (2008). Learning about the world through long-term query logs. *ACM Transactions on the Web*, **2**(4), 1–27.

Richardson, M., Dominowska, E., and Ragno, R. (2007). Predicting clicks: estimating the clickthrough rate for new ads. In *Proceedings of the* 16^{th} *International World Wide Web Conference*, pages 521–530.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, **12**(1), 77.

Rose, D. and Levinson, D. (2004). Understanding user goals in Web search. In *Proceedings of the* 13^{th} *International World Wide Web Conference*, pages 13–19.

Sakai, T. (2009). On the robustness of information retrieval metrics to biased relevance assessments. *Information and Media Technologies*, **4**(2), 547–557.

Sculley, D., Malkin, R., Basu, S., and Bayardo, R. (2009). Predicting bounce rates in sponsored search advertisements. In *Proceedings of the* 15^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1325–1334.

Sellen, A., Murphy, R., and Shaw, K. (2002). How knowledge workers use the Web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 227–234.

Shen, D., Li, Y., Li, X., and Zhou, D. (2009). Product query classification. In *Proceedings of the* 18^{th} *ACM Conference on Information and Knowledge Management*, pages 741–750.

Shneiderman, B., Byrd, D., and Croft, W. (1997). Clarifying search: A user-interface framework for text searches. *D-Lib Magazine*, **3**(1), 18–20.

Sparck Jones, K., Walker, S., and Robertson, S. (2000). A probabilistic model of information retrieval: Development and comparative experiments (part 2). *Information Processing and Management*, **36**(6), 809–840.

Tan, B. and Peng, F. (2008). Unsupervised query segmentation using generative language models and Wikipedia. In *Proceedings of the* 17^{th} *International World Wide Web Conference*, pages 347–356.

Teevan, J., Alvarado, C., Ackerman, M., and Karger, D. (2004). The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 415–422.

Teevan, J., Dumais, S., and Liebling, D. (2008). To personalize or not to personalize: modeling queries with variation in user intent. In *Proceedings of the* 31^{st} *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 163–170.

Von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 319–326.

Welch, M. J., Cho, J., and Olston, C. (2011). Search result diversity for informational queries. In *Proceedings of the* 20^{th} *International Conference on World Wide Web*, pages 237–246.

White, R. W. and Drucker, S. M. (2007). Investigating behavioral variability in Web search. In *Proceedings of the* 16^{th} *International Conference on World Wide Web*, pages 21–30.

Xu, W., Manavoglu, E., and Cantu-Paz, E. (2010). Temporal click model for sponsored search. In *Proceedings of the* 33^{rd} *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 106–113.

Yahoo! Search Marketing (2013). Getting Started Guide: Glossary. `http://developer.searchmarketing.yahoo.com/docs/V7/gsg/glossary.php`.

Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y., and Chen, Z. (2009). How much can behavioral targeting help online advertising? In *Proceedings of the* 18^{th} *International Conference on World Wide Web*, pages 261–270.

Zhang, W. and Jones, R. (2007). Comparing click logs and editorial labels for training query rewriting. In *Proceedings of the WWW Workshop on Query Log Analysis: Social And Technological Challenges*.

Zhang, Y., Chen, W., Wang, D., and Yang, Q. (2011). User-click modeling for understanding and predicting search-behavior. In *Proceeding of the* 17^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1388–1396.

Zheng, Z., Zha, H., Zhang, T., Chapelle, O., Chen, K., and Sun, G. (2007). A general boosting method and its application to learning ranking functions for Web search. *Advances in Neural Information Processing Systems*, **20**, 1697–1704.

Zhu, Z., Chen, W., Minka, T., Zhu, C., and Chen, Z. (2010). 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.