Search, Interrupted: Understanding and Predicting Search Task Continuation

Eugene Agichtein*
Emory University
Atlanta, GA, USA
eugene@mathcs.emory.edu

Ryen W. White, Susan T. Dumais, and Paul N. Bennett
Microsoft Research
Redmond, WA, USA
{ryenw, sdumais, pauben}@microsoft.com

ABSTRACT

Many important search tasks require multiple search sessions to complete. Tasks such as travel planning, large purchases, or job searches can span hours, days, or even weeks. Inevitably, life interferes, requiring the searcher either to recover the "state" of the search manually (most common), or plan for interruption in advance (unlikely). The goal of this work is to better *understand*, *characterize*, and *automatically detect* search tasks that will be continued in the near future. To this end, we analyze a query log from the Bing Web search engine to identify the types of *intents*, *topics*, and *search behavior* patterns associated with long-running tasks that are likely to be continued. Using our insights, we develop an effective prediction algorithm that significantly outperforms both the previous state-of-the-art method, and even the ability of human judges, to predict future task continuation. Potential applications of our techniques would allow a search engine to pre-emptively "save state" for a searcher (e.g., by caching search results), perform more targeted personalization, and otherwise better support the searcher experience for interrupted search tasks.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process; selection process*

Keywords

Search session analysis; Search behavior; Personalization.

1. INTRODUCTION

As Web search becomes increasingly important for planning and decision making, the complexity and scope of search tasks performed on search engines is increasing. Search engines are now often used for tasks such as travel planning, job hunting, or real estate searching. However, these tasks require significantly more effort and time to complete [10][21][24][25], potentially spanning days, weeks, or even months. While existing commercial Web search engines such as Bing and Google now provide tools to help users maintain and manage their search histories, the support they provide is not sufficient and the tools are not specifically designed to allow searchers to resume tasks that may been interrupted.

A challenge for search engines is to detect when a searcher is performing a long-running search task and *predict* whether they will continue it in the future. To this end, we analyze a query log from Bing to understand the types of *intents*, *motivations*, *topics*, and *search behaviors* associated with long-running tasks that are likely to be continued. Specifically, we try to understand search task continuation by analyzing tasks that were and were not continued by over a thousand Web searchers.

For example, consider the task of planning a wedding. The searcher might begin by checking recommended venues and their availabilities. However, at that point the task could be interrupted, as it requires checking dates and venues with the immediate family. When the task is continued the next day, the searcher has to restart from the beginning, unless the user planned for this event, and manually saved the most promising intermediate results. Indeed, there has been previous work on system support that lets users explicitly record promising content [10][27]. However, a perfect search engine could save the user the trouble if it could reliably detect that a suspended search session is likely to be continued at a later time.

While previous studies have considered long running tasks spanning multiple sessions (e.g., [10][21][24][25]), we dive deeper into the problem of task continuation to analyze the intent, motivation, and topics of these tasks. The more extensive analysis we perform allows for a fuller understanding of which tasks are most commonly resumed, in turn resulting in more accurate task continuation prediction. Potential applications include pre-emptively "saving state" for a searcher (e.g., by caching search results), more targeted personalization, and otherwise better supporting the searcher experience for long-running searches.

More formally, our problem is *predicting task continuation*:

> *Given* an active search task that has been suspended, *predict* whether the searcher will continue the task in the near future (e.g., within the next five days).

This problem is challenging, since it requires a search engine to make predictions about the kinds of tasks that tend to be continued, which intuitively would require substantial knowledge about the world. Yet, this work presents techniques to make these predictions automatically as well as, and often better than, experienced human annotators. Our contributions are threefold:

- A large-scale characterization of the intents, motivations, and topics associated with long-running search tasks (Section 3).
- Novel features to effectively capture these characteristics for automated prediction of task continuation (Section 4).
- Techniques for accurate prediction of continuation that outperform both a state-of-the-art automatic baseline and human predictions, coupled with the analysis of the most effective features used by the predictive algorithms (Section 5).

Next, we present related work to put our contributions in context.

2. RELATED WORK

Prior research that relates to what we describe in this paper falls into four main areas: (i) behavioral analysis and modeling of search, (ii) understanding search intent, (iii) analysis of cross-session tasks, and (iv) task switching and interruptions.

* Work done while visiting Microsoft Research.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.

Search behavior has been studied intensely in recent years. Log data from search engines have proven to be extremely valuable in studying how people search in naturalistic settings across a wide variety of different search intents. Most previous work has focused on search behavior analysis and prediction within a single search session [1][7][42], and related queries within a session can be part of a *search goal* [16][19], which try to represent the more abstract concept of search intent given only observable events. However, there is growing interest in using long-term search log data to build models of users' interests [39] and improve search result ranking [34].

An important part of representing search intent is understanding the various types of search tasks and the different motivations that searchers may have for pursuing their information goals. Earlier work on understanding search behavior focused on classifying queries into high-level search goals, such as informational, navigational and transactional [6][8][32]. Kellar *et al.* [20] conducted a field study in which they logged detailed Web usage and asked participants to provide task categorizations of their Web usage based on the following categories: fact finding, information gathering, browsing, and transactions. They showed differences in search behavior per task type. In particular, information gathering tasks were the most complex; participants spent more time completing this task, viewed more pages, and used the Web browser functions most heavily during this task. Li and Belkin [23] review and discuss previously-proposed task classifications and develop a faceted classification that can be used to describe searchers' work tasks and information search tasks. They identify essential facets and categorize them into generic task facets (e.g., source, product, and goal) and common task attributes (e.g., task characteristics and user perceptions). Rather than characterizing the nature of the search intent, Radlinski *et al.* [30] model search intent from queries and clicks in a way that could be directly consumed by search engines. Goals and related constructs (such as search intent) have also been widely studied in psychological research. Austin and Vancouver [4] review the theoretical development of the structure and properties of goals, goal establishment and striving processes, and goal-content taxonomies, which we use to motivate the selection of task dimensions to analyze. In fact, to our knowledge, our research is the first attempt to bring theory of motivation from psychology to bear on search intent analysis.

In this paper we focus on tasks extending across multiple sessions. Search behavior can be analyzed over time to identify queries that express the same underlying information need. Previous work has tried to automatically identify queries on the same task. Mei *et al*. [26] proposed a framework to study sequences of search activities and focused on simple prediction and classification tasks, ranging from predicting whether the next click will be on an algorithmic result to segmenting the query stream into goals and missions. Teevan *et al*. [37] showed, via query log analysis, that nearly 40% of queries were attempts to re-find previously encountered results. Aula *et al*. [3] studied the search and information re-access strategies of experienced Web users using a survey. They found that people often have difficulty remembering the queries they used originally to discover information of interest. MacKay and Watters [25] explored a variety of Web-based information seeking tasks and found that almost 60% of complex information gathering tasks continued across sessions. Liu and Belkin [24] examined the structure (parallel or dependent) of tasks that extend across different search sessions. Jones and Klinker [19] proposed methods to partition a query stream into research missions and goals, where each mission corresponds to a set of related information needs and may include multiple search goals. Morris *et al.* [27] developed *SearchBar*, a system that proactively and persistently stores query histories, browsing histories, and users' notes and ratings. SearchBar supports multi-session investigations by assisting with task context resumption and information re-finding. Donato *et al.* [10] developed *SearchPad*, a system that automatically identifies research missions and presents a search workspace comprising previous queries and results related to the mission. SearchPad uses measures of topic coherence between pairs of consecutive queries and user engagement to identify such research missions. This work was further extended by Aiello *et al.* [2] to group queries into mission-coherent clusters based on searcher behavior. However, none of the research described so far specifically addressed the important challenge of predicting search task continuation.

The most similar research to this paper is that of Kotov *et al.* [21]. In that paper, the authors describe research on modeling cross-session information needs, and address the challenge of identifying all previous queries in a user's search history on the same task as the current query, and predicting whether a user will return to the task in future sessions. Kotov *et al.* developed classifiers for these two tasks and through evaluation using labeled data from search logs showed that their classifiers can perform both tasks effectively. We use these classifiers as a baseline for some of the analysis presented later in the paper.

Also relevant to this work is previous research on task switching and interruptions. Multi-tasking and external factors such as interruptions have been previously associated with prolonged search tasks. Spink [35] studied the multi-tasking behavior of a single searcher in a public library using diary, observation and interviews and found that switching between tasks was common. On the basis of that study, she then developed a model of information multi-tasking and information task switching. Czerwinski *et al.* [9] present the findings of a week-long diary study of task interleaving admist interruptions, following eleven information workers in a non-search setting. They show that task complexity, task duration, length of absence, interruption count, and task type influence the perceived difficulty of switching back to tasks with participants reporting that it was most difficult to recommence complex tasks. The features we devise to represent tasks adapt and operationalize these ideas.

The research presented in this paper extends previous work in a number of ways. First, we perform a detailed descriptive analysis of the cross-session search tasks that maps task intents and motivations derived from the information science and psychology literatures to evidence of task continuation mined from search logs and labeled by trained human annotators. Second, we propose new features to model characteristics of cross-session search tasks, focusing on future task continuation, using features of search behavior mined from annotated log data. Third, we show that these features can improve continuation modeling and prediction over a previously-reported state-of-the-art baseline, and even over experienced human annotators attempting to perform the same prediction task.

3. UNDERSTANDING CROSS-SESSION TASKS: DESCRIPTIVE ANALYSIS

This section describes the data collection (Section 3.1), the human data annotation for the dimensions hypothesized to be related to search task continuation (Section 3.2), and presents analysis of task characteristics (Sections 3.4-3.5) based on both manual annotation of the tasks and an extended set of search log data.

3.1 Data Collection

The data were gathered from the Microsoft Bing commercial Web search engine by sampling a set of sessions over a one-week period for more than 1,000 users. Similar to [21], we study what have been previously defined as "early dominant" tasks identified for each user. An early dominant task is defined as having at least two distinct queries issued within a two-day period at the beginning of the week of interest. Some of these tasks are continued later during the week, while others are not. The data that we used for our study are summarized in Table 1.

Users and tasks (1 early-dominant task per user)	1,191
Unique queries	28,474
Active period	Last week of February 2010
Prior history	2 weeks prior
Continued tasks	683 (57%)

Table 1. Search log data used in this study.

Additionally, the data above were augmented by extracting up to an additional two weeks of prior history for each user in the sample, from the two weeks immediately before the week of interest. This history contained search sessions determined based on a 30-minute inactivity timeout [40], as well as the queries and URLs issued and visited. This allowed us to study the potential for utilizing additional profile information to predict task continuation.

3.2 Data Annotation

We annotated the characteristics of early dominant search tasks (defined above) according to a range of dimensions derived from information and cognitive science literatures, following the procedure in Section 3.3. Our goals were: (1) to analyze the relationship of task characteristics to task continuation and (2) to learn to automatically identify these characteristics for better search continuation modeling and prediction. In particular, we wished to investigate how task *intent* and *motivation*, as well as other contextual factors such as task *urgency,* relate to the likelihood of continuing a search task (within the one-week horizon that we used in our study). In the remainder of this subsection we define the dimensions on which we annotated tasks.

Intent Type: The type of the task, derived from previous studies in the information science literature (e.g., [20][23]). The hypothesis is that some task types, such as information gathering or transactions, are associated with task continuation. The specific intent types chosen for labeling were:

- *Fact finding (focused)*: Find specific piece(s) of information (e.g., a query such as "mc gilvery oil wolsey").
- *Information gathering (exploration)*: Find information on a topic rather than for a specific fact (e.g., "english comedy").
- *Undirected browsing*: Explore a site or the Web without an obvious goal (e.g., "portland craigs list").
- *Transaction*: Accomplish a task or perform a transaction online (e.g., "pay discover card bill")
- *Communication (social)*: Read or interact in online social sites such as forums.
- *Information maintenance or update*: Monitor information on a running topic and possibly update a Web resource.

Motivation: The cognitive or affective motivation inferred to be behind the task, derived and simplified from cognitive science and psychology literature [4]. Our intuition was that some motivations are more likely to associate with task continuation than others. The motivations selected for labeling were:

- *Affective*: Based on emotion or feeling, with sub-categories of *Arousal* (e.g., adult content), *Tranquility* (e.g., viewing art), *Happiness*, and *Physical well-being* (e.g., verifying health information).
- *Cognitive*: Learning about the world or about the self, with sub-categories of *Exploration*, *Understanding*, and *Positive self-evaluation*.
- *Self-assertive*: Individual relationship between person and the environment, with sub-categories of *Individuality, Self-Determination, Superiority,* and *Approval* (e.g., posting on a support forum).
- *Social*: Integrative social relationships, with sub-categories of *Belongingness* (maintaining social relationships), *Social Responsibilities*, or providing *Social Support.*

If none of the specific subtypes seemed appropriate, the annotators had an option to pick a generic motivation (e.g., "Social").

Complexity: The complexity of the task, measured by the number of goals required to find the needed information. We hypothesized that more complex tasks, with multiple goals, are more likely to be continued. The options for complexity were:

- *Single goal*: A task that can be theoretically satisfied by a single web page (e.g., "women's suffrage 1922").
- *Multiple goals*: A task that is expected to require aggregating information from multiple web pages (e.g., "cheap flights").
- *Undirected*: No evident goal (may be undirected exploration).

We asked annotators to specify the number of goals (if the task was not labeled as "undirected") based on their estimates of the number of Web pages required to fulfill the searcher's information need (one=single goal, many=multiple goals).

WorkOrFun: Does the task appear to be necessary for work or life or is it more for fun? We hypothesized that fun-related tasks are more likely to be continued than those considered to be work-related.

Time Sensitivity: How urgent or time sensitive is the information need, and is it likely to disappear/expire in a short time? Naturally, we hypothesized that highly time-sensitive tasks are less likely to be continued.

Continue or Not?: Finally, we asked the annotators to predict how likely they think a task is to continue within the week's data horizon. The following four response options were available: [*very likely, likely, unlikely, very unlikely*]. We hypothesized that human judges would be able to use their world knowledge and intuition to reasonably estimate the likelihood of task continuation, given the information available to them from the first two days of search behavior (e.g., all of the queries that users had issued, the URLs they had clicked, and the time of these events). These manually-generated estimates serve as a baseline for the performance of the predictive models developed in this paper.

3.3 Annotation Procedure and Agreement

The human annotations were performed at the *task* level, where each task was previously identified as "early dominant" by a human annotator (defined in Section 3.1 above), using a separate manual annotation process described in detail in reference [21]. For each of these tasks, the annotators were shown the sequences of queries, clicks, and date/times, with corresponding session identifiers, as well as all other search actions of that user (regardless of the task)The actual labeling was performed only for the early-dominant tasks. The four annotators reviewed the guidelines

for the above intents and motivations and worked through more than 20 example search tasks together, to ensure consistent interpretation and application of the guidelines. Annotators labeled an average of nearly 300 search tasks each, with three of them contributing over 90% of the labels.

An additional sample of 100 tasks was labeled by the three annotators responsible for the bulk of the labeling, for the purposes of computing inter-annotator agreement statistics. The average annotator agreement and the free-marginal Fleiss Kappa statistic [31] are reported in Table 2[1].

Dimension	*Average Agreement*	*Free-marginal Kappa*
Intent	0.649	0.591
Motivation	0.649	0.532
Complexity (# goals)	0.712	0.568
Time sensitivity	0.698	0.547
Work or Fun?	0.677	0.516

Table 2. Inter-annotator agreement for goal/intent labels (across the additional 100 tasks).

The agreement ranges from 0.65 to 0.71, with *Kappa* values between 0.52 for "WorkOrFun" to 0.59 for "Intent". These values are acceptable for such a difficult and potentially-subjective task.

3.4 Task analysis: Intent and Motivation

The majority of tasks were labeled as *information gathering (exploratory)* (56%), examples of which included research, school work, shopping, and travel planning. The other tasks were labeled as *fact finding (focused)* (20%), and *transaction* (13%), with the remainder of the search intents comprising 2-4% each.

The task continuation statistics for these intents are reported in Figure 1. Information maintenance tasks were most likely to be continued (85%), followed by undirected browsing (78%). Both of these may reflect hobbies and other longer term interests of the users in our study. Interestingly, *transaction* and *communication* were also likely to be continued (both over 70-75%). One possible confound is that transaction tasks include a small fraction of *navigational re-finding*, even though by requiring at least two unique queries we attempted to filter out navigational queries. With 52% and 48% return rates, *information gathering* and *fact finding* tasks were less likely to be continued, perhaps because most these tasks were fairly simple and could be completed within a single session.

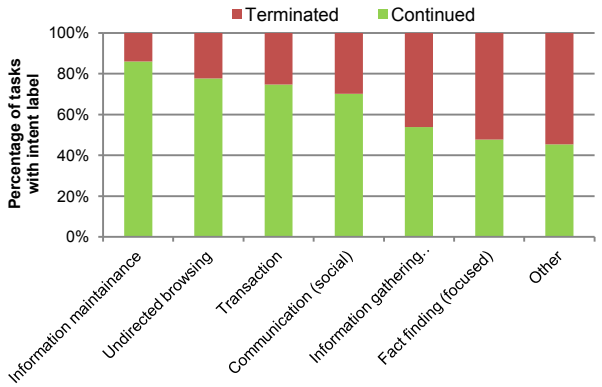

Figure 1. Task continuation for broad search intents.

[1] We use free-marginal Multi-rater Kappa since it is appropriate for typical agreement studies in which raters' distributions of cases into categories are unrestricted.

Figure 2 reports the task continuation statistics for different motivations for the tasks, in decreasing order by the likelihood of continuation. It appears that affectively motivated tasks are more likely to be continued, with arousal (typically, adult content) the most likely to be continued. Interestingly, *self-assertive* and *social* motivations were almost equally likely to result in task continuation, while tasks motivated by *cognitive: understanding* and *affective: physical wellbeing* were the least likely tasks to be continued. Tasks with these motivations do not typically persist over time, presumably because they involved episodic lookups of facts or health-related information that does not require follow-up.

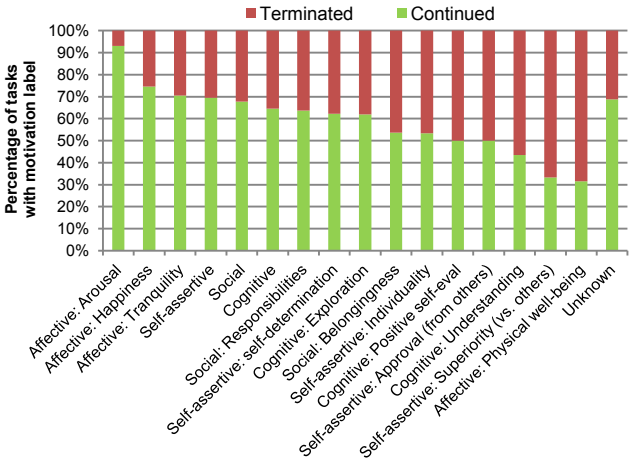

Figure 2. Task continuation for different search motivations.

In addition to analyzing variations in task continuation likelihoods associated with different intents and motivations, we were also interested in the impact of task complexity on the likelihood that users would continue. Figure 3 shows the relationship between the number of goals identified and the task continuation likelihood.

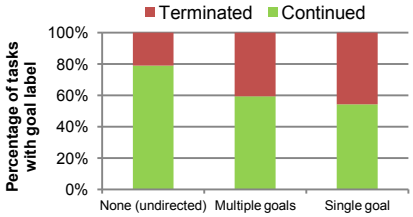

Figure 3. Task continuation by task complexity (number of goals).

Interestingly, the number of task goals (Figure 3) is not strongly associated with task continuation. In fact, the tasks that appear to be undirected (e.g., without a clear goal page or information nugget), are more likely to be continued. These include browsing employment opportunities, real estate listings, or adult content. Furthermore, tasks judged to be time-sensitive (Figure 4a), are more likely to be continued, compared to tasks judged to be not time-sensitive. Also tasks being attempted for pleasure (fun) rather than necessity (work-related) are also slightly more likely to be continued (Figure 4b). While this seems counter-intuitive, one explanation could be that when searching by necessity, users are more likely to satisfice once the (minimum) sufficient information is found, whereas curiosity- or pleasure-driven exploration are less likely to be satisfied as quickly, and is likely to be more aligned with the searcher's long-term interests. We explore this observation further in the next section.

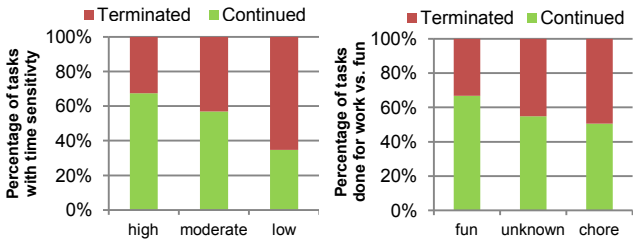

Figure 4. Task continuation by (a) time sensitivity and (b) work or fun task types.

We have seen in this section that several factors are associated with the likelihood of task continuation. In particular, tasks that give searchers pleasure and align with the users' interests are more likely to be continued, at least within the one-week period analyzed in this study, as we explore in more detail next.

3.5 Search Topic Analysis: Repeat History

We hypothesized that certain topical categories of tasks are more likely to be resumed than others (see also [10]). To identify topical category, we use automatic query classification into the top two levels of the Open Directory Project (ODP, dmoz.org) hierarchy. The classifier has a micro-averaged F1 value of 0.60 and is described more fully in reference [5]. To obtain a topic representation for queries labeled as belonging to the task of interest, we obtained the top ten results for each query from Bing and categorized each result by running the text classifier on its content. The result is a vector of topic probabilities, which we restricted to the three most probable classes. For each task, we obtained the most probable ODP category by merging the distributions for all associated queries.

Figure 5 reports that search tasks in some ODP categories, such as "adult", "kids and teens" and "news", are very likely to be continued, while search involvement with other topics, such as "home", "health", and "science" appear to be more episodic and less likely to be continued over time. Note that the search topic is distinct from the search intent (e.g., a task associated with "news" topics may be either *information maintenance*, or *fact finding*). Importantly, this demonstrates that the ODP category labels may be useful for automatically predicting task continuation. We explore the utility of this representation for prediction later in the paper.

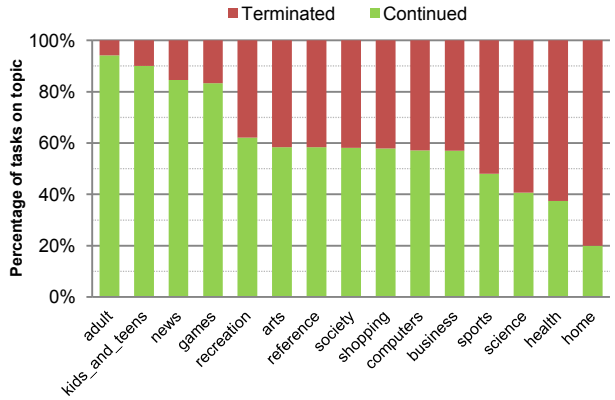

Figure 5. Task continuation by top-level ODP category.

The potential utility of the ODP category labels for task continuation prediction is not surprising, and indeed we observed anecdotally in our data that some topics were more likely to be repeated over time. These topic repeatability statistics could be considered as a "prior" for the task continuation likelihood, and could be exploited in the absence of any other information about the user.

Class Repeats: Specific Categories	*Repeat Prob.*
Most Likely	
Computers/Internet	0.639
Arts/Television	0.562
Adult/Computers	0.543
Arts/Radio	0.521
Adult/Image_Galleries	0.515
Games/Board_Games	0.503
Shopping/Antiques_and_Collectibles	0.483
Games/Video_Games	0.482
Arts/Music	0.469
Adult/World	0.469
Games/Card_Games	0.431
Shopping/General_Merchandise	0.431
Sports/Baseball	0.421
Adult/Arts	0.415
Shopping/Vehicles	0.413
Least Likely	
Shopping/Visual_Arts	0.047
Recreation/Living_History	0.042
Computers/Consultants	0.041
Recreation/Birding	0.041
Recreation/Climbing	0.040
Science/Instruments_and_Supplies	0.039
Arts/Writers_Resources	0.038
Reference/Museums	0.035
Society/Holidays	0.032
Recreation/Scouting	0.029
Health/Animal	0.029
Society/Gay,_Lesbian,_and_Bisexual	0.027
Business/International_Business_and_Trade	0.027
Arts/Illustration	0.022
Sports/Tennis	0.020

Table 3. Highest and lowest repeat probabilities for different ODP topical categories (large-scale sample).

To examine this observation in more detail we analyzed the probability that a given topical category will be observed in a future session for the same user within a week (similar to the setting used for this study). To do this, we used a separate set of Bing search logs for a period of three weeks that did not overlap with the one week of data used for our study. From these logs we extracted over 100 million search sessions for over five million unique users. Search sessions were defined using a 30-minute inactivity timeout [40]. The results are summarized in Table 3, and show that topics such as *Computers/Internet*, *Arts/Television*, and *Adult/Computers* are the most likely to be observed in subsequent search sessions, while topics such as *Sports/Tennis* or *Reference/Museums* are likely to be used in one session but not to appear in future sessions for the same user within the following week. The former set of categories may be more likely to reflect users' longer-term, persistent interests, whereas the latter may be more transient and affected by immediate social responsibilities e.g., specific events such as a museum visit or a tennis tournament.

4. MODELING TASK CONTINUATION

In the previous section, we analyzed the task continuation data with a focus on the characteristics of the search tasks that are associated with task continuation. We now turn to modeling and automatically predicting task continuation. As described earlier, this is an important area for search providers trying to help users perform cross-session searching. We first describe the features

used for task representation and then describe the algorithms and training procedure that we adopted in this study (Section 4.2).

4.1 Features

We represent a task using *topical*, *user engagement*, *user history profile*, and *topic and query priors* feature groups, described in more detail below and shown in Table 4. We use these features to predict task continuation.

Baseline features. We began by re-implementing the most important features reported in [21], which forms our baseline system in the prediction experiments. These features capture the basic lexicographic and behavior properties of the search session, such as query overlap, number of clicks on results returned by the search engine, and time between queries. Reference [21] provides more detailed descriptions of these features.

In addition to the baseline features, we also added four groups:

Search topic. These new features aim to capture the topical categories of the task derived from the automated classifier trained on ODP data and described in Section 3.5. Additional measures include the entropy of the topic distribution (for both the first- and second-level categories of the ODP hierarchy) to capture the degree of topical focus in the task. We conjectured that tasks that span fewer distinct ODP topics are more likely to be continued

User engagement. These new features aim to capture the searcher's level of engagement in the task they are performing, going far beyond the baseline features described above. Features of note include the estimated satisfaction and dissatisfaction with the results (based on estimates of the amount of time that users spent dwelling on clicked results, per [13]), the span of time and effort invested in the task, the amount of "multi-tasking" interspersed with the task, as well as other metrics of effort and user activity. We hypothesized that if a user is heavily engaged with a task and that effort is focused, they will be more likely to continue.

User profile history. In addition to analyzing the current search task, we also aim to capture historical information about the user. To do this we used two weeks of log data from the time period before the week of interest for each of the users in our study. Features generated from this profile include the topic distribution of previous search sessions, queries, overlap with the current task, and other profile information such as the time of the day and day of the week when the task was started. We hypothesized that topics or query terms that interested the user in the past, are more likely to be continued in the future.

Repeat priors on topic and query repetition. In addition to the random sample of the nearly 1,200 users under study, we make use of global query and ODP category statistics computed over the query log described in Section 3.5. We hypothesized that topics and query terms that tend to re-appear globally could provide additional evidence for task continuation.

4.2 Classifiers

We experimented with two different classifiers for the problem of predicting task continuation. The two classifiers used were Logistic Regression [15] (which was shown to be effective for task continuation prediction in reference [21]). We refer to this method as **Baseline** in subsequent experiments.

Our main experiments were performed using a gradient-**B**oosted Decision **T**ree classifier, based on the MART algorithm [14], with a logistic penalty, so that we can evaluate the importance of richer feature combinations. We refer to this classifier as **BT** (for Boosted Tree) in subsequent experiments, typically listed in combination with either all the features in Table 4 ("BT: All") or feature subsets. The classification task is to predict whether a search task, previously identified to be early-dominant for a user, will be continued in the future (positive class) or not (negative class). All experimental results reported below were performed using 5 runs of 10-fold cross validation, randomized for each method.

Name	*Description*
Baseline features	
BASE_SameQueryHist, BASE_NumSessHist, BASE_NumDomQueriesHist, BASE_AvgInterQTimeHist, BASE_FreqDomQueriesHist, BASE_NumDwell30Hist, BASE_NumQueryHist, BASE_NumTop10ClickQuery BASE_AvgInterQTimeSess BASE_NumClickHist BASE_NumQueryChars BASE_SubQueryHist BASE_SupQueryHist BASE_SubQuerySess BASE_SupQuerySess	Implemented as described in reference [21]
Topic	
NumClassifierLeafs	Number of distinct classifier topics clicked on task
NumODPCats	Number distinct ODP categories clicked
NumODPLeafs	Number distinct ODP leafs clicked
TopClassifierLeaf	Most frequent topic
TopOdpCat	Most frequent ODP category
TopODPLeaf	Most popular ODP leaf
OdpDomCatEntropy	Entropy of dominant task ODP categories
OdpDomLeafEntropy	Entropy of dominant task ODP leafs
ClassifierDomEntropy	Entropy of dominant task classifier leafs
Engagement Effort and Focus	
AvgClickPosQuery	Average position of result clicks for last query
TotalSkipPosQuery	Sum of skip positions for last query
AvgDomClickPos	Avg click pos for all dom task queries
TotalDomSkipPos	Sum of skip positions for all dom task queries
AvgClickPosSess	Avg click position for all queries in session
TotalSkipPosSess	Sum of skip positions for all queries in session
NumDomClickBacks	Total number of on-task click-backs
AvgDomClickBacks	Avg number of click backs per session
NumDomTaskSessions	# sessions that had at least one dom task query
NumOffTaskQueriesSess	Number off-task queries in last session
NumTaskSwitchHist	#off/on- task switches over recent history
NumTaskSwitchSess	#off/on- task switches over last session
OnTaskQueriesRatioSess	Fraction of on-task to total queries in last session
OnTaskSessionsRatio	Fraction of on-task to off-task sessions
TaskSpanTime	Time from first to last occurrence of dominant task
TaskSpanSessions	Sessions spanned by the dominant task
TaskSpanDays	Number of days for the task
NumDomResClicks	Total number of clicks on dominant task queries
AvgDomResClicks	Avg # clicks per query on dominant task
NumSATClicks	Clicks on dominant task with > 60 sec dwell time
NumDSATClicks	Clicks on dominant task with < 30 sec dwell time
User Profile	
SameQueryPriorHist	1 if last dom task query appeared in prior history
DomQueriesPriorHist	Number of dom task queries in prior history
NumDomTopicPriorHist	Number of dom topics in prior history
FracDomTopicsPriorHist	Fraction of dom task topics in prior history
FracPriorHistDomTopics	Fraction of topics in prior history also in dom task
NumTopicPriorHist	Total number unique topics in prior history
ProbPriorHistDomTopics	Sum of dom task topic probabilities
TopDomTopicsInPriorHist	Top K (k=10) most popular topics in prior history
TaskStartDayOfWeek	Day of week (Sun=0) for the start of the task
TaskStartTimeOfDay	Time of day (midnight=0) for the start of the task.
Repeat Priors	
CatRepeatPrior	Probability of category repeat for the same user
QueryRepeatPrior	Probability of query repeat for the same user
TermRepeatPrior	Probability of term repeat for the same user

Table 4. Features used to represent cross-session search tasks.

4.3 Evaluation Metrics

To compare the performance of the classification methods we use the following standard performance measures: (1) accuracy, (2) precision and recall for the positive class (task continuation), and (3) area under the receiver-operator-characteristic curve (AUC). Statistical significance between performance values was calculated using two-tailed independent sample *t*-tests where appropriate.

5. RESULTS AND DISCUSSION

This section first reports the human performance on the task, to indicate that predicting continuation is challenging even for human judges (Section 5.1). We then report the results of automatic continuation predictors (Section 5.2), followed by extensive analysis of the feature groups and individual features that are most strongly predictive of task continuation (Section 5.3). Finally, we present the findings of a failure analysis which suggests future improvements to our predictive models.

5.1 Human Prediction of Task Continuation

Table 5 reports the performance of the classifier trained on individual task dimensions (manually labeled as described above), as well as on the explicit human judgment of task continuation, and a classifier trained on the combination of all of the manual annotations. In other words, we attempt to create the best "hybrid" human and machine prediction possible, by using the labels provided by the human judges as features. These labels were expected to augment the explicitly labeled "continue or not" prediction (which was considered positive when the response was "very likely" or "likely", and negative otherwise). Recall, that the annotators were able to see the first two days of the user's search history, but did not have access to the longer-term User Profile features above). So, the humans' predictions were performed based on two days of data as well as world knowledge and intuition about the nature of search tasks. As Table 5 indicates, humans can definitely predict continuation more accurately than the naïve Majority baseline that always picks "continue", or than any individual intent or motivation label. However, there is an even stronger signal in the *combination* of the manual dimension labels, resulting in the best prediction possible based on the human judgments data.

Method (Human Annotations)	*Accuracy*	*Precision*	*Recall*	*AUC*
Majority baseline ("continue")	0.573	0.573	1.000	0.573
Task type	0.586	0.591	0.842	0.566
Motivation	0.629	0.660	0.738	0.628
Complexity	0.572	0.575	0.986	0.506
WorkOrFun	0.550	0.574	0.852	0.579
Continue prediction	0.677	0.730	0.703	0.692
BT: All human labels	**0.678**	**0.704**	**0.764**	**0.729**

Table 5. Human task annotations vs. search continuation.

In addition to computing the predictive value of the task dimensions and their combination, we were also interested in the relationship between the nature of the human judges' estimation of continuation likelihood and whether users were observed to be continuing the search task. We did this to help us to understand their ability to make the explicit prediction (rather than using their prediction as a feature for learning). Figure 6 shows the predicted versus actual outcomes for each of the four rating options.

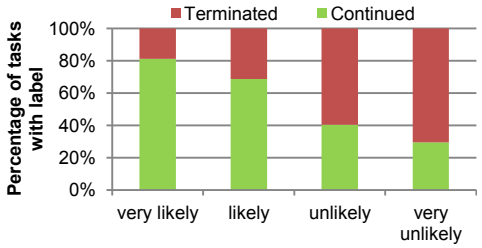

Figure 6. Predicted vs. actual task continuation by human annotators.

As Figure 6 indicates, when human judges were sure of the prediction (i.e., rated a task to be "very likely" or "very unlikely" to be continued), their prediction accuracy was 80% and 75%, respectively. However, for the majority of cases, the annotators provided more tentative labels ("likely" and "unlikely"), and in those cases the predictions had substantially lower accuracy..

We now focus on the predictive performance of the trained models, comparing them with the human predictive performance.

5.2 Comparing Prediction Methods

Table 6 reports the performance of a human prediction of task continuation, against our implementation of the state-of-the-art baseline described in [21], and our extended method *BT:All* (using all classifier features and two weeks of prior history as described above). Both classifiers substantially outperform the human annotators; furthermore, *BT:All* substantially and significantly outperforms the baseline in terms of accuracy, precision, and AUC metrics with $p < 0.01$.

Method	*Accuracy*	*Precision*	*Recall*	*AUC*
Human prediction	0.677	0.730	0.703	0.692
Baseline	0.697	0.728	0.761	0.752
BT: All Features	**0.751** (+8%)**	**0.786** (+8%)**	**0.783 (+3%)**	**0.829** (+10%)**

Table 6. Predicting search continuation (* and ** indicate statistical significance at $p \leq 0.05$ and $p \leq 0.01$ compared to the Baseline, respectively, using unpaired t-test).

We augment the quantitative analysis in Table 6 by plotting the precision-recall curve for each of the methods in Figure 7.

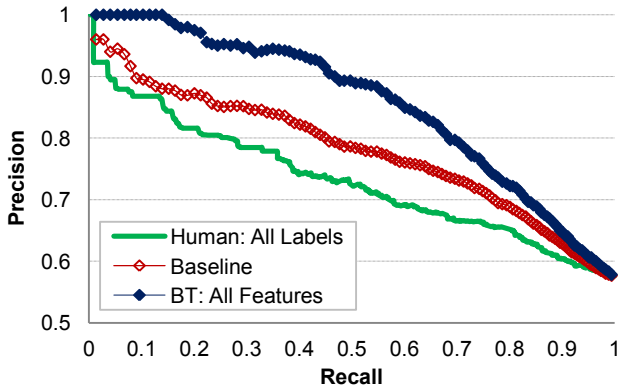

Figure 7. Precision-Recall curve for predicting search continuation for the Baseline, BT:All, and the Human prediction methods.

As Figure 7 indicates, both the automated *Baseline* and our *BT:All Features* classifier substantially outperform human predictions. Furthermore, *BT:All* provides the biggest lift in AUC over the Baseline, at precision of at least 0.8 and remains acceptably high (≥ 0.75) until nearly 0.8 recall levels.

One factor that may affect the performance of the *BT:All* classifier is the availability of user history information. This information was not available to humans (although they can draw upon general world knowledge and their own search experiences) or Baseline. To quantify the contribution of user's history (profile) information to use for prediction, Table 7 reports the results of varying the amount of history data for each user included in the model from *None* (i.e., no user profile information prior to the two days at the beginning of prediction), to one week and two weeks.

Method	*Accuracy*	*Precision*	*Recall*	*AUC*
BT: All: No History	0.721	0.758	0.761	0.788
+ 1 Week History	0.731	0.766	0.768	0.791
+ 2 Weeks History	**0.751****	**0.786****	**0.783***	**0.829****

Table 7. Varying amount of history for user profile computation (* and ** indicate statistical significance at p≤0.05 and p≤0.01 compared to the *BT: All: No History* method, respectively, using unpaired t-test).

While adding one week of prior history improves performance slightly on all metrics, the improvements are not significant. However, the effects of adding an additional week of history (for two weeks total) are striking, providing substantial and significant improvements (with $p < 0.01$ for accuracy, precision, and AUC metrics, and $p < 0.05$ for Recall), compared to the same method with no prior user history. Having multiple weeks may more effectively capture the users' long-term interests or allow for recurring tasks (e.g., those that happen biweekly) to be observed and used to make predictions for the current week. More research is needed to determine whether such gains consistently increase with history length. Also note that even when we remove the history features, the performance of *BT:All: No History* is still substantially better than the performance of both Baseline and the human annotators. We discuss the differences between human and machine performance in more detail later in Section 5.4.

An important question in understanding the success of *BT:All* model is determining which feature groups contributed the most toward its strong performance. To this end, we now present some feature ablation analysis.

5.3 Feature Ablation Analysis

In order to determine the contribution provided by each of the feature groups, we perform feature ablation experiment, by starting with the full set of features (Table 4), and then systematically removing feature groups from the set, one group at a time. The results are reported in Table 8, averaging over five runs of randomized 10-fold cross validation.

Surprisingly, removing text features such as the most frequently used query terms, has negligible effect on performance. In contrast, removing the user profile features computed over the user history degrades performance significantly on the accuracy, precision, and AUC metrics. While other feature groups also appear to contribute, the single most valuable feature group is the user engagement effort and focus (listed in Table 4). Removing these features degrades performance significantly to roughly that of the original baseline. It appears that the more engaged the user is with the search task, the more likely they are to continue it in the one-week time span of our study.

Feature set	*Accuracy*	*Precision*	*Recall*	*AUC*
BT: All Features	**0.751**	**0.786**	**0.783**	**0.829**
– Topic	0.747	0.786	0.774	0.823
– **Engagement Effort and Focus**	0.703**	0.742**	0.746**	0.773**
– **User profile**	0.732**	0.770*	0.765	0.798**
– Repeat Priors	0.750	0.785	0.783	0.824

Table 8. Feature group ablation: task continuation prediction when removing one feature group at a time (* and ** indicate statistical significance at p≤0.05 and p≤0.01 compared to the *BT: All* method, respectively, using unpaired t-test).

5.3.1 Individual Features: Engagement and Focus

With the importance of the engagement features apparent in the feature-group ablation study, we set out to investigate how each of the individual engagement features correlates with task continuation. We computed the Pearson's correlation coefficient (r) between each of the engagement feature values and the task continuation label. The r values for the most and least correlated features are shown below in Table 9.

Engagement Feature	*r*
TaskSpanTime	**0.412**
NumDomTaskSessions	**0.387**
TaskSpanSessions	**0.364**
TaskSpanDays	**0.321**
OnTaskSessionsRatio	**0.230**
NumSATClicks	**0.178**
NumDomResClicks	0.162
TotalDomSkipPos	0.120
…	
NumOffTaskQueriesSess	-0.029
NumDomRepeatQueriesSess	-0.040
AvgDomClickBacks	-0.044
TotalSkipPosQuery	-0.052
OnTaskQueriesRatioSess	-0.055
NumTaskSwitchSess	**-0.148**

Table 9. Individual search engagement features vs. task continuation.

Table 9 reports that the most strongly correlated engagement features (from Table 4) include *TaskSpanTime*, *NumDomTaskSessions* (the number of on-task domain sessions), features such as the ratio of on-task vs. off-task sessions, and the number of SAT clicks (defined as clicks with dwell time ≥ 60). These features indicate that searchers tend to be more strongly focused on tasks that would be continued, and less involved in multi-tasking. In contrast, the *lack* of focus (e.g., increased multi-tasking during search as measured by the *NumTaskSwitchSess* feature), correlates negatively with task continuation. Interestingly, there is also a moderate correlation between search satisfaction (measured by the number of SAT clicks (defined in Table 4) and task continuation, and slightly negatively correlated to the number of "click-backs" or bounce-backs (*AvgDomClickBacks*), which are indicative of unsatisfying results. Indeed, Hu *et al.* [18] found a positive correlation between searcher satisfaction and search engine re-use on a week by week basis over a six-month period, although they just studied query volume and not task continuation.

To understand how the individual features contribute *in combination* with the other features for the BT classifier, Table 10 reports the individual features that contributed the most to reducing error during training – those with highest average relative reduction in residual squared error (averaged across the cross validation folds).

Feature	*Average Gain*
TaskSpanTime	**1**
DomQueriesPriorHist	**0.624**
BASE_NumQueryChars	**0.371**
TermRepeatPrior	**0.362**
BASE_AvgInterQTimeSess	**0.342**
BASE_AvgInterQTimeHist	**0.338**
BASE_NumQueryHist	**0.321**
TaskStartTimeOfDay	**0.289**
OnTaskQueriesRatioSess	**0.276**
OdpDomCatEntropy	**0.275**
FracPriorHistDomTopics	0.273
ClassifierDomEntropy	0.271
ProbPriorHistDomTopics	0.264
ClassRepeatPrior	0.260
BASE_FreqDomQueriesHist	0.252
NumTopicPriorHist	0.248
AvgDomClickPos	0.233
BASE_NumDwell30Hist	0.232
BASE_NumClickHist	0.224
BASE_NumSessHist	0.224
QueryRepeatPrior	0.218
AvgDomResClicks	0.210
BASE_SameQueryHist	0.208
NumDomTaskSessions	0.208
AvgDomClickBacks	0.201

Table 10. Feature importance: Top features, averaged across folds.

The single strongest predictor of continuation is *TaskSpanTime* – the amount of time the searcher already has spent on the task in the first two days. More interestingly, the next strongest indicator is *DomQueriesPriorHist*, the number of queries in the dominant task that were observed in the prior history for that user (hence pulling in information about the longevity of the task). The previously studied feature *BASE_NumQueryChars* (the number of characters in the query) is another strong indicator, perhaps because longer queries say something about the nature of the task or are suggestive of users' knowledge of a particular domain of interest. Previous studies have shown that users issue longer queries when searching in a domain about which they are knowledgeable [40]. Another new indicator of continuation is the *TermRepeatPrior* – the distribution of term repeat probabilities computed over more than five million users. This suggests that prior likelihoods of repetition computed independent of user may also be useful for predicting task continuation. Other strong indicators include the task start time of day, as well as the task focus and engagement features discussed above.

5.4 Analysis and Study Limitations

To gain additional insights into the task continuation prediction problem, we compared the task continuation predictions made by human annotators (using the labeling procedure described in Section 3.3), with those of the automated classifier, by focusing on the cases where the human and the classifier predictions differed. In these cases, humans were able to predict task *termination* more accurately (by about 25%) than task continuation. We conjecture that this could be partly explained to the data collection constraints, in that we required for a task to be resumed within one week in order to be marked "Continued". Furthermore, humans were able to predict continuation more accurately (by about 30%) for tasks labeled as difficult, compared to those labeled as easy or moderate. We also found that humans were most accurate in predicting task continuation within a day, which intuitively covers many transaction and planning tasks. Most remarkably, in cases where a task query previously occurred in a user's prior history, the classifier was 73% more likely to be correct than human judges, who did not have access to the long-term user profiles and had to rely on their intuition and world knowledge instead.

At this point it is appropriate to point out limitations of this study. As discussed earlier, to make this study feasible, the time horizon of task continuation was limited to one week. It is possible that some tasks are continued more than a week later. Another limitation is methodological: without asking the users directly, it was not possible for us to investigate the underlying causes of the search interruptions, i.e., whether the searchers were waiting for external input or simply ran out of time and had to switch to another task first. To get that information we will need to work with users directly and employ in-situ or retrospective methods to better understand the factors affecting their observed behaviors.

6. CONCLUSIONS

Many search tasks such as travel planning, making large purchases or job searches can span multiple search session extending over hours, days, or even weeks. The research reported in this paper is aimed at better understanding, characterizing, and automatically predicting search tasks that will be continued in the future – a task complementary to that of identifying previous related search sessions after the fact. We first annotated a query log from Bing to identify the types of search intents, motivations, and topics associated with search tasks that are continued. We then developed new features of search engagement and focus in search sessions, which we use as input for prediction. Finally we developed effective prediction algorithms that significantly outperform both the previous state-of-the-art method, as well as the ability of human judges to predict search continuation.

Our analysis of the task continuation yielded effective features that allow our prediction method to significantly outperform both a state-of-the-art baseline and predictions based on human judgments. We identified the groups of features (user prior history and task engagement) that most strongly predict task continuation. We also identified individual task characteristics, such as time span, user focus, and task start time that are most strongly correlated with continuation.

Future research directions include studying task continuations occurring beyond the one-week time frame studied here, as well as looking back further in time to build more complete historic profiles of users' search interests given the observed value of such historic information for task continuation prediction. We are also interested in applying task continuation predictions to improve search result ranking (e.g., by personalizing the results to the active task and not to the whole user's profile); and in better supporting searchers in resuming tasks (e.g., if a task is likely to be continued, saving the results already found for more convenient access once the task is continued). Thus, the work presented here forms an important advance towards developing a more personal-

ized and intelligent solution for long-running, complex tasks – a key challenge for the information retrieval community.

7. ACKNOWLEDGMENTS

We thank Alexander Kotov and Jaime Teevan for collecting and labeling earlier data for this paper, and Dan Liebling and Shane Williams for assistance with data processing. We also thank Filip Radlinski for insights on quantifying complexity of search tasks.

REFERENCES

[1] E. Agichtein, E. Brill and S. Dumais. Improving Web search ranking by incorporating user behavior information. *SIGIR '06*, 19–26, 2006.

[2] L. M. Aiello, D. Donato, U. Ozertem, and F. Menczer. Behavior-driven clustering of queries into topics. *CIKM '11*, 1373–1382, 2011.

[3] A. Aula, N. Jhaveri and M. Käki. Information search and re-access strategies of experienced Web users. *WWW '05*, 583–592, 2005.

[4] J.T. Austin and J.B. Vancouver. Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin, 120(3)*: 338–375, 1996.

[5] P.N. Bennett, K. Svore and S.T. Dumais. Classification-enhanced ranking. *WWW '10*, 111–120, 2010.

[6] A. Broder. A taxonomy of Web search. *SIGIR Forum*, *32(2):* 3–10, 2002.

[7] 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. *KDD '08*, 875–883, 2008.

[8] Y.-S. Chang, K.-Y. He, S. Yu and W.-H. Lu. Identifying user goals from Web search results. *WWW '06*, 1038–1041, 2006.

[9] M. Czerwinski, E. Horvitz and S. Wilhite. A diary study of task switching and interruptions. *CHI '04*, 175–182, 2004.

[10] D. Donato, F. Bonchi, T. Chi and Y. Maarek. Do you want to take notes? Identifying research missions in Yahoo! Search Pad. *WWW '10*, 321–330, 2010.

[11] D. Downey, S.T. Dumais, D. Liebling and E. Horvitz. Understanding the relationship between searchers' queries and information goals. *CIKM'08,* 449–458, 2008.

[12] S. Dumais, G. Buscher and E. Cutrell. Individual differences in gaze patterns for web search. *IIiX '10*, 185–194, 2010.

[13] S. Fox, K. Karnawat, M. Mydland, S. T. Dumais and T. White. Evaluating implicit measures to improve the search experience. *TOIS, 23(2)*: 147–168, 2005.

[14] J. Friedman. MART boosted decision trees: Greedy function approximation: a gradient boosting machine. *Annals of Statistics, 25(5)*: 1189–1232, 2001.

[15] J. Friedman, T. Hastie and T. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, *28(2)*: 337–407, 2000.

[16] A. Hassan, R. Jones and K. Klinkner. Beyond DCG: User behavior as a predictor of a successful search. *WSDM '09*, 221–230, 2010.

[17] A. Hassan, Y. Song and L. He. A task level metric for measuring web search satisfaction and its application on improving relevance estimation. *CIKM'11,* 125–134, 2011.

[18] V. Hu, M. Stone, J. Pedersen and R.W. White. Effects of search success on search engine re-use. *CIKM'11*, 1841–1846, 2011.

[19] R. Jones and K. Klinkner. Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. *CIKM '08*, 699–708, 2008.

[20] M. Kellar, C. Watters and M. Shepherd. A field study characterizing Web-based information-seeking tasks. *JASIST, 58(7)*: 999–1018, 2007.

[21] A. Kotov, P.N. Bennett, R.W. White, S.T. Dumais, and J. Teevan. Modeling and analysis of cross-session search tasks. *SIGIR'11,* 5–14, 2011.

[22] U. Lee, Z. Liu and J. Cho. Automatic indentification of user goals in Web search. *WWW '05*, 391–400, 2005.

[23] Y. Li and N.J. Belkin. A faceted approach to conceptualizing tasks in information seeking. *IP&M*, 44(6): 1822–1837, 2008.

[24] J. Liu and N.J. Belkin. Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. *SIGIR '10*, 26–33, 2010.

[25] B. MacKay and C. Watters. Exploring multi-session Web tasks. *CHI '08*, 1187–1196, 2008.

[26] Q. Mei, K. Klinkner, R. Kumar and A. Tomkins. An analysis framework for search sequences. *CIKM '09*, 1991–94, 2009.

[27] D. Morris, M. Ringel Morris and G. Venolia. SearchBar: A search-centric Web history for task resumption and information re-finding. *CHI '08*, 1207–1216, 2008.

[28] B. Piwowarski and H. Zaragoza. Predictive user click models based on click-through history. *CIKM '07*, 175–182, 2007.

[29] F. Radlinski and T. Joachims. Query chains: Learning to rank from implicit feedback. *KDD '05*, 239–248, 2005.

[30] F. Radlinski, M. Szummer and N. Craswell. Inferring query intent from reformulations and clicks. *WWW '10*, 1171–1172, 2010.

[31] J.J. Randolph. Free-marginal multirater Kappa (multirater K free): an alternative to Fleiss' fixed-marginal multirater Kappa. *JULIS'05*, 2005.

[32] D.E. Rose and D. Levinson. Understanding user goals in Web search. *WWW '04*, 13–19, 2004.

[33] M.D. Smucker, J. Allan and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. *CIKM'07*, 623-632, 2007.

[34] D. Sontag, K. Collins-Thompson, P.N. Bennett, R.W. White, S.T. Dumais and B. Billerbeck. Probabilistic models for personalizing web search. *WSDM '12*, 433–442, 2012.

[35] A. Spink. Multitasking information behavior and information task switching: an exploratory study. *J. Documentation*, *60(4)*: 336–351, 2004.

[36] B. Tan, X. Shen and C. Zhai. Mining long-term search history to improve search accuracy. *KDD '06*, 718–723, 2006.

[37] J. Teevan, E. Adar, R. Jones and M.A.S. Potts. Information re-retrieval: Repeat queries in Yahoo's logs. *SIGIR '07*, 151–158, 2007.

[38] E.G. Toms, L. Freund, R. Kopak and J.C. Bartlett. The effect of task domain on search. *CASON'03,* 303-312, 2003.

[39] R.W. White, P. Bailey and L. Chen. Predicting user interests from contextual information. *SIGIR '09*, 363–370, 2009.

[40] R.W. White and S.M. Drucker. Investigating behavioral variability in Web search. *WWW '07*, 21–30, 2007.

[41] R.W. White, S.T. Dumais, and J. Teevan. 2009. Characterizing the influence of domain expertise on Web search behavior. *WSDM '10*, 132–141.

[42] B. Xiang, D. Jiang, J. Pei, X. Sun, E. Chen, and H. Li. Context-aware ranking in Web search. *SIGIR'10*, 451–458, 2010.