Ready to Buy or Just Browsing? Detecting Web Searcher Goals from Interaction Data

Qi Guo
Emory University
qguo3@mathcs.emory.edu

Eugene Agichtein
Emory University
eugene@mathcs.emory.edu

ABSTRACT

An improved understanding of the relationship between search intent, result quality, and searcher behavior is crucial for improving the effectiveness of web search. While recent progress in user behavior mining has been largely focused on aggregate server-side click logs, we present a new class of search behavior models that also exploit fine-grained user interactions with the search results. We show that mining these interactions, such as mouse movements and scrolling, can enable more effective detection of the user's search goals. Potential applications include automatic search evaluation, improving search ranking, result presentation, and search advertising. We describe extensive experimental evaluation over both controlled user studies, and logs of interaction data collected from hundreds of real users. The results show that our method is more effective than the current state-of-the-art techniques, both for detection of searcher goals, and for an important practical application of predicting ad clicks for a given search session.

Categories and Subject Descriptors

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

General Terms

Design, Experimentation, Performance

Keywords

user intent inference; search behavior modeling; search advertising

1. INTRODUCTION

An improved understanding of searcher needs and interests is crucial for search engines to generate satisfactory search results, with applications ranging from search evaluation to improving search ranking, presentation, and usability. What makes the problem particularly daunting is that the same query may reflect different goals not only different users [29], but even for the same user at different times. For example, a user may search for "blackberry" initially to learn about the Blackberry smartphone; however, days or weeks later the same user may search for "blackberry" to identify the best deals on actually purchasing the device. Thus, identifying the most popular or majority meaning for a query is not sufficient; rather, the challenge is to identify the intent of the given *search*, contextualized within a search task (e.g., buying a smartphone, which may involve goals such as researching the device, comparing data plans, lookup of customer reviews, and eventual purchase).

While previous studies have shown the effectiveness of eye tracking to identify user interests (e.g., [9]), unfortunately, it requires expensive equipment, limiting the applicability. However, recent work has shown the existence of coordination between the searcher gaze position and mouse movement over the search results [19, 28]. Our hypothesis is that *searcher interactions such as mouse movement and scrolling can help more accurately infer searcher intent and interest in the search results*. That is, like eye movements, such interactions can reflect searcher attention. These interactions can be captured with Javascript code that could be returned as part of a Search Engine Result Page (SERP). This would allow estimating which parts of the SERP the user is interested in (e.g., whether the searcher is paying more attention to the organic or the sponsored results), and provide additional clues about the search intent.

To test this hypothesis, we develop a novel model of inferring searcher intent that incorporates both *search context* and *rich interactions* with the results, such as mouse movements, hovering, scrolling, and clicking on the results (Section 3). The model is operationalized by converting these interactions into features, which can then be used as input to classification algorithms to infer the search intent from the interaction data (Section 4).

While many other dimensions of search intent have been studied (e.g., [29]), in this paper we focus on characterizing two general types of commercial intent: *research* and *purchase*, illustrated in the examples above. We focus on these intent classes for two reasons: a) these are broad classes of informational queries, and b) distinguishing between the two has significant practical applications, in particular, for search advertising. For example, a searcher issuing a seemingly commercial query (e.g., "blackberry") may not be interested in the search ads if the organic (non-sponsored) search results are sufficient for their needs. In this case, showing ads could annoy the searcher, and contribute to "training" them to ignore the ads [8]. Thus, knowing the searcher intent (and consequently, interest in viewing sponsored results) would allow search engines to target ads better; and for advertisers to better target the appropriate population of "receptive" searchers. So, if we could infer a user's current interests based on her search context and behavior, a search engine may then show more or fewer ads (or none at all) if the current user is in the "research" mode.

The experiments in this paper follow a similar progression. First, we show that the proposed interaction model helps distinguish between known "research" and "purchase" search intents in a controlled used study (Section 5.2). Then, we show that when our

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

model is applied to the search data of real users, the searches predicted to have "purchase" intent indeed have significantly higher ad clickthrough rates than those predicted to have "research" intent (Section 5.3). Finally, we apply our model to the task of *predicting* ad clickthrough for an individual user within the *current* search session - which could have significant practical applications for commercial search engines (Section 6).

In summary, the contributions of this paper include:

- A richer model of searcher intent, that incorporates searcher interactions with session-level *state* for jointly modeling searcher goals and behavior.
- Empirical evidence that the predictions of the searcher's goals correlate with empirical ad clickthrough rates.
- A large-scale experimental evaluation of our model on predicting ad clickthrough, an important problem in its own right.

Next, we briefly survey the background and related work to put our contribution in context and to motivate our approach.

2. BACKGROUND AND MOTIVATION

Search intent detection, and more generally information seeking behavior, has been an active area of research in information science and information retrieval communities. We briefly summarize these efforts next. We then motivate our approach by considering two concrete kinds of commercial intent, "research" and "purchase" (Section 2.3). Finally, we describe a particularly important application of distinguishing between research and purchase intent to search advertising (Section 2.4).

2.1 Related Work

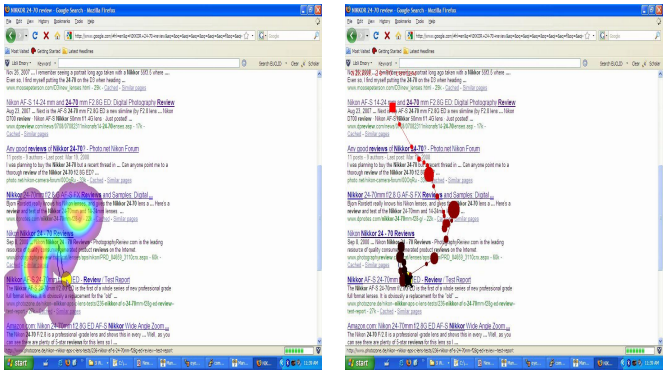
The origins of user modeling research can be traced to library and information science research of the 1980s. An excellent overview of the traditional "pre-Web" user modeling research is available in [5]. With the explosion of the popularity of the Web, and with increasing availability of large amounts of user data, Web usage mining has become an active area of research. In particular, inferring user intent in Web search has been studied extensively, including references [22, 1]. There is a broad consensus on the top 3 levels of intent taxonomy, namely the navigational, transactional and informational intents introduced by Broder[7]. Recently, a more specific query intent classification was presented in [29], where informational and resource (transactional) user goals are further divided into specific sub-goals.

Previous research on user behavior modeling for Web search focused on aggregated behavior of users to improve Web search or to study other general aspects of behavior [16]. Another approach is to model random walk on the click graph [13], which considers a series of interactions within a search page, but not session-level context or behavior information. Most of the previous research on predicting ad clickthrough focuses on learning from the content of displayed ads (e.g., [12, 27]), but did not take into account the session-level search context and the individual user behavior. Reference [4] considered the result (and ad) relative position and presentation features to improve clickthrough estimation, within a single page. Reference [15] addressed the detection of commercial intent in the aggregate using page-level context modeling. Another dimension of work somewhat similar in approach to ours considered query chains and browsing behavior to infer document relevance (e.g., [26]). Earlier references [11], [10] and [25], attempted to capture and identify user goals based on the query context. Most recently, a model to estimate searcher's viewing behavior based on observable click data was introduced in [32]. Our work expands on these efforts by exploiting additional evidence (namely, richer user interaction features) for both general intent classification and for applications such as ad click prediction.

Furthermore, it has been shown that specific user goals and experience vary widely and have substantial effect on user behavior [33]. Some queries have substantial variation in intent [31], and searcher behavior can help distinguish user intent in such ambiguous cases, as we attempt to do in this paper. Recently, eye tracking has started to emerge as a useful technology for understanding some of the mechanisms behind user behavior (e.g., [14, 21]). Our work expands on the observations described in [28] and operationalized in [19], which explored eye-mouse coordination patterns. Other previous work considered mouse movement in a different setting (e.g., windows desktop operations) for biometric identification based on interactions [2], and for website usability evaluation [3]. In contrast to previous work, we address different problems - that is, we aim to predict the searchers *current goal* (for research goal identification), or *future behavior*, e.g., whether the user is likely to click on an ad sometime during the current session.

2.2 General Search Intent Detection

While it has been shown previously that the most popular intent of a query can be detected for sufficiently frequent queries (e.g., [20, 23, 6]), out aim is to detect the intent of the specific *search* – that is, for queries that could plausibly be navigational or informational in intent. Previous work has shown preliminary indications that at a coarse level, the navigational vs. informational intent of a query can be distinguished for a given search by mining user's behavior [18]. However, informational intent combines many different types of search, varying from direct lookup to exploratory search (e.g., [34]). Furthermore, specific kinds of informational intent are more difficult to identify due to (typically) lower frequency of informational queries, and consequently smaller amounts of interaction data available. Finally, informational searchers tend to exhibit higher variation in intent [31], making informational search modeling a particularly challenging problem, as described next.



(a) (b)

Figure 1: Searcher gaze position and corresponding mouse trajectory (Research intent)

2.3 Deconstructing Informational Searches

Unlike navigational intent, informational query intent encapsulates many different types of important search tasks, for example *Directed* vs. *Undirected* (*Exploratory*) searches, as well as more specific *Lookup* or *Locate* goals, e.g., to verify that a particular product can be obtained. Many other dimensions of informational queries have been identified, including topical [29], exploratory vs. specific [29], commercial vs. non-commercial [15]. In particular,

we focus on one important intersection of informational and commercial intent categories, namely *Research* vs. *Purchase* intent.

As a concrete example, consider how users with research intent examine the search engine result page (SERP) for a query "nikkor 24-70 review". This query is commercial (the searcher is probably considering whether to buy this digital camera model), but could also be research-oriented (the searcher is interested in reviews, and not yet in making an immediate purchase). Figure 1 (a) shows the gaze position "heat map" (different colors represent amount of time spent examining the corresponding page position). Figure 1 (b) shows the mouse movements performed by the subject as they were examining the SERP. This example illustrates the possible connection between user interactions on the SERP and interest in the specific results. Thus, it is important to model not only the "popular" intent of a query, but also the searcher's immediate intent based on the context (within a search session) as well as on the interactions with the search results. In addition to research importance, this capability has important practical applications to search advertising, as described next.

2.4 Application: Search Advertising

An important practical application of our methods is predicting whether the user is likely to click on search ads shown next to the "organic" results. This problem is related to the research vs. purchase orientation of the user's goals: a user is more likely to click on a search ad if they are looking to make a purchase, and less likely if they are researching a product. We verify this observation empirically by comparing the ad clickthrough of "research" and "purchase" searches as classified by our model.

Furthermore, one could *predict* whether the searcher is more or less likely to click on an ad in *future* searches within the current session. This idea is illustrated in Figure 2, which shows an example where the user hovers the mouse over the ads before she clicks on an organic result in her first search for the query "green coffee maker". In her following search for the same query, in the same session, she clicks on an ad. We call this predisposition "advertising receptiveness", and show that the user's interest in a search ad shown for a *future* search within the same session can be predicted based on the user interactions with the *current* search result page.

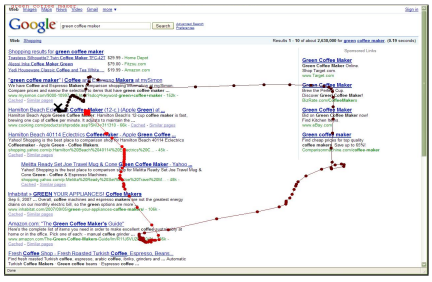


Figure 2: Mouse trajectory on a SERP for query "green coffee maker" with an ad click on the *next* search result page

If a search engine could be notified that a searcher is (or is not) interested in search advertising for their current task, the next results returned could be more accurately targeted towards this user. For example, if the user appears interested in buying a hybrid car, ads for hybrids as well as deals in the organic results should be returned. In contrast, if the user appears to be just *researching* hybrid technology, then the search engine should privilege customer reviews or technical articles. To achieve this real-time behavioral targeting, we argue that contextualized user interaction models are required.

3. SEARCH AND USER MODEL

In this section we first describe the definitions of search tasks and search goals. We then introduce our approach to mine the contextualized fine-grained interactions.

3.1 Search Model: Tasks and Goals

Our work assumes a simplified model of search following recent information behavior literature, where a user is attempting to accomplish an overall search task by solving specific search goals, as illustrated in Figure 3.

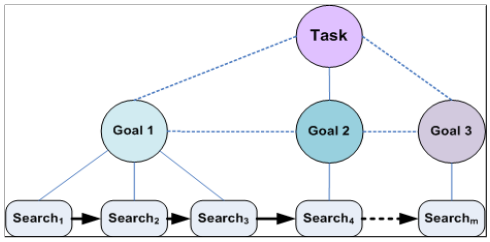


Figure 3: Relationship between a search task, immediate goals and specific searches to accomplish each goal.

Many user information needs require multiple searches until the needed information is found. Thus, it is natural to organize individual queries into overall tasks and immediate goals, which correspond to a common information need. For this, we use the idea of a search *task*, which, in turn, requires more *immediate goals* to accomplish by submitting and examining related *searches*. Our operational definition of a search task is that it consists of a consecutive sequence of queries that share at least one non-stopword term with any previous query within the task. A example search session consisting of two search tasks is reported in Figure 4. We verified this simple definition of a search task manually, and out of more than 100 tasks examined, in all but 3 tasks the searches shared at least one non-stopword term with some other search in the task. In our dataset, while the 30-minute sessions tend to be 6.77 searches long on average, tasks tend to contain 2.71 searches on average, which is consistent with previous finding [33] that users perform on average only two or three query reformulations before giving up.

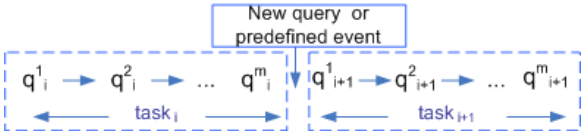


Figure 4: An example user session, consisting of two consecutive disjoint search tasks.

3.2 User Model: Goal-driven Search

Our user model naturally follows our search model. A user, while solving a search task, has a number of *immediate goals*. While these goals are "hidden" - that is, not directly observable, the user *searches* (queries) and their *interactions* on the corresponding search results can be observed. Thus, we model a user as a non-deterministic state machine with *hidden states* representing user goals, and *observable actions* that depend on the user's current state. Our model is illustrated in Figure 5: searcher actions such as queries, result clicks, and mouse movements are observations generated by the hidden states corresponding to the user's goals. We restrict the interactions to those on the SERP to make our work more realistic: search engines are able to capture user interactions

over their own results, but capturing actions on other pages require significant additional effort.

For example, if the immediate user goal is informational, then longer mouse trajectories are more likely to be observed on search result pages (as the user is more likely to examine the results to decide which one is most relevant); in contrast, if the immediate user goal is navigational, he or she can quickly recognize the target site, resulting in shorter mouse trajectory and faster response time. Similarly, ad clicks are more likely to be emitted if the user is in a receptive state to search advertising (e.g., has a Purchase goal), and less likely if the user is in a non-receptive state (e.g., has a Research goal). Hence, observations including the search context and user interactions are related to the states (goals) of the users. If we can infer the hidden states using the observations, we can both recover the user's immediate search goal and potentially the overall task, as well as *predict future user actions* such as ad clicks, in *subsequent* searches.

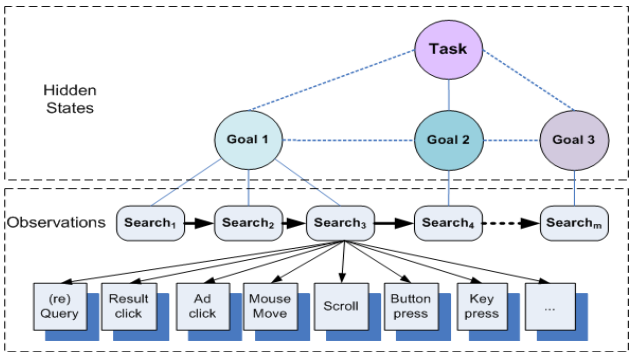


Figure 5: Sample states and observations for a single search within a task.

Note that the predictions in our model are dependent on both the current and the *previous* goal state of the user, thus naturally allowing us to maintain the user's "mental state" across individual searches. Furthermore, this formalism allows for arbitrary number of hidden states which may correspond to different types of goals and more complex tasks. For example, this would allow us to model different variations of ad receptiveness. These sequential user models can (and have been) implemented in previous work (see Section 2.1). However, what makes our model unique is the rich representation of user actions *on* the search result pages, allowing us to potentially capture the mental state of the user while he or she is examining the search results, as we describe next.

4. INFRASTRUCTURE, FEATURES AND ALGORITHMS

We now describe the actual implementation of our system. First, we describe the infrastructure for extracting and storing user interactions (Section 4.1); then, we describe the concrete representation of these interactions as features (Section 4.2) for incorporating the interaction information into classification algorithms (Section 4.3).

4.1 Infrastructure

The user interaction data was captured by instrumenting the LibX toolbar [1]. The toolbar was modified to buffer the GUI events such as mouse movements, scrolling, and key press events, and send them to our server for logging. These instrumented Web browsers were installed on approximately 150 public-use computers (mostly Windows PCs) at the Emory University library. The usage was tracked only for users who have explicitly opted in to participate in our study. No identifiable user information was stored.

Note that the instrumentation described above for the search result pages does not necessarily require a user download or installation: JavaScript code similar to what we run in the toolbar can be easily returned in the header of a Search Engine Result Page (SERP). Also, we are only modeling searcher behavior on a SERP, and do *not* consider the external result pages visited by clicking on the results; therefore, all the data we collected would be available to the search engine via light-weight server-side instrumentation.

In our prototype implementation we sample mouse movements and scroll events at every 5 pixels moved, or every 50 ms, whichever is more frequent, and keep all other events (e.g., MouseDown, KeyPress events) without downsampling.

4.2 Features

We now describe the types of interactions captured and the corresponding feature representations. The major feature groups and representative features are summarized in Table 1. Each feature group is described below to the extent permitted by space limitations. Complete feature specification and the dataset from our user study experiments described in Section 5.1 are available online[2]. The features used to represent each search are illustrated in Figure 5, with some features spanning multiple searches to provide session-level context, as described below.

Query group: This group of features is designed to capture the same information as was used in the previous studies of capturing clicks in the context of previous and subsequent queries [25, 26]. Specifically, we include the tokens from the text of the query (after frequency thresholding); the length of the query in characters and words, and binary features such as *IncludesTLD*, which is 1 if the query includes a TLD token such as ".com" or ".org".

SERPContent group: These features represent the text (and markup) content of the SERP. Specifically, both the organic results and the sponsored results are allocated a separate feature space, and include the tokens from *OrganicText* and *AdText*, respectively (after frequency filtering). Additionally, tokens and markup is extracted from the whole SERP, regardless of the type of result (organic or sponsored) and represented as the *SERPText* features (after frequency filtering).

ResultQuality group: These features aim to capture coarse information about the SERP relation to the query, namely how many words in the organic result summaries match the query terms (*SnippetOverlap*); how many words in the text of the ads match the query terms (*AdOverlap*); as well as the normalized versions of these features computed by dividing by the query length, in words. We also capture the number of ads, number of ads at the top of the SERP (*NorthAds*), and number of ads on the side (*EastAds*). These features has been shown in previous work to correlate with the degree of commercial interest in the query.

Interaction group: The interaction features aim to capture the client-side interaction events. Specifically, the features include: the number of SERP GUI events, such as number of mouse events (*TotalMouse*), scroll events (*TotalScroll*) and keypress events (*TotalKeypress*); time features, such as SERP deliberation time, measured as seconds until first GUI event (*DeliberationTime*), the time until the first result click (*SERPDwellTime*); and hovering features, that

[1]Original LibX toolbar available at `www.libx.org`

[2]`http://ir.mathcs.emory.edu/intent/sigir2010/`.

Feature group	*Count*	*Description*
Query	4	QueryTokens* (unigram), QueryLengthChars, QueryLengthWord, IncludesTLD (1 if contains ".com", ".edu").
SERP Content	3	AdText* (unigram), OrganicText* (unigram), SERPText* (unigram). Each feature contains 100 most frequent terms from each area of the SERP (e.g., 100 most frequent tokens in the ads).
Result Quality	7	TotalAds, NorthAds, EastAds, SnippetOverlap, SnippetOverlapNorm, AdOverlap, AdOverlapNorm
Interaction	99	MouseRange, MouseCoordinates, MouseSpeed, MouseAcceleration, TotalMouse, TotalScroll, TotalKeypress, SERPDwellTime, DeliberationTime, HoverEastAd, HoverNorthAd, HoverOrganic, etc (see main text)
Click	7	ClickUrl* (unigram), NumBrowseAfterClick, AverageDwellTime, TotalDwellTime, SAT, DSAT, ClickType
Context	7	IsInitialQ, IsSameQ, IsReformulatedQ, IsExpansionQ, IsContractedQ, RepeatQ, SERPIndex
All	127	All features and feature classes used for experiments

Table 1: Summary of the features used for representing searcher context and interactions. The full feature specification and sample data are available at `http://ir.mathcs.emory.edu/intent/sigir2010/`**.**

.

measure how the time that the mouse hovers over an area of interest such as north ads, east ads, and organic results regions.

Our interaction features also aims to capture the physiological characteristics hidden in mouse movements, following reference [24]. In particular, the mouse trajectory representation is split into two subgroups, (*Mouse (Global)*) and (*Mouse (Local)*):

- Mouse (Global): the features include the length, vertical and horizontal ranges of mouse trajectory, in pixels; also, the features describing the general statistics of the trajectory, namely, the means and the standard deviations of the mouse coordinates, the difference in distance and time between two adjacent mouse points, the velocity, acceleration, slope and rotation (computed from the difference of slopes between neighboring points).
- Mouse (Segment): to distinguish the patterns in different stages of the user interactions with the search results, we split each mouse trajectory into five *segments*: initial, early, middle, late, and end. Each of the five segments contains 20% of the sample points of the trajectories. We then compute the same properties (e.g., speed, acceleration, slope etc.) as above, but computed for each segment individually. The intuition is to capture mouse movement during 5 different stages of SERP examination (e.g., first two segments correspond to the visual search stage, and last segment corresponds to moving the mouse to click on a result).

Click group: Captures the types and properties of result clicks and SERP revisits. Specifically, the features include tokens in the clicked URL (ClickUrl); the number of URLs visited after a result click (NumBrowseAfterClick), the average and total dwell time on each visited result URL. We also identify satisfied URL visits (those with dwell time greater than 30 seconds [17]), and "dissatisfied" visits (those with dwell time less than 15 seconds [30]), as well as the number and the position of the satisfied and dissatisfied URL visits within the task. Finally, we capture the type of the click, such as a click on organic result, on a menu item, or on a search ad (ClickType).

Context group: Captures where the search belongs to within a task context, the features include: whether the query is initial in session (*IsInitialQ*), whether the query is identical to previous query (*IsSameQ*), whether the query overlap with previous query submitted; respectively true if a word is replaced (*IsReformulatedQ*), or added(*IsExpansionQ*), or removed (*IsContractedQ*); whether the query was issued within same session (*RepeatQ*); the current position (progress) within a search session, e.g., whether this was a first, second, or 5th search in the session (*SERPIndex*). In summary, the features attempt to capture properties of the query, the search context, and the interaction on the SERP page itself.

4.3 Classifier Implementation

We now describe the details of classifier implementations we considered. We experiment with two different families of classifiers: Support Vector Machine (SVM) that supports flexible feature representation, and Conditional Random Fields (CRF), which naturally supports modeling sequences.

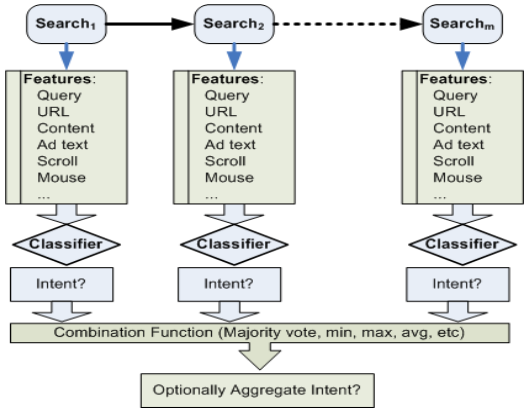


Figure 6: Our SVM classifier configuration.

Support Vector Machine (SVM): We chose the Weka implementation of the SVM implementation with the SMO optimization algorithm, using polynomial kernel with degree 4 (chosen during preliminary development experiments).

Conditional Random Field (CRF): The SVM representation allows only limited representation of the search *state*: that is, whether the searcher is still in "exploratory" stage of the search or is now on the next goal of verifying specific information found during exploration "lookup" stage. To explicitly model these different states (goals) within a session, CRF allows us to define a conditional probability over hidden state sequences given a particular observation sequence of searches. Take predicting future ad clicks as an example: at training, the hidden state is assigned according to whether an ad click was observed in the future searches within the task. Note that an ad click on *current* SERP is simply an observation, and is not necessarily an indication whether a user remains likely to click on future search ad in the same task. At test time, we identify the intent sequence that maximizes the conditional probability of the observation sequence. In the current system we used the Mallet[3] implementation.

[3]Available at `http://mallet.cs.umass.edu/`.

5. EXPERIMENTAL RESULTS: BUYING OR BROWSING?

This section presents the experimental setup and results for our case study of informational intent detection, namely *research* vs. *purchase* intent. The goal of the experiments is to validate whether interaction data can help more accurately distinguish searcher intent classes. We first describe the data gathering and construction of the "ground truth" data and our evaluation metrics. Then, we report and analyze the experimental results (Section 5.2) obtained in a small user study. Finally, we verify our results over a large log of real user interaction data.

5.1 Search Task and Data Collection

Task: the problem is to detect, given a user's behavior on a SERP, whether the query had *research* or *purchase* intent.

Data collection: We performed a user study with 10 subjects, who were graduate and undergraduate students and university staff, that is, were technically savvy and had some experience with Web search. The subjects were asked to perform two search sessions. Each subject was asked first to research a product of interest to them for potential future purchase. Then, the same subject was asked to attempt to "buy" an item of immediate interest to the subject, which may or may not be the same item the subject was researching in the previous stage. The subjects were not restricted on time, and could submit any queries (usually, to the Google search engine) and click on any results.

All the interactions were tracked using our Firefox plugin. At the same time, the searcher gaze position was tracked using the EyeTech TM3 integrated eye tracker at approximately 30Hz sampling rate, for subsequent analysis. Additionally, each search and corresponding SERP interactions were labeled as parts of a *research* or *purchase* session, according to the explicitly stated intent of the corresponding session.

Methods Compared:

- **Baseline**: always guesses the majority class (Research).
- **SVM(Query)**: similar to the state-of-the-art models using query features (e.g., [26, 25]), implemented using Query group features described in Section 4.2, and trained using the SVM model.
- **SVM(All)**: the SVM classifier implemented using the features described in Section 4.2 to infer the user goal for each search (independently of other searches in the session).

5.2 Results

We now report the results for classifying whether the searcher had research or purchase intent. We split the data by time, using the first 90% of searches for each subject's data for training, and the rest for testing (recall, that each subject had two sessions, one research, and one purchase). In this experiment, the intent of each search is predicted independently of other searches in the session (we will use a full session-level model in Section 6). To evaluate classification performance, we use the standard Precision, Recall and Macro-averaged F1. Table 2 shows that our system, SVM(All), outperforms both baselines, resulting in accuracy of almost 97%.

To identify the most important features contributing to the classification, we performed feature ablation by removing one feature group at a time from the classifier(Table 3). All the feature groups provide significant contributions, but the most important features appear to be SERPContent and Interaction features: with these features removed, accuracy degrades to 86.7% from 96.7% with these features included. This makes sense since the SERP content can help enrich the context of a query, while the Interaction features provide additional clues about the searcher interest. However, since this user study was done over a rather small number of subjects, further investigation and additional user study is needed to fully understand the connection between various feature groups. To complement these results, we validate our model on an objective ad clickthrough metric on a much larger user population, as described next.

Method	*Acc.*	*Research*		*Purchase*		*F1*
		Prec.	*Recall*	*Prec.*	*Recall*	
Baseline	56.7	56.7	100	0	0	36.2
SVM(Query)	86.7	93.3	82.4	80.0	92.3	86.6
SVM(All)	**96.7**	**100**	**94.1**	**92.9**	**100**	**96.6**

Table 2: Classification performance for research vs. purchase.

Method	*Acc.*	*Research*		*Purchase*		*F1*
		Prec.	*Rec.*	*Prec.*	*Rec.*	
SVM(All)	**96.7**	**100**	**94.1**	**92.9**	**100**	**96.6**
SVM(-Query)	93.3	94.1	**94.1**	92.3	92.3	93.2
SVM(-SERPContent)	86.7	93.3	82.4	80.0	92.3	86.6
SVM(-ResultQuality)	93.3	**100**	88.2	86.7	**100**	93.3
SVM(-Click)	90.0	93.8	88.2	85.7	92.3	89.9
SVM(-Interaction)	86.7	**100**	76.5	76.5	**100**	86.7
SVM(-Context)	93.3	**100**	88.2	86.7	**100**	93.3

Table 3: Feature ablation results for intent classification.

5.3 Ad Clickthrough on Real Search Data

To better understand the effectiveness of our classifier, we evaluated our model on a large dataset of real user searches collected in the Emory University libraries using the infrastructure described in Section 4.1. We hypothesize that for *research* searches, clickthrough on search ads should be lower than for *purchase* searches. Therefore, we can evaluate the effectiveness of our intent classification model by comparing the ad clickthrough on the searches classified as *research* by our model, to those classified as *purchase*. To avoid "cheating", no click group or result URL features were used, as they could provide information to the classifier about the ad click on the SERP.

Data: The data was gathered from mid-August through mid-December 2008. To ensure data consistency, we generated a longitudinal dataset of the usage for 440 opted-in users, who clicked a search ad at least once during this period. For this universe of users we include all the search sessions attempted during this period. The resulting dataset contains 4,377 login sessions, comprising 6,476 search sessions, 16,693 search tasks and 45,212 searches.

Results: The predicted *purchase* searches have substantially higher ad clickthrough rates (9.7%) compared to *research* searches (4.1%), and all searches with at least one ad displayed (5.9%). These statistics are summarized in Table 4. As hypothesized, our *research* and *purchase* predictions indeed correlate with ad clickthrough of real users. What makes this result remarkable is that our model was trained on a small dataset compiled from just 10 subjects in the user study (with clear intent labels), yet still provides promising performance on unconstrained user data obtained "in the wild".

Search class	*#ACLK (%)*	*#SERP with Ads*	*Ad CTR (%)*
All	854	14545	5.9
Research	417	10054	4.1 (-29%))
Purchase	437	4491	9.7 (+66%)

Table 4: Search ad clickthrough statistics on all search pages (All), and for searches classified as "Research" and "Purchase".

6. PREDICTING AD CLICKTHROUGH

We now turn to the practical application of predicting future search ad clicks for the *current* user session. We first define the problem more formally, then describe the data and metrics (Section 6.2) used for this task, the methods compared (Sections 6.3 and 6.4), followed by the empirical results and discussion (Section 6.5, which concludes this section.

6.1 Problem statement

This problem of predicting future ad clickthrough for the *current user* is distinct from predicting ad clickthrough in aggregate for many users. We define our problem as follows: *Given* the first i searches in a search task $S(s_1, ..., s_i, ..., s_m)$, and the searcher behavior on these first i SERPs, *predict* whether the searcher will click on an ad on the SERP within the current search task S, for any of the future searches $s_{i+1}, s_{i+2}, ..., s_m$.

6.2 Data and Evaluation Metrics

For this task, the dataset was based on the interaction data collected from the opted-in users in the Emory Libraries, and consists of the same log data as described in Section 5.3.

Evaluation Metrics: To focus on the ad click prediction, we report the results for the positive class, i.e., the "advertising-receptive" state. Specifically, we report Precision (P), Recall (R), and F1-measure (F1) calculated as follows:

- **Precision (P)**: Precision is computed with respect to the positive (receptive) class, as fraction of true positives over all predicted positives. Specifically, for each search task, the precision is the fraction of correct positive predictions over all positive predictions for the task, averaged across all the search tasks.
- **Recall (R)**: for each task, the recall is computed as the fraction of correct positive predictions over all positive labels in the task. This value is then averaged over all the tasks.
- **F1-measure (F1)**: F1 measure, computed as $\frac{2P \cdot R}{P+R}$.

6.3 Methods Compared

- **CRF(Query)**: CRF model, implemented using the Query group features as described in Section 4.2.
- **CRF(Query+Click)**: CRF model, implemented using Query group and Click group features as described in Section 4.2.
- **CRF(All)**: CRF model, implemented using all features as described in Section 4.2.
- **CRF(All-Interaction)**: Same as above, but with the Interaction group features removed.

6.4 Classifier Configuration

We configured the CRF to have two hidden states, *A+* and *A-*, corresponding to "Receptive" (meaning that an ad click is expected in a future search within the current session), and "Not receptive" (meaning to not expect any future ad clicks within the current session).

6.5 Results and Discussion

To simulate an operational environment, we split the data by time, and use the first 80% of the sessions for training the system, and the remaining 20% of the sessions for test. The results on the test set are reported in Table 5. As we can see, our system achieves the highest performance on all metrics, compared to the baselines. Specifically, CRF(Query+Click) outperforms CRF(Query) on the ad receptiveness prediction task by incorporating the click information, and the CRF(All) system further increases both precision and recall by incorporating additional behavior features. Interestingly, removing the Interaction group of features from the full system (CRF(All-Interaction)) degrades the recall and overall F1 performance of the system, while precision is somewhat improved. This suggests that interaction features help detect additional cases (compared to query and click information alone) where a searcher may be interested in the ads, while occasionally introducing additional false positives. We discuss the performance of the system in more detail, next.

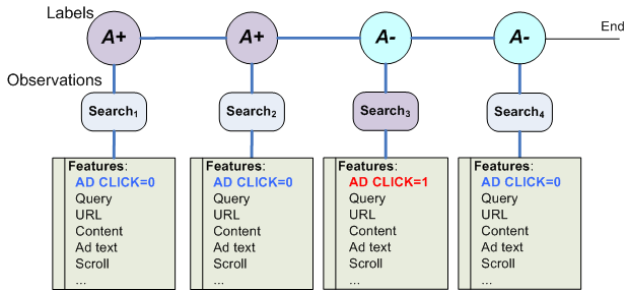


Figure 7: CRF configuration Model with two hidden states, A+ (receptive), and A- (non-receptive), with labels assigned according to the observed *future* ad clickthrough - here on the third search result pages within the session.

Method	*Precision*	*Recall*	*F1*
CRF(Query)	0.05 (-)	0.11 (-)	0.07 (-)
CRF(Query+Click)	0.14 (+153%)	0.12 (+11%)	0.13 (+77%)
CRF(All)	0.15 (+170%)	**0.21 (+99%)**	**0.17 (+141%)**
CRF(All-Interaction)	**0.16 (+206%)**	0.14 (+32%)	0.15 (+112%)

Table 5: Precision, Recall, and F1 for predicting ad receptiveness within a search task

Potential limitations: While our ad click prediction experiments were performed over a relatively large dataset collected over thousands of real search sessions for hundreds of users, we acknowledge some limitations of our study. Specifically, our user population is relatively homogeneous (college and graduate students, and faculty and staff), and substantially more training data may be required to achieve this performance for the general population. Another limitation is lack of conversion data: ad clickthrough is just one evaluation metric, and may not be predictive of the ultimate intent of the searcher (e.g., a searcher may click on an ad out of curiosity). Despite the limitations above, our population is large enough that useful conclusions could be drawn. To better understand the system performance and guide follow-up research, we describe representative case studies to provide better understanding of our system's performance:

Not using a mouse as a reading aid: this is the most frequent source of error introduced by the interaction features: when a mouse is not used to mark or focus user interest, interaction information could be misleading. One possible approach is to classify users into different groups according to mouse usage patterns, and train separate prediction models for each group.

Long difficult research sessions with ad clicks: in such cases, the searcher began with a research intent, and in her first several searches, no interest in ads were shown. However, as the session progresses, the user eventually clicks on an ad as the promising organic results are exhausted. For example, one searcher submitted

a query "comcast basic cable channels" in her task to find Comcast's basic cable line-up, and finally clicked on an ad because of the unsatisfactory organic results. Such ad clicks appear to be different from cases where a user clicks on ads because of a premeditated purchasing intent. We plan to investigate the different types of clicks in our future work.

Commercial purchase sessions without ad clicks: in such cases, a searcher examined the ads but did not click on any. This could be due to poor quality of search ads, or to availability of more promising organic search results. For example, one searcher submitted a query "george boots" and clicked on a Google's Product Search result. In this case, the searcher might be actually receptive to search advertising. However, we label such sessions as "non-receptive" since there's no future ad click to use as evidence. One natural extension of our model is to expand our labels by considering clicks on product search results to be similar to ad clicks with respect to purchasing intent. Another possibility may be that particular users could be generally "less-receptive" to advertising. To tackle this problem, personalizing our user models is a promising direction for future work.

7. CONCLUSIONS

In this paper we introduced a rich searcher behavior model that captures not only the queries and clicks, but also the fine-grained interactions with the search results, contextualized within a search session. Our experimental results on three related tasks demonstrate the generality and flexibility of our approach. The first task, predicting *research* vs. *purchase* goal of the user, provided insights about the feature groups most important for distinguishing these variants of commercial intent. In the second task, we validated our model by showing correlation of the predicted search intents with the ad clickthrough rates of real users. Finally, we demonstrated the performance of our model for an important practical application of predicting *future* search ad clickthrough within the *current* search session of each user.

In the future, we plan to expand our model to consider user interactions and page context other than the search result pages. In particular, we plan to incorporate the interactions on the intermediate result pages visited between successive searches, which may provide additional contextual information. We also plan to incorporate the user's history to enable the personalization of intent inference and search behavior prediction.

Acknowledgments

The authors thank Microsoft Research and Yahoo! Research for partially supporting this work through faculty research grants.

8. REFERENCES

[1] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proc. of SIGIR*, 2006.

[2] A. Ahmed and I. Traore. Detecting computer intrusions using behavioral biometrics. In *Proc. of PST*, 2005.

[3] R. Atterer, M. Wnuk, and A. Schmidt. Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. In *Proc. of WWW*, 2006.

[4] H. Becker, C. Meek, and D. Chickering. Modeling Contextual Factors of Click Rates. In *Proc. of AAAI*, 2007.

[5] N. J. Belkin. User modeling in information retrieval. *Tutorial at UM97*, 1997.

[6] D. J. Brenes, D. Gayo-Avello, and K. Pérez-González. Survey and evaluation of query intent detection methods. In *Proc. of WSCD workshop*, pages 1–7, 2009.

[7] A. Broder. A taxonomy of web search. *SIGIR Forum*, 2002.

[8] A. Z. Broder, M. Ciaramita, M. Fontoura, E. Gabrilovich, V. Josifovski, D. Metzler, V. Murdock, and V. Plachouras. To swing or not to swing: Learning when (not) to advertise. In *Proc. of CIKM*, 2008.

[9] G. Buscher, A. Dengel, and L. van Elst. Query expansion using gaze-based feedback on the subdocument level. In *Proc. of SIGIR*, 2008.

[10] H. Cao, D. H. Hu, D. Shen, D. Jiang, J.-T. Sun, E. Chen, and Q. Yang. Context-aware query classification. In *Proc. of SIGIR*, 2009.

[11] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li. Context-aware query suggestion by mining click-through and session data. In *Proc. of KDD*, 2008.

[12] M. Ciaramita, V. Murdock, and V. Plachouras. Online learning from click data for sponsored search. In *Proc. of WWW*, 2008.

[13] N. Craswell and M. Szummer. Random walks on the click graph. In *Proc. SIGIR*, 2007.

[14] E. Cutrell and Z. Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *Proc. of CHI*, 2007.

[15] H. K. Dai, L. Zhao, Z. Nie, J.-R. Wen, L. Wang, and Y. Li. Detecting online commercial intention (oci). In *Proc. of WWW*, 2006.

[16] D. Downey, S. T. Dumais, and E. Horvitz. Models of searching and browsing: Languages, studies, and application. In *Proc. of IJCAI*, 2007.

[17] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23(2), 2005.

[18] Q. Guo and E. Agichtein. Exploring mouse movements for inferring query intent. In *Proc. SIGIR*, 2008.

[19] Q. Guo and E. Agichtein. Towards predicting web searcher gaze position from mouse movements. In *Proc. CHI Extended Abstracts*, 2010.

[20] B. J. Jansen, D. L. Booth, and A. Spink. Determining the user intent of web search engine queries. In *Proc. of WWW*, 2007.

[21] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.*, 25(2), 2007.

[22] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *Proc. of WWW*, 2005.

[23] X. Li, Y.-Y. Wang, and A. Acero. Learning query intent from regularized click graphs. In *SIGIR*, pages 339–346, 2008.

[24] J. G. Phillips and T. J. Triggs. Characteristics of cursor trajectories controlled by the computer mouse. *Ergonomics*, 2001.

[25] B. Piwowarski, G. Dupret, and R. Jones. Mining user web search activity with layered bayesian networks or how to capture a click in its context. In *Proc. of WSDM*, 2009.

[26] F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. In *Proc. of KDD*, 2005.

[27] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proc. of WWW*, 2007.

[28] K. Rodden, X. Fu, A. Aula, and I. Spiro. Eye-mouse coordination patterns on web search results pages. In *Proc. of CHI*, 2008.

[29] D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proc. of WWW*, 2004.

[30] D. Sculley, R. G. Malkin, S. Basu, and R. J. Bayardo. Predicting bounce rates in sponsored search advertisements. In *Proc. of KDD*, 2009.

[31] J. Teevan, S. T. Dumais, and D. J. Liebling. To personalize or not to personalize: modeling queries with variation in user intent. In *Proc. of SIGIR*, 2008.

[32] K. Wang, N. Gloy, and X. Li. Inferring search behaviors using partially observable Markov (POM) model. In *Proc. of WSDM*, 2010.

[33] R. W. White and S. M. Drucker. Investigating behavioral variability in web search. In *Proc. of WWW*, 2007.

[34] R. W. White and R.A. Roth. Exploratory Search: Beyond the Query-Response Paradigm. Morgan & Claypool Synthesis Lectures on Information Concepts, Retrieval, and Services, 2009.