

Rain or Shine? Forecasting Search Process Performance in Exploratory Search Tasks

Chirag Shah

School of Communication and Information, Rutgers University, 4 Huntington Street, New Brunswick, NJ 08901-1071. E-mail: chirags@rutgers.edu

Chathra Hendaheewa

Department of Computer Science, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854-8019. E-mail: chathrah@cs.rutgers.edu

Roberto González-Ibáñez

Departamento de Ingeniería Informática, Universidad de Santiago de Chile, Santiago, Chile. E-mail: roberto.gonzalez.i@usach.cl

Most information retrieval (IR) systems consider relevance, usefulness, and quality of information objects (documents, queries) for evaluation, prediction, and recommendation, often ignoring the underlying search process of information seeking. This may leave out opportunities for making recommendations that analyze the search process and/or recommend alternative search process instead of objects. To overcome this limitation, we investigated whether by analyzing a searcher's current processes we could forecast his likelihood of achieving a certain level of success with respect to search performance in the future. We propose a machine-learning-based method to dynamically evaluate and predict search performance several time-steps ahead at each given time point of the search process during an exploratory search task. Our prediction method uses a collection of features extracted from expression of information need and coverage of information. For testing, we used log data collected from 4 user studies that included 216 users (96 individuals and 60 pairs). Our results show 80–90% accuracy in prediction depending on the number of time-steps ahead. In effect, the work reported here provides a framework for evaluating search processes during exploratory search tasks and predicting search performance. Importantly, the proposed approach is based on user processes and is independent of any IR system.

Introduction

Predicting how people perform in their information search processes is a hard problem. The prediction problem becomes even more complex when considering exploratory searches. Exploratory search is typically described as open-ended and multifaceted, where the goals may be unclear and there may be no or multiple satisfactory answers (Kules & Capra, 2009; Marchionini, 2006; Wildemuth & Freund, 2012). Searchers engage in exploratory search when they commence researching a new topic, when they form a new idea, in problem identification, and other forms of information seeking for creative discovery. The searcher shifts out of an exploratory search when they have defined the issues and problems in the new topic area and are now sure of their information need. MacKay (1969) and Taylor (1968) explain this difference in terms of issuing a command to the information system and asking a question of the information system. A command from the user wishes to affect the “goal-setting” levers in the information system, while a question from the user wishes the system to affect the goal-settings, the range of state of readiness, in the user (MacKay, 1969, p. 101; see also, Cole, 2012, p. 20).

Although modern information retrieval (IR) systems aim to support exploratory search, such systems are often unable to perform dynamic and timely predictions of their users' search performance. Exploratory search as a combination of browsing and focused search involves a variety of key factors such as context, intentions, motivations, prior knowledge, feelings, expectations, and strategies, which are often ignored by these IR systems. In the absence of such

Received March 1, 2014; revised November 25, 2014; accepted November 26, 2014

© 2015 ASIS&T • Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23484

information, IR systems often rely on limited data (e.g., pages visited, queries used) and measurements (e.g., precision, recall) to perform evaluations of users' search processes. Therefore, predicting one's success or failure in an exploratory search situation is a challenging task. In this article, we address this challenge using data from several user studies and building statistical models of prediction based on several search features. The outcomes of this work include a new framework for evaluating exploratory search, and a predictive model that can predict with a high accuracy the likelihood of one's search process leading to success. Although capturing and using a comprehensive context of the searcher still remains unaddressed, the current article provides the first step in a new direction that investigates exploratory search performance from a user's side.

The rest of the article is organized as follows. An extensive literature review of the related work to present the relevance and importance of the problem being addressed and the research questions being tackled in this article are presented in the Background section. Based on our review of the literature we identify two research questions (RQs)—one relating to evaluation and the other relating to search features for prediction. Next, our method for running experiments, which included features to extract data from user actions during a search process, is given in the Evaluation Method section (addressing RQ1). This is followed by the details of our experiments in the Experiments section. This section includes a description of the data used, the tasks that generated the data through laboratory studies, feature selection, and the comparisons between the proposed method of performance prediction and the baseline method. This section not only shows the effectiveness of our method in terms of successfully predicting one's performance with a high accuracy, but also presents new measures for computing search effectiveness and efficiency more suitable for process-based evaluations (addressing RQ2). These results, their implications, and impacts on IR along with their limitations, are discussed in the Discussion and Conclusion sections.

Background

Research on predictive models related to evaluating search performance of information-seeking tasks can be grouped into two main categories: behavioral studies and IR approaches. The former focuses on how users of IR systems and services work with the search task at hand by emphasizing their intentions, strategies, and sensemaking (e.g., Kuhlthau, 1991; Dervin, 1998). The latter focuses on IR system performance (i.e., effectiveness, efficiency, precision, and recall), search predictions, results evaluations, and users' interactions with the system. In the end, both of these approaches/perspectives have the same/similar goal of helping a searcher, but they both achieve this in different ways. Furthermore, Jansen and Rieh (2010) have shown the division between information search and information

retrieval mainly in terms of where the emphasis is placed on in each domain. They have shown that "Information searching refers to people's interaction with information retrieval systems, ranging from adopting search strategy to judging the relevance of information retrieved" (p. 1517) whereas "Information retrieval is finding material of an unstructured nature that satisfies an information need from within large collections stored on computers" (p. 1517). In their discussion about information search and information retrieval, "Search is iterative" (p. 1529) is one of the constructs that is highly critical to the information seeking process being addressed in the article, because the notion of iterative process leads to being able to observe, evaluate, and understand the underlying process of search conducted by users.

Because the domain of interest for the work described here is search process and exploratory search, a brief overview covering aspects of user search process analysis and exploratory search tasks that relate to this article is also presented in this section.

Behavioral Modeling Perspective

Different models of information seeking have been introduced in the past few decades with different assumptions, structure, purposes, scope, and intended uses (Case, 2012, p. 135). One such information-seeking model that is widely used and applicable to the context of search process-based analysis is Kuhlthau's (1991) ISP model. (Kuhlthau, 1991). The "Information Search Process" (ISP) model consists of six stages of a student seeking information over the course of a school assignment: Initiation, Selection, Exploration, Focus formulation, Collection, and Presentation. (Note that there is also a seven-stage version of the ISP model that includes an Assessment stage, after the students have presented their finished assignment, when the students assess their learning.) Each of these stages explains a different type of user search behavior and the model is based on changes in affective, cognitive, and physical actions of an information seeker. Kuhlthau's insight into Stage 3, where the student explores information on their selected topic, is particularly pertinent to the broader topic of exploratory search. According to Kuhlthau (1991), Stage 3 is characterized by feelings of confusion, uncertainty, and doubt while investigating information on a general topic to improve personal understanding of that topic. Therefore, actions such as locating information, reading information, and relating new information to existing knowledge are pertinent within the exploration stage of Kuhlthau's ISP model that is also in line with the actions exhibited when conducting exploratory search.

In investigating information behavior for learning conducted on a group of middle school students, Beheshti et al. (2014) found factors that constitute information behavior characteristics in relation to the ISP model such as goals, knowledge of information management, consultation, positive and negative emotions. Shah and González-Ibáñez (2010) were able to identify clear distinctions between

different ISP stages along the search process except for the exploration, formulation, and collections stages where the distinctions were vague when information seeking is conducted in a collaborative setting. They found interesting insights about the relationship between the positive and negative feelings of participants working in collaboration and the level of affective relevance in each of the ISP stages by analyzing the chat messages exchanged while performing the information-seeking tasks.

Behavioral studies have been conducted to explore user models that help anticipate specific aspects of the search process. One goal in this context has been the determination of whether a search process will be completed in single or multiple sessions. For example, Agichtein, White, Dumais, and Bennett (2012) investigated different patterns that can be identified in tasks that require multiple sessions. As a result, the authors devised an algorithm capable of predicting whether users will continue or abandon the task. Similar work is described in Diriye, White, Buscher, and Dumais (2012), which focuses on predicting and understanding why and when users abandon web searches. To address this problem, Diriye et al. (2012) studied features such as queries and interactions with search result pages. Based on this approach, they were able to determine reasons for search abandonment such as accidental causes (e.g., web browser crashing), satisfaction levels, and query suggestions. Some researchers have also focused on search engine switching detection, which involves determining when and why searchers decide to move from one search engine to another. In this regard, Savenkov et al. (2013) presented an approach that predicts search engine switching based on pattern analyses of search logs.

There have also been attempts to understand past users' behaviors in order to predict future actions in similar conditions. For example, Adar, Weld, Bershad, and Gribble (2007) explored behavioral aspects using large-scale data sets containing queries and other information objects produced by users. The authors were able to identify different behavioral patterns that seem to appear consistently within different data sets. Although not directly related to performance prediction, this work focused on attributes of the search process instead of the final products derived from it.

Research such as this often relies on historic data from large populations and the applications of trend and seasonal components, which are used to model long-term direction and periodicity patterns of time-series (Ord, Hyndman, Koehler, & Snyder, 2008). For example, researchers have explored the significance of seasonal aspects in web search (e.g., weekly, monthly, or annual behaviors), which provides useful information to predict and suggest queries (Dignum et al., 2010).

Exploratory Search

According to White and Roth (2009), exploratory search is defined as "a type of information seeking and a type of sense-making focused on the gathering and use of

information to foster intellectual development" (p. 38). Cole (2011) argued that the theory of information need goes beyond the simple answer retrieval for well-defined questions in IR to show how Taylor's model of information need (Taylor, 1968) based on the unspecifiable and unknowable query definition by the user for knowledge acquisition from an information system considers the complexity of the information need. This gives rise to the importance of type of search such as exploratory search where the user conducts the information search without knowing the answer or the form of answer he will receive from the system.

Five aspects that distinguish exploratory search from other types of search tasks are: uncertainty, ambiguity, level of specificity, and the need for discovery and learning from multiple sources (Kules & Capra, 2009; Marchionini, 2006; Wildemuth & Freund, 2012). Another distinguishable aspect of these types of tasks is that they are open-ended (Wildemuth & Freund, 2012). Despite the complexities associated with the design of exploratory search tasks for research purposes, these tasks are commonly used by researchers in information-related fields (Kules & Capra, 2009). Uncertainty associated with information seeking has been extensively studied by Wilson, Ford, Ellis, Foster, and Spink (2002). They found that there exists significance in the relationship between the level of uncertainty and the problem stage that the user is engaged in during the information-seeking task and between level of uncertainty and the domain knowledge possessed by the user at exploration. This theory of information need ties in well with the exploratory search tasks where exploration is a major phase in the information-seeking process and thus the importance of uncertainty in the user's mind has to be considered. Borlund and Dreier (2014) performed an empirical analysis of how three types of information needs (verification, conscious topical, muddled topical) are related to the information-searching behavior characteristics by Ingwersen (1986). They found that muddled topical information need was mostly exhibited in exploratory search where the participants demonstrated highest average number of query modifications (compared to other two information needs), spent longer time on searching depicting the difficulty in expressing the specific information need and finding relevant information exhibiting exploratory search behavior.

Exploratory search evaluation has mainly focused on system-based assessment comprised of evaluation measures such as system utility. This measures how successful the system was in facilitating the key elements of search exploration (e.g., helping users obtain new insights, assisting learning, offering support for critical decision making), depth and effectiveness of learning, recall (Tunkelang, 2009), level of subject-matter learning as a function of exploration time (Rao & Card, 1994), information novelty and subjective measures such as engagement, enjoyment, and task success (White, Kules, Drucker, & Schraefel, 2006a).

There have been fewer attempts to evaluate exploratory search from the user perspective. Most research up until this

point has been from on a systems perspective. Spink (2002) proposed a user-centered approach for the evaluation of exploratory search that used a new search tool. This new tool highlighted aspects of effectiveness and usability. Spink's analysis was based on relevance judgments, users' pre- and postsearch questionnaires, and search transaction logs using a qualitative approach.

In addition to the challenges experienced by searchers when working on exploratory search tasks, researchers and developers face other problems. According to Kules and Capra (2009) evaluation is considered one of the major challenges. In particular, the evaluation of the systems that assist searchers when working on these types of tasks (White, Muresan, & Marchionini, 2006b) in addition to the evaluation of how searchers perform these types of tasks are particularly difficult (Shah & González-Ibáñez, 2011). Some approaches in this regard include the use of traditional information retrieval measures such as precision and recall, and non-traditional measures such as unique coverage and query diversity (Shah & González-Ibáñez, 2011).

IR Perspective

From an IR perspective, Radinsky et al. (2012) explored models to predict users' behaviors in a population of web searchers in order to improve results retrieved by IR systems. The authors also developed a learning algorithm capable of selecting an appropriate predictive model depending on the situation and time. As described by the authors, applications of this approach could go from click predictions to query-URL predictions. In contrast to this approach, our method presented considers both the population trends and an individual user's behavior.

On a similar track, several works have been conducted on query performance prediction, focusing on developing techniques that help IR systems anticipate whether a query will be effective or not in providing results that satisfy users' needs (Cronen-Townsend, Zhou, & Croft, 2002; Guo, White, Dumais, Wang, & Anderson, 2010; He & Ounis, 2006). For example, Guo et al. (2010) found that features derived from search results and interactions offer better prediction results than a prediction baseline defined in terms of query features. These results have direct implications for individual users by aiding the auto evaluation process of IR systems.

In information search, users may be unaware of their individual performance when solving an information search task. For instance, Shah and Marchionini (2010) showed how lack of awareness about different objects involved in searching (queries, visited pages, bookmarks) could result in mistaken perceptions about search performance during an exploratory search task. Even if an IR system is highly effective, users may adopt techniques such as query reformulation, query expansion and query abandonment and evaluate several search paths before finding what they need. This process, which can be related to search strategies, implies that the effort and time required is usually

underestimated by the users themselves. In this sense, instead of predicting end products (i.e., overall performance), the approach we introduce in this article is oriented toward predictions at different time intervals during the search tasks in order to increase the user's level of awareness about their own search process. Similar to weather forecasts, this information could help users become aware of possible trends based on past and current behaviors.

A detailed discussion about IR evaluations and their shortcomings is beyond the scope of this article, but the reader is referred to a recent article by Järvelin (2012) for a comprehensive treatment of this subject. To the best of our knowledge, search process performance prediction at different time intervals from a user perspective has not been explored. Previous work has emphasized exploiting temporal information embedded in documents for presentation, clustering, and exploration of search results (Alonso, Gertz, & Baeza-Yates, 2007) where the temporal aspect is used on the documents or at the information objects level but not at the search process level.

Another recent area of focus from an IR perspective is the problem of effective search trend prediction. By analyzing the query logs of search engines, researchers seek to accurately predict the most trending items on the web. This feature has been actively implemented in almost all of the current popular search engines such as Google Trends¹, Yahoo!, Trending Now², and Bing Trends³. Golbandi, Katzir, Koren, and Lempel (2013) provided a method to search trend prediction more accurately with low latency relative to the actual event that sparked the trend. Similar approaches can be found in weather and Stock Market studies. For example, using machine-learning approaches such as support vector machine (SVM), several models have been implemented to predict the trends of two different daily stock price indices using NASDAQ and Korean Stock prices (Kyoung-jae, 2003; Ming-Chi, 2009). In a similar fashion, our approach is oriented to forecast a user's search performance *n*-steps-ahead with the aim to aid their search process awareness and performance trends.

Modeling user search behaviors when conducting web searches has also been a major topic of interest among researchers. Adar et al. (2007) compared users' search and blogging behavior across multiple search engines by modeling the time-varying data using dynamic time warping and correlation analysis. They also created an interface for analyzing vast amounts of search log data through the visualization and summarization of the differences between time series. One could then analyze user search behavior patterns and thereby make predictions. Liao, Song, He, and Huang (2012) conducted web log data analyses to evaluate the effectiveness of task trails in three different search applications related to user satisfaction, user search interest prediction, and query suggestion. In order to find search goal

¹<http://www.google.com/trends/>

²<http://news.yahoo.com/blogs/trending-now/>

³<http://www.bing.com/blogs/>

success, Hassan, Jones, and Klinkner (2010) have proposed modeling user search behavior using a Markov model with transition times based on user search process-based features. Such work has shown empirically that models for predicting user search goal success are significantly more accurate than traditional relevance-based models. Zhang, Chen, Wang, and Yang (2011) proposed a new task-centric click model called TCM, which aims to characterize the user click behavior related to a task as a whole by using an expectation-maximization based learning algorithm. They were able to show that their proposed model provided better insights into user search behavior and prediction than the earlier probabilistic models like dynamic bayesian networks or user browsing models.

Research Questions

Based on the literature review we conducted on exploratory search and IR perspective, we found that exploratory search evaluation from a user perspective has not been studied adequately. The search process that the users undergo during information seeking is given less focus in past research and more emphasis is placed on the outcomes of the search process such as queries, search result pages, and content. It is clear to us that measuring exploratory search from a user's perspective by focusing on a users' processes, independent of the system(s) employed, is not addressed sufficiently in the literature. Therefore, in this article, we approach the exploratory search evaluation from a user perspective and evaluate the search performance across time and along the user's exploratory search process, taking a purely quantitative approach.

Furthermore, from the user search behavior analysis literature, we understood that there exists different information-seeking models and successful attempts to predict useful information objects (such as web pages, queries) based on the past user search behavior analysis. One important aspect of user behavior analysis that has not been extensively studied to the best of our knowledge is dynamically evaluating the performance of the user search process and thereby making predictions of how the users are going to perform and achieve their search goal. Thus, we believe that incorporating the different observed user search behavior related features (based on actions) to measure the effectiveness of the user search process could lead to much richer predictions to assist the user to improve their search performance.

Therefore, we began our investigation by asking how we could leverage the valuable lessons obtained from both IR evaluations to not only evaluate an ongoing search process, but also predict how well it will unfold and suggest a better path to the searcher if he is likely to underperform. The motivation behind this investigation was based on the following assumptions and realizations grounded in the literature as well as findings from our previous work.

1. The underlying rational processes involved in information search are reflected in the actions users make while

searching. These actions include entering and reformulating search queries, skimming the results, as well as selecting and collecting useful information (Gwizdka, 2008; Liu, Gwizdka, Liu, Xu, & Belkin, 2010a; Liu, Gwizdka, Liu, & Belkin, 2010b).

2. A searcher's performance is a function of these actions performed during a search episode (González-Ibáñez, Haseki, & Shah, 2012; Shah & González-Ibáñez, 2011).

With these assumptions, we propose to quantify a user's exploratory search process by tracking various features calculated around two dimensions that are pivotal in exploratory search behavior analysis according to the past research explained in the Background section: (a) expression of information need, and (b) coverage of information, and then use these indicators for user performance (henceforth, "search performance" or "performance") prediction. Specific research questions (RQs) constructed from this perspective are listed below.

- RQ1. How can we measure search performance as a function of search-related user actions and products?
- RQ2. What features, if any, can help predict the performance of a user at different times along the search process?

To address these RQs, we used data from three different user studies in which the participants were asked to work on exploratory search tasks. Specifically, we consolidated various evaluation measures from the literature and our own past work for addressing RQ1 (see section Evaluation Method), and used various search features to build a predictive model for addressing RQ2 (see section Experiments).

Uniqueness of Our Approach

Unlike previous works in IR that uses seasonal components of historic data to predict future trends, we use time series of user-specific search process-based features to predict the future performance of users. Moreover, we investigate predictive models based on the machine learning (ML) classification technique, in particular SVM trained over a set of features such as number of queries, and page dwell time in comparison to an IR based baseline method.

In contrast to most IR evaluations, our method focuses on the user process. Unlike most recommender systems, our approach could also output alternative strategies instead of similar and relevant products to help the searcher. In essence, the work reported here takes several lessons from traditional IR evaluations, recommender system designs, and weather and stock forecasting to come up with a new approach for evaluating and predicting search performance.

The method described in this article, especially the way evaluation is performed, incorporates many of the suggestions for setting up exploratory search tasks and measuring various quantities. The differences, however, are found in the way team-related measures are employed as opposed to individual-based measures. Some of these differences and

the rationale behind them are discussed in the following subsection.

In the next section we provide a detailed description of our evaluation method, feature selection, and the measures we used in order to create the ML-based predictive model.

Evaluation Method

There are several ways to evaluate the *search performance* of a user performing a search task. Most of these look at the objects retrieved as results of the search process (Saracevic, 1995). For instance, it is very common in IR to use measures such as precision, recall, mean average precision (MAP), and non-discounted cumulative gain (NDCG), which are all based on analyzing the relevance of documents retrieved. Similarly, recommender systems analyze objects (documents, movies, books, music) found to be relevant by the user (content filtering) or the community (collaborative filtering) to suggest objects to a user's current search process (Adomavicius & Tuzhilin, 2005; Ricci, Rokach, Shapira, & Kantor, 2011). Instead of relying on ranking and relevance of retrieved objects (which will serve as the baselines), we focus on user processes for our evaluation. In order to analyze the search processes followed by different users or teams, we assume that the underlying dynamics of the search processes are expressed by a collection of activities that take place from the beginning to the end of the search processes.

The first part of our method is a feature extraction step in which we obtain a wide array of features relating to two distinct sets of activities that guide and define an exploratory search: (a) expression of information need, and (b) coverage of information. Once again, we note that although *information* is treated as an artifact (e.g., Buckland, 1991), we are considering the representation and coverage of information, which are indicative of the underlying user behaviors as assumed in the Introduction of this article. Features that represent these two fundamental activities are extracted for each unit of time, t . This step is performed in order to evaluate how well we could use those features to capture the underlying dynamics that would lead to recognizing the level of performance at the future time steps at $t + n$, $n = 1, 2, \dots, N$, where N is the furthest time step.

The decision to include or exclude a feature was based on literature as well as our past experience with representing and evaluating search objects and processes (e.g., Borlund & Dreier, 2014; Ingwersen, 1986; González-Ibáñez, Shah, & White, 2012; Shah & González-Ibáñez, 2011; Shah, Liu, González-Ibáñez, & Belkin, 2012). Each feature is extracted for each user or team, u up to time t from the search processes. They are explained in detail as follows.

Expression of Information Need

- *Number of Queries* (u, t) – $\mathbf{NQ}_{u,t}$: Total number of distinct queries executed by a user up to time t . This feature implicitly relates to how much effort and cognitive thinking a user has put in to this task.

- *Length of Query* (u, q, t) – $\mathbf{LQ}_{u,q,t}$: Length of each query, q executed by a user, u based on the character count of the query up to time t . This feature captures how the user constructed the queries and how long they were at different times of the search process.
- *Number of Tokens in each Query* (u, q, t) – $\mathbf{LTQ}_{u,q,t}$: This is the count of tokens/words in each query, q executed by user u up to time t . This query-based measure takes into account how specific a user was in defining the query. By inspecting log data from various data sets, we realized that queries with a fewer number of tokens tend to get general results. On the other hand, composed queries with multiple terms are related to more specific searchers. We also observed that typically the users started with general queries with few words at the beginning of the search process but then went into more detailed queries to find more specific information later. For all these reasons, we found it to be useful to capture the number of tokens used in a query.
- *Query Entropy* (u, q, t) – $\mathbf{QN}_{u,q,t}$: This measures the information content in a given query, q by finding the expected amount of information contained in a query. We used the widely recognized notion of Shannon entropy (Shannon & Weaver, 1963) in information theory to calculate the information content of a query. We calculated the number of unigrams appearing in each of the queries, q for each user, u up to time t , which represent the observed counts of the random variable as shown below. This was used as the input to Shannon entropy calculation and we used to the maximum-likelihood method to calculate the entropy. Query entropy feature has been used in the past to predict *goodness* of a query for making a query expansion decision (Lv & Zhai, 2009; Shah & Croft, 2004).

$$\mathbf{QN}_{u,q,t} = \sum_{y=1}^{n\{\text{unigrams}_{Qu,q,t}\}} p_y \log_2 p_y \quad (1)$$

Coverage of Information

- *Total Coverage* (u, t) – $\mathbf{TC}_{u,t}$: The total number of distinct web pages visited by a user, u up to time, t . This feature captures the web page-based activity performed by a user and provides a measure to see how much distinct information has been found by the user up to this time.

$$\mathbf{TC}_{u,t} = \{wp_1, wp_2, \dots, wp_i\}_{u,t} \quad (2)$$

- *Useful Coverage* (u, t) – $\mathbf{UC}_{u,t}$: The total number of distinct web pages in which a user spent at least 30 seconds, up to time t . This measure evaluates out of the total pages the user visited how many of them were useful in finding relevant information leading to satisfaction with their context in completing the exploratory search task. This implicit measure based on dwell time on a web page has been empirically proven to provide a good indication of usefulness (Fox, Karnawat, Mydland, Dumais, & White, 2005; White & Huang, 2010).
- *Number of Saved Snippets* (u, t) – $\mathbf{NS}_{u,t}$: Total number of snippets saved by user u up to time t . This measures the amount of information the user retained for its possible future relevance to completing the task. This feature is an indication of explicit relevance judgments made by the user.

The method used to assess the search performance of a user is described below. First, we define Coverage Effectiveness of user, u up to time t , $CE_{u,t}$ as the ratio of useful coverage and total coverage (both defined earlier). A similar measure was used by González-Ibáñez et al. (2012) and also by Shah and González-Ibáñez (2011).

$$CE_{u,t} = \frac{UC_{u,t}}{TC_{u,t}} \quad (3)$$

We define a measure called Query Effectiveness (u,t), $QE_{u,t}$ for each user, u up to time t in order to predict the level of performance yielded by the search process in the future.

$$QE_{u,t} = \frac{CE_{u,t}}{NQ_{u,t}} \quad (4)$$

Therefore, Query Effectiveness is defined as the Coverage Effectiveness obtained per query, or how effective a query is in terms of achieving a certain level of useful coverage. Query Effectiveness, in turn, represents success of the underlying search process. The proposal of this measure is also a reflection and reinforcement of the two assumptions listed in the Introduction section in that the QE shows the effectiveness of one's search strategy.

The performance for each user, u at each time t was classified into the four classes based on the following criteria:

$$Class_{u,t} = \begin{cases} 1 & \text{if } QE_{u,t} \leq \text{first quartile}(QE_{u,t}) \\ 2 & \text{else if } QE_{u,t} \leq \text{second quartile}(QE_{u,t}) \\ 3 & \text{else if } QE_{u,t} \leq \text{third quartile}(QE_{u,t}) \\ 4 & \text{else} \end{cases} \quad (5)$$

We then constructed the feature matrices, which consist of all aforementioned features for each minute of time t for all the users in each data set. This was converted into a long vector of features, which we then fed as the input to the classification models used. The class labels were generated as 1–4 at minute $t + n$ based on the above mentioned criteria and used as the output class labels to be used in the n -steps-ahead prediction model. If a class label at n -steps-ahead was correctly predicted based on the features extracted up to time t from the classification model, it was considered as correctly classified and, if not, as misclassified.

Baseline Evaluation

Furthermore, to construct a baseline for comparison purposes, we calculated IR evaluation based precision and recall measures (Kent, Berry, Luehrs, & Perry, 1955) considering time-based evaluation for each user u for time t as described below.

$$Coverage_{u,t} = \{wp_1, wp_2, \dots, wp_i\}_{u,t} \quad (6)$$

In Equation (6) wp_i denotes the distinct web pages visited by user u until time t .

The universe of distinct web pages visited by all users up to time t are considered as the set of documents covered up to time t as defined in Equation (7).

$$Coverage_t = \bigcup_u Coverage_{u,t} \quad (7)$$

The universe of relevant web pages visited by all users/teams up to time t are denoted in Equation (8) where $RelevantCoverage_{u,t}$ means the is the set of web pages that user u visited and collected snippets from until time t .

$$U_r = \bigcup_u RelevantCoverage_{u,t} \quad (8)$$

Based on equations (7) and (8), we calculated the precision and recall measures for each user u until each time t as will be defined here in Equations (9) and (10).

$$Precision_{u,t} = \frac{|RelevantCoverage_{u,t}|}{|Coverage_t|} \quad (9)$$

$$Recall_{u,t} = \frac{|RelevantCoverage_{u,t}|}{|U_r|} \quad (10)$$

Experiments

In this section, we present the details about the data, experiments, and a discussion about the results obtained.

Data

The experiments were based on log data collected from three user studies. The three data sets contain search sessions logs of: (a) single users performing an exploratory search task about the Gulf oil spill (Shah et al., 2012), (b) single users performing an exploratory search task about three different revolutions in history (Shah et al., 2012), and (c) collaborative pairs performing an exploratory search task about the Gulf oil spill (Shah & González-Ibáñez, 2011, 2012). The following are the descriptions of these two tasks taken from Shah and González-Ibáñez (2012), Shah and González-Ibáñez (2011), and Shah et al. (2012). All the participants in these studies used a version of the Coagmento system (Shah, 2010).

The tasks used in the three user studies spanning two topics used in this analysis are exploratory in nature. The search tasks assume that the underlying topics are of general interest, however, the participants in the study have no specific knowledge. This aligns with White et al.'s (2006a) investigation of exploratory search. Their research focused on answering the question, "what if we want to find something from a domain where we have a general interest but not specific knowledge?" (p.37). Our first task focuses on the Gulf oil spill of 2010, an event that grabbed the attention of people around the world. The second task is about different European revolutions. In both tasks the participants are

provided with specific objectives, for example, finding out the causes and consequences of the oil spill. Despite the existence of well-defined goals, searchers need to explore several sources before providing an answer.

Oil spill task description “A leading newspaper has hired you to create a comprehensive report on the causes, effects, and consequences of the recent gulf oil spill. As a part of your contract, you are required to collect all the relevant information from any available online sources that you can find.

To prepare this report, search and visit any website that you want and look for specific aspects as given in the guideline below. As you find useful information, highlight and save relevant snippets. Make sure you also rate a snippet to help you in ranking them based on their quality and usefulness. Later, you can use these snippets to compile your report, no longer than 200 lines in length, as instructed.

Your report on this topic should address the following issues: description of how the oil spill took place, reactions by BP as well as various government and other agencies, impact on economy and life (people and animals) in the gulf, attempts to fix the leaking well and to clean the waters, long-term implications and lessons learned.”

Revolutions task description “As a history buff, you have heard of the quiet revolution, the peaceful revolution, and the velvet revolution. For a skill-testing question to win an iPod you have been asked how they differ from the April 19th revolution.

Search and visit any website that helps you find information on this topic. As you find useful information, highlight and save relevant snippets. Make sure you also rate a snippet to help you in ranking them based on their quality and usefulness.”

Logs for each user and team correspond to approximately twenty minutes of active search. The log data consist of the time stamped records of all the user activities given here.

- User ID / Team ID (only for team data)
- Timestamp
- Action (query, visited URL, saved snippets)

We analyzed the team data of 60 teams of collaborative user pairs (total 120 users). Because the tasks were exploratory and divisible in nature, the team members (pairs) were able to divide up the work and search independently. Therefore, we also considered the team data for independent user analysis, treating each team member as an individual user completing part of the task. Table 1 summarizes the number of users/teams and number of outliers used in this analysis. Note that there were two outliers and missing data-points because of data logging errors that we had to exclude from these data sets.

Experimental Details

It was observed from the data analysis that most of the users and teams completed the exploratory search task

TABLE 1. Data sets from various user studies used for experiments.

Data set	Acronym	Number of users or teams	Number of outliers
Single Users: Oil Spill	Single_OS	59	1
Single Users: Revolution	Single_Rev	37	0
Teams: Oil Spill	Team_OS	60	0
Team Members as Independent users: Oil spill	TeamInd_OS	120	1

assigned to them within the first 20 minutes and spent a few more minutes writing down the answers and formalizing the final output, which we considered as unrelated to the search process but only pertinent to the completion of the task. Therefore, we consider the first 20 minutes of the task as part of the active search process for the experiments presented here. The unit of time in the analysis was minutes and the overall analysis spanned from minute 1 to minute 20 for all the users and teams. The n -steps-ahead prediction method described in the Evaluation Method section was executed as follows.

We calculated the feature values for all the seven features, which were explained in the features of coverage information and information need, for each minute ($t = 1, 2, \dots, 20$) and for each user/team, u . A two-dimensional feature matrix (with dimensions $7 \times t$) was constructed for each user/team u corresponding to feature values of each time point (each minute). In order to perform further classification-based analysis, we transformed these two-dimensional matrices to one-dimensional vectors representing features corresponding to each user. Then we converted this matrix to get a vector based on the means (averages) of each feature up to time t . All of these vectors for all users were then combined to generate the total feature matrix for all users up to time t .

Equation $FeatureM_{u,t}$ (11) represents the total feature matrix for each user u up to time t and Equation $FeatureV_{u,t}$ (12) represents the corresponding feature vector for each user u up to time t constructed by using the means of the time based features.

$$FeatureM_{u,t} = \begin{bmatrix} f1_{u,1:t} \\ f2_{u,1:t} \\ \vdots \\ f7_{u,1:t} \end{bmatrix}$$

where

$$fn_{u,1:t} = [fn_{u,1}, fn_{u,2}, \dots, fn_{u,t}] \quad (11)$$

$$FeatureV_{u,t} = [\text{mean}(f1_{u,1:t}), \text{mean}(f2_{u,1:t}), \dots, \text{mean}(f7_{u,1:t})] \quad (12)$$

Using this formulation, we evaluated performance of the search process of a user u at each time point t based on the

measure of $QE_{u,t}$, which considered how much useful coverage could be established by incurring a minimal amount of cost or effort, which is linked implicitly to the distinct number of queries attempted by a user up to that time. We categorized the feature matrix obtained into four categories based on quartiles of distribution. Then, we performed n -steps-ahead prediction ($n = 1, 2, 3, 4, 5, 10, 15$) based on the generated feature matrix up to time t to evaluate whether we could predict in which quartile a user's or team's performance would fall.

Classification was performed using support vector machines (SVM) with kernel function as radial basis function (RBF) with default parameter settings using *svm* function available in the *e1071* library in R. In all cases the results reported are based on 10-fold cross validation.

To compare our proposed approach we used the baselines evaluation measures, precision and recall as defined in the subsection, Baseline Evaluation evaluation in the Evaluation Method section. By performing this comparison, we are analyzing what we could gain by acknowledging features

TABLE 2. Classification accuracies for different data sets at different time and prediction steps. Results with significance at $p < 0.05$ is indicated by *, $p < 0.01$ with **, $p < 0.005$ with ***, and $p < 0.001$ with ****.

Data set	Time $t = 3$ mins, $n = 3$ -steps-ahead		Time $t = 10$ mins, $n = 5$ -steps-ahead		Time $t = 10$ mins, $n = 10$ -steps-ahead	
	Baseline	Proposed	Baseline	Proposed	Baseline	Proposed
Single_OS	45.0%	64.0%****	48.0%	72.0%***	63.0%	73.0%*
Single_Rev	25.0%	32.0%*	40.0%	78.0%***	45.0%	65.0%*
Teams_OS	50.0%	87.5%****	72.0%	83.0%*	48.0%	78.0%****
TeamsInd_OS	45.0%	64.0%****	30.0%	78.0%****	44.0%	60.0%*

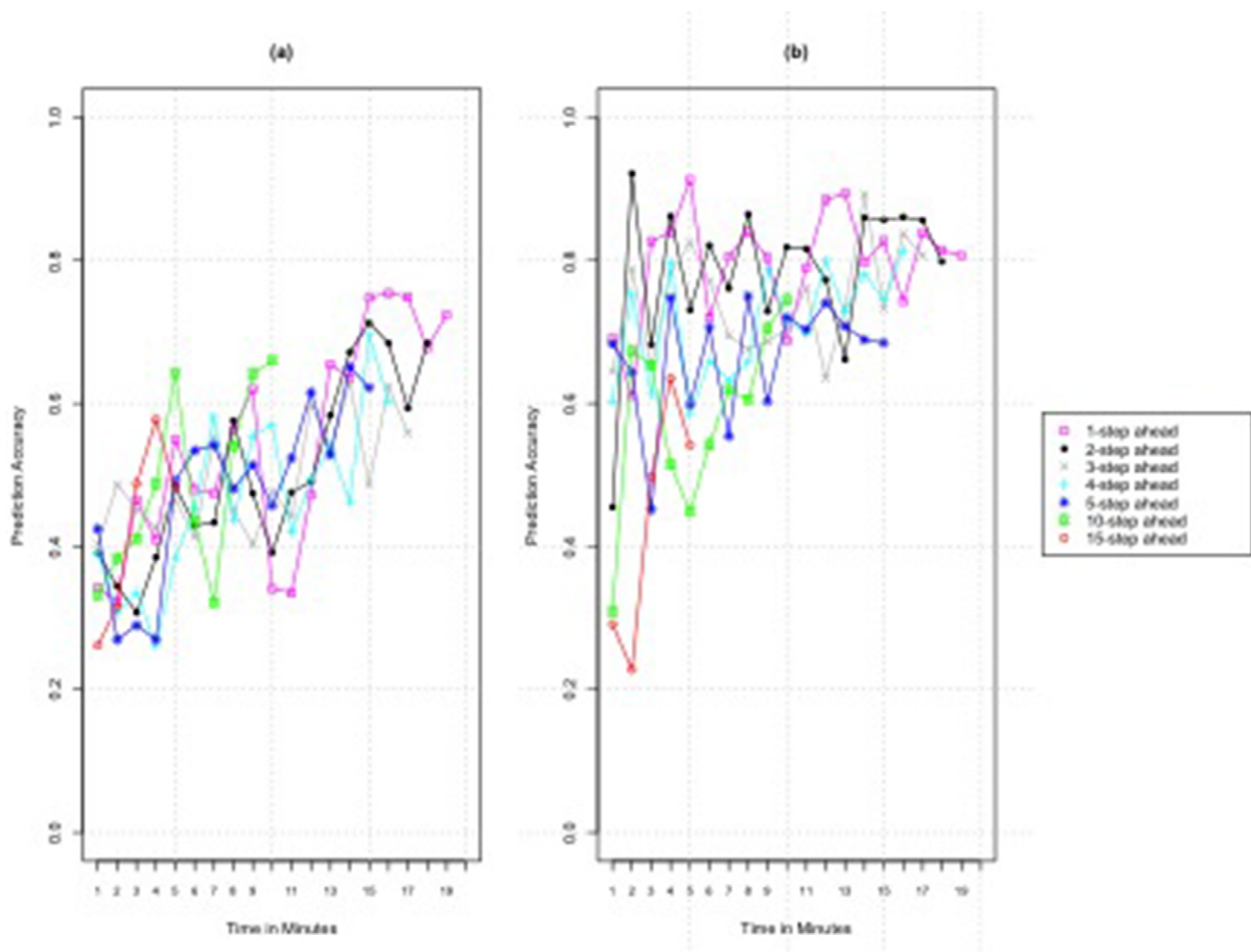


FIG. 1. Prediction accuracy comparison for Single Users: Oil Spill (a) Using precision and recall features, (b) Using proposed Search process based features. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

that consider user actions with respect to their expression about information seeking (queries) and information exposure (coverage) instead of the features that only evaluate retrieval performance.

Feature Selection

In order to identify features that were not significant in distinguishing between the four classes, we conducted a simple feature selection procedure as described by Weiss and Indurkha (1998). This procedure is known as *independent significance features test*. After calculating the statistical significance of each independent feature in predicting the corresponding class label (according to the four quartiles based performance), the features, which had a significance value below the mean of the overall significance, were considered as less important. The features with higher significance were considered of high importance.

In most cases toward the end of the sessions ($t = 10$ to 20 minutes), it was observed that most of the features were

highly significant except for Number of Snippets and Query Length, which had lower significance scores. In predicting at earlier times ($t = 1$ to 3 minutes), features such as Number of Snippets, and all query-related features such as Number of Queries, Query Length, Query Entropy, and Number of Tokens in Query, had lower significance values. It can be presumed that at the beginning of a session the number of queries used or snippets saved by a user are very low compared to later in the session because the user may still be in the *Initiation* or *Selection* stages of their search processes as described by Kuhlthau (1991) and may not be ready to start formulating meaningful queries or collect useful information.

After performing this feature selection, we conducted the same n -steps-ahead prediction tasks for all data sets and observed that the classification results were only varied slightly at the beginning of the session because of lower impact from query-related features mentioned earlier, leading to overall prediction accuracies that were not significantly changed.

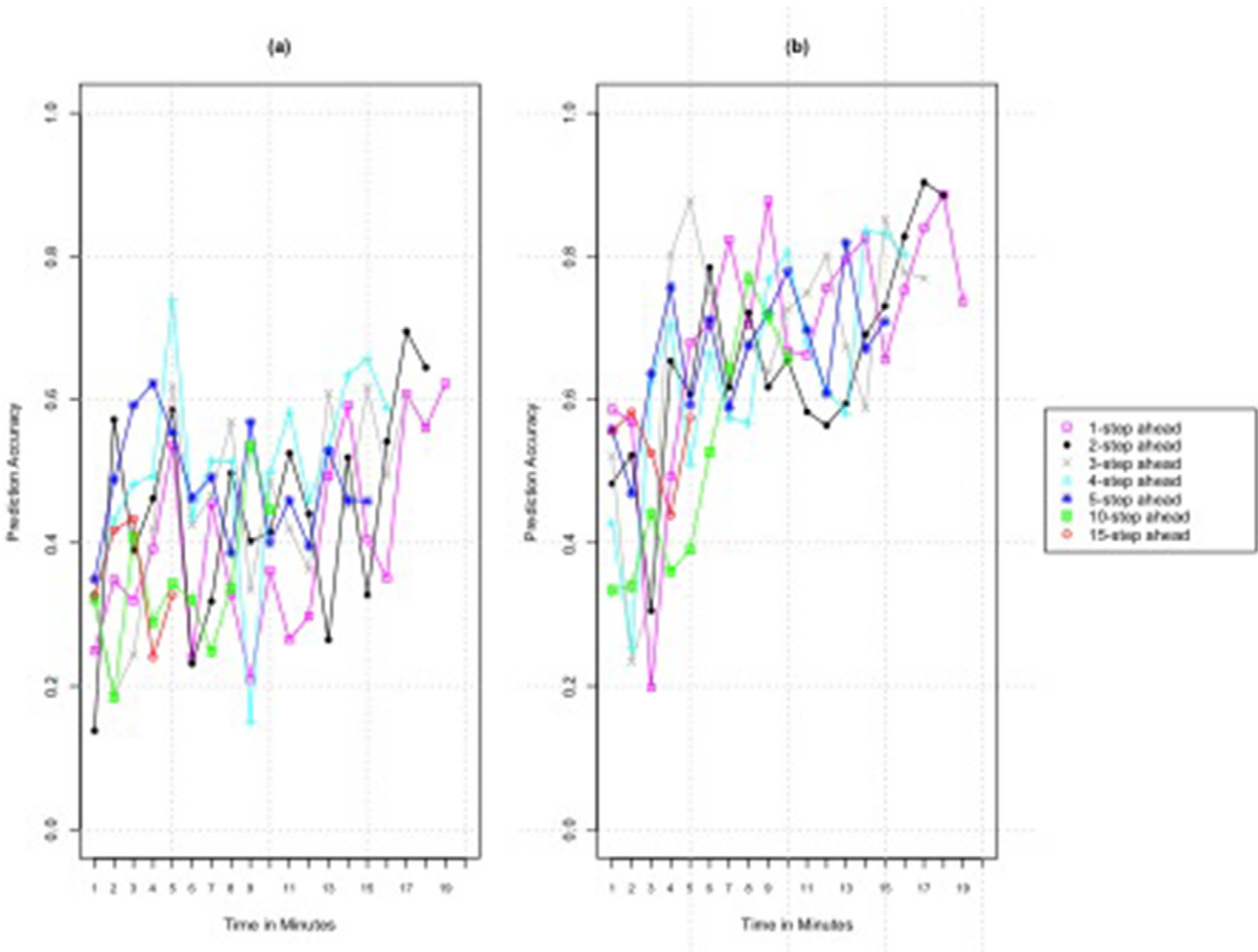


FIG. 2. Prediction accuracy comparison for Single Users: Revolution (a) Using precision and recall features, (b) Using proposed Search process based features. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

The classification accuracies for some of the time points along the prediction steps for the different data sets comparing the baseline method (with precision and recall as features) and features using the proposed method are shown in Table 2. The stars above the proposed methods classification accuracy depict the significance of p -value at significance levels ranging from 95% to 99% for independent samples t -test conducted for classification accuracies generated by baseline and proposed method over the 10-folds. The two separate sets of independent samples for the independent t -test comprised of prediction accuracies at each fold of the test data for the baseline method and the proposed method, respectively. This enabled us to identify whether the means of the classification accuracies for the test data would be significantly different when comparing the baseline method and the proposed method.

Prediction Results

The prediction accuracies using the proposed search process-based features and the baseline features (precision

and recall) for Single Users: Oil Spill are shown in Figure 1a and b, respectively. The results reflect the findings reported in Table 2, that is, using the proposed process-based method for prediction n -steps-ahead provides higher accuracy than the prediction based on precision and recall.

Prediction accuracies using the proposed search process-based features and the traditional IR-based features (precision and recall) for Single Users: Revolution is shown in Figure 2a and b, respectively.

For teams and independent members of teams, the prediction with proposed search process features beats the baseline as shown in Figures 3 and 4.

Predicting Precision/Recall Using Proposed Features

The results prompted an additional research question: what would happen if we use our proposed features to predict the performance based on precision and recall. To answer that, we also analyzed how the prediction would vary if we try to anticipate the precision and recall based class labels using the proposed search process-based features.

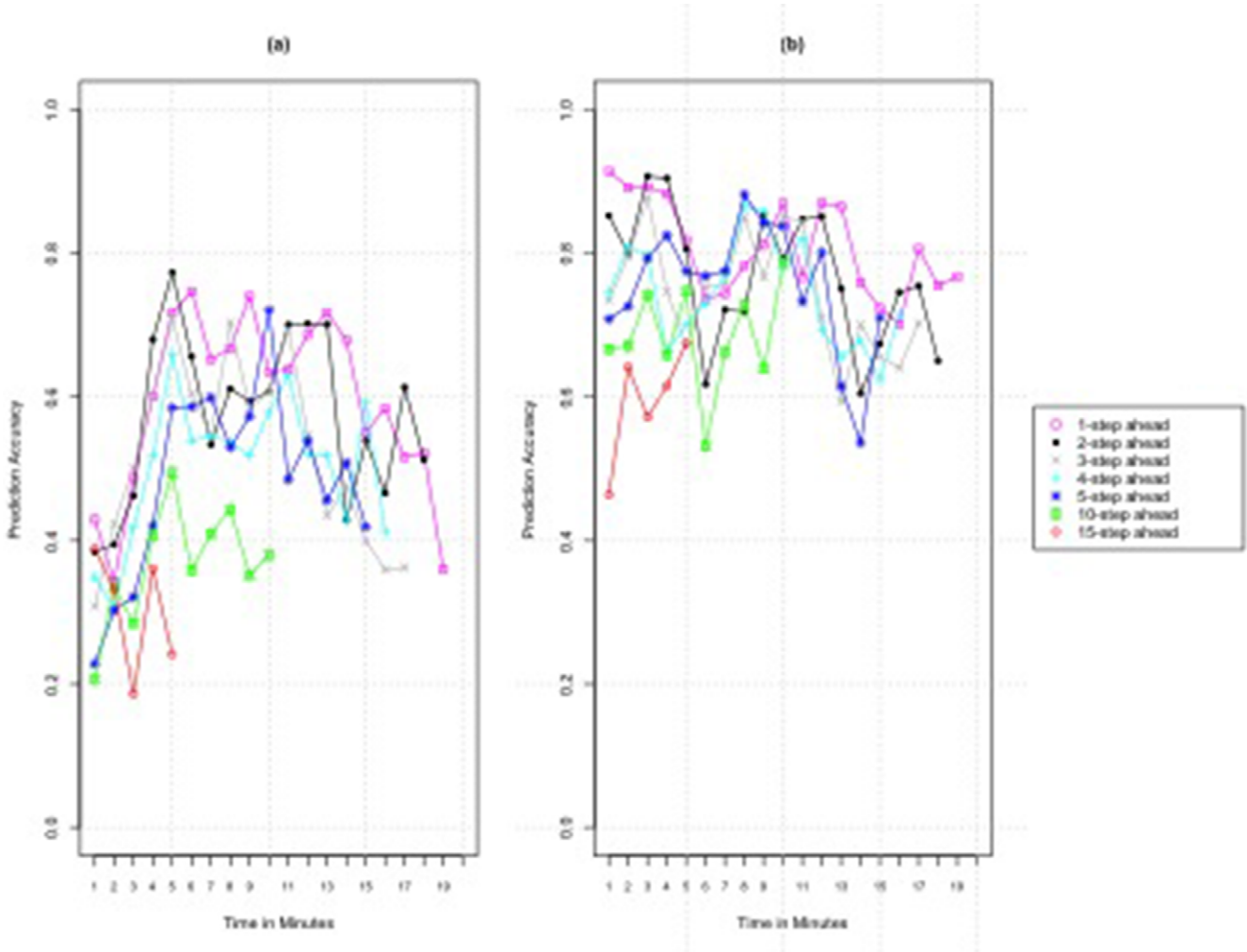


FIG. 3. Prediction accuracy comparison for Teams: Oil Spill (a) Using precision and recall features, (b) Using proposed Search process based features. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

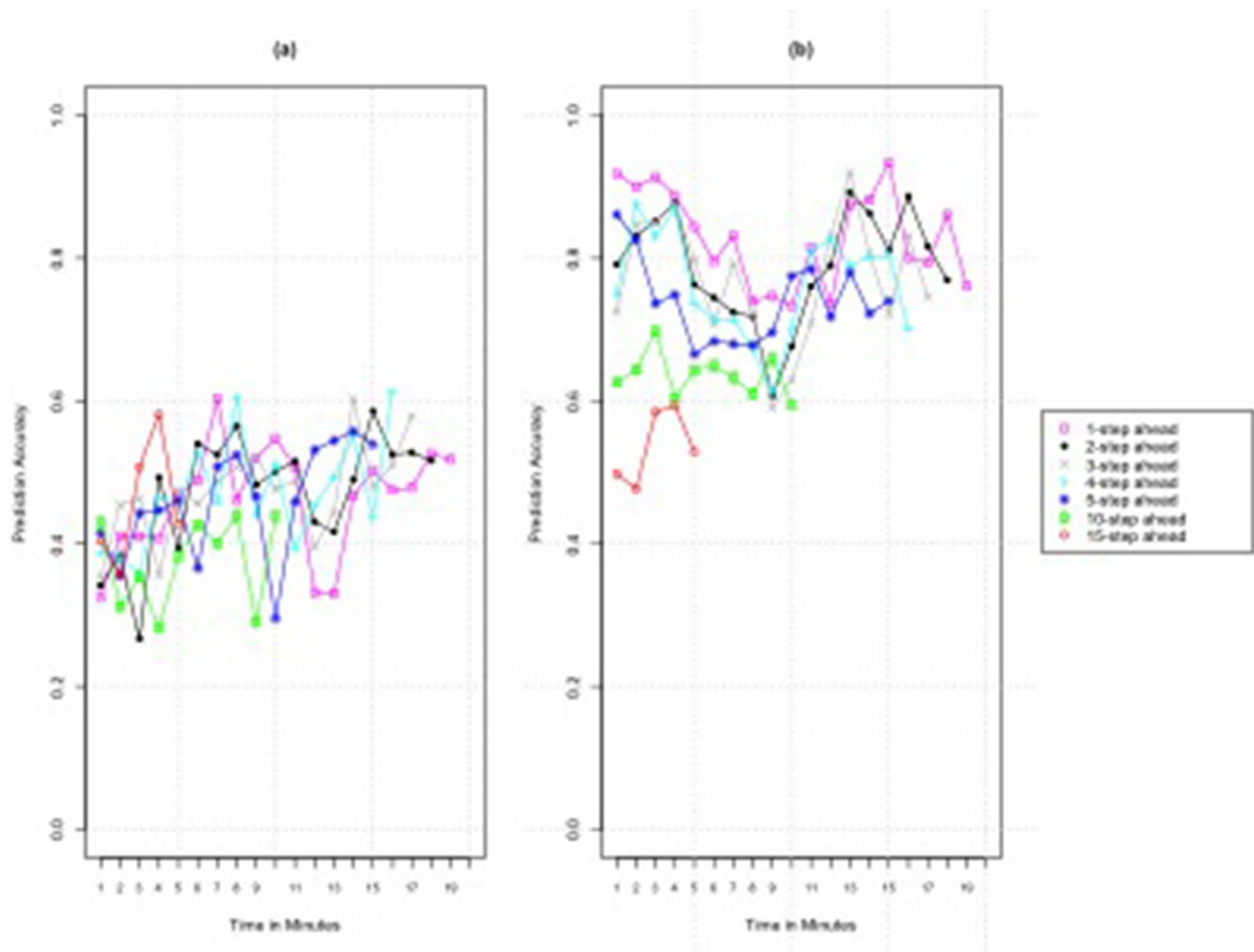


FIG. 4. Prediction accuracy comparison for Team Members as Independent Users: Oil Spill (a) Using precision and recall features, (b) Using proposed Search process based features. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figures 5–8 show the prediction accuracies with class labels based on precision and recall for the four data sets used in this analysis, respectively. Even with such analysis we can see that we are able to achieve above 70% of accuracy in most n -steps-ahead predictions—still a lot better than using traditional IR features for prediction, thus demonstrating the robustness of our features.

Discussion

When it comes to prediction, information retrieval and filtering systems are primarily focused on objects in conjunction with assessing what and whether something could assist the users. These approaches are often system-dependent even though the process of information seeking is usually user-driven. Personalization and recommendation are frequently employed as methods to address user-specific IR and filtering, but still limited to comparing and recommending objects without focusing on underlying IR processes carried out by the searchers. We developed a new

approach to address these shortcomings. We began by asking how we could model a user's search process based on the actions he is performing during an exploratory search task and then forecast how well that process will perform in the future. This was based on a realization that an information seeker's search goal or task can be mapped out as a series of actions, and that a sequence of actions or choices the searcher makes, especially the search path he takes, affects how well he will do. Thus, in contrast to approaches that measure the *goodness* of search *products* (e.g., documents, queries) as a way to evaluate the overall search effectiveness, we measured the likelihood of an existing search *process* to produce *good* results.

The findings from this work have larger implications on both evaluation and prediction frameworks in IR and on exploratory search. Given that the proposed technique is independent of any specific kind of system, and solely focused on user-based processes, it has potential to be applied to a variety of IR systems and situations irrespective of retrieval, ranking, or recommendation algorithms. One,

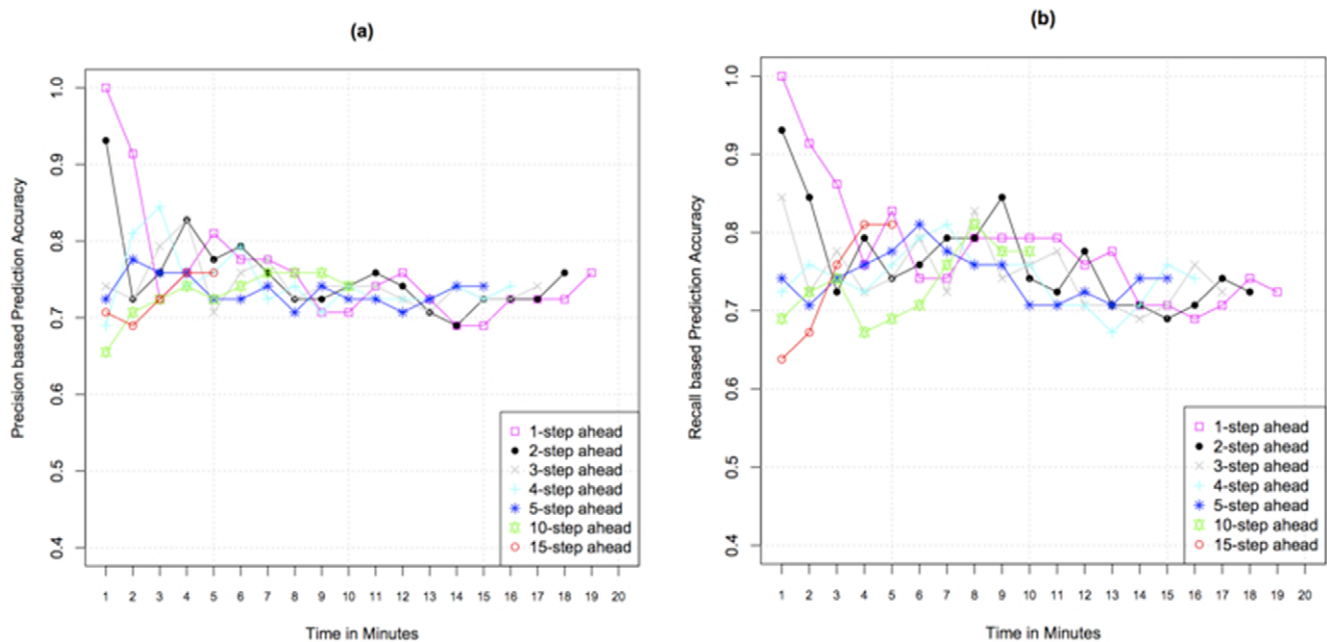


FIG. 5. Prediction accuracy comparison for Single Users: Oil Spill (a) Class labels based on precision, (b) Class labels based on recall. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

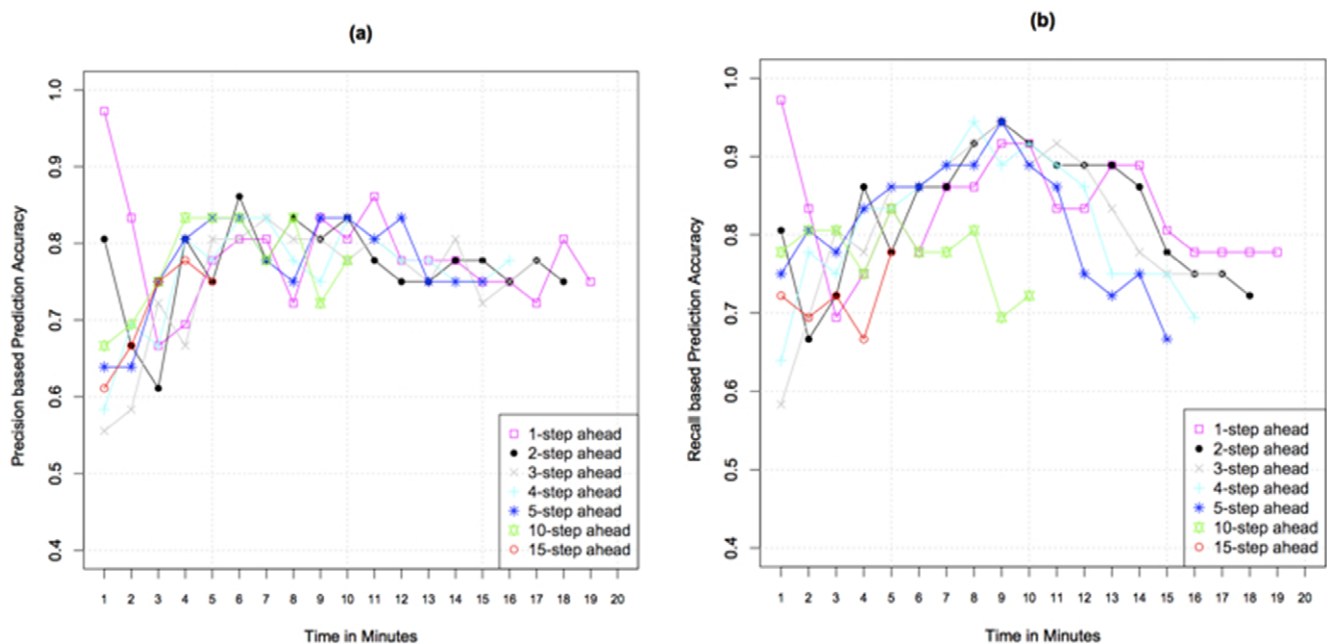


FIG. 6. Prediction accuracy comparison for Single Users: Revolution (a) Class labels based on precision, (b) Class labels based on recall. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

however, needs to understand that a different setup or data set may have different set of user actions available as features for our model. For example, Kules et al. (2009) studied searchers interacting with web-based faceted library catalog and found different amounts of time people spent on formulating the queries, looking at the results, and

examining the facets. One could use these features in the prediction model proposed here to see how those behaviors (e.g., query formulation, examining facets) help one understand and predict a searcher's performance. This particular example also demonstrates that depending on the context, one may have various forms of search-related actions

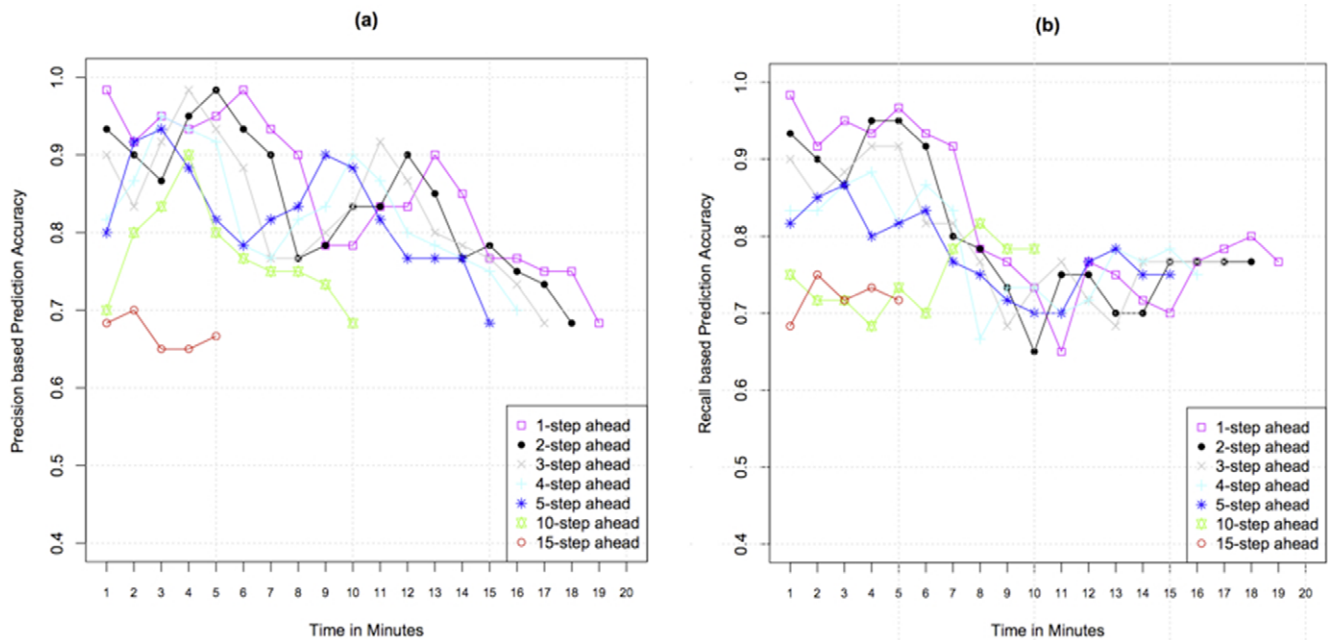


FIG. 7. Prediction accuracy comparison for Teams: Oil Spill (a) Class labels based on precision, (b) Class labels based on recall. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

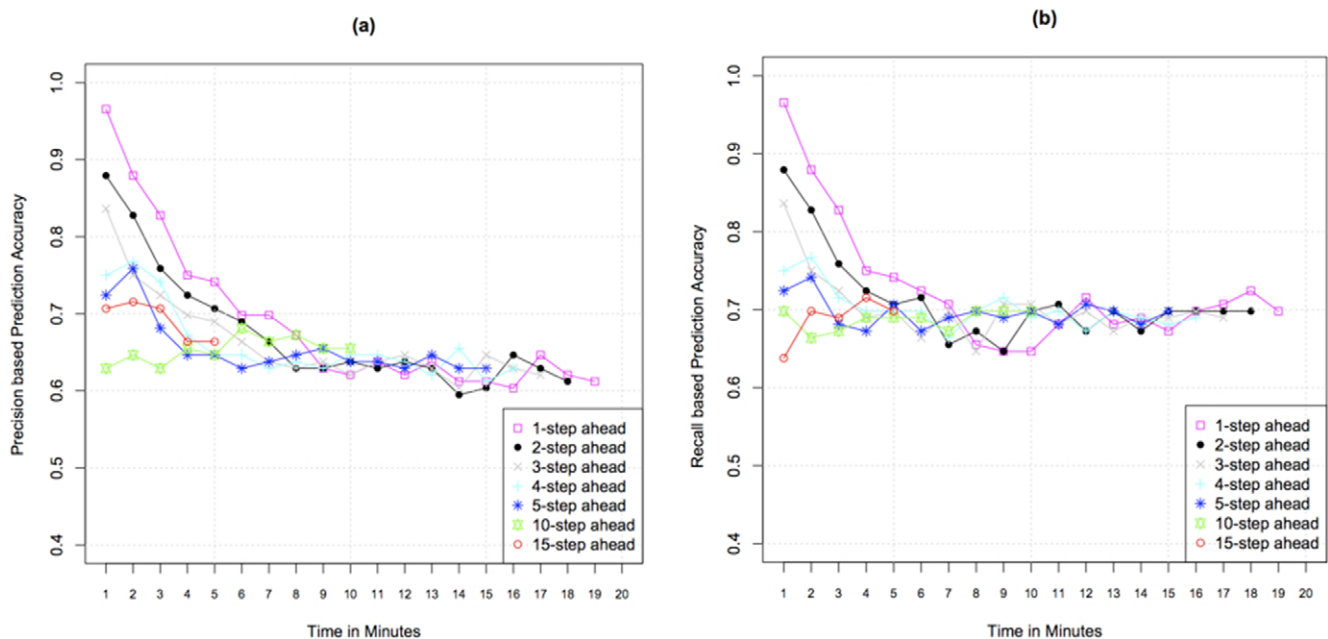


FIG. 8. Prediction accuracy comparison for Team Members as Independent Users: Oil Spill (a) Class labels based on precision, (b) Class labels based on recall. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

available in an exploratory search task, but the same model presented here can be used, accounting for those actions as features. Finally, although we have used data sets borrowed from previous user studies, one could easily apply the proposed method to web logs, Text REtrieval Conference (TREC) data, and other forms of data sets with various user actions recorded over time.

The proposed approach is not without its limitations. Here we assumed that the underlying search process is expressed as a set of physical actions and objects, referred to as *search path*. This assumption can be violated in some situations, including when users may not be behaving rationally (e.g., children or other special populations). The data sets we used also had limitations with respect to the task

(time-bound exploratory search), and motivation (extrinsic, monetary). A different kind of task or motivation may affect the way search processes are represented and evaluated. We also acknowledge that “information” is viewed here as artifact, even when we were studying information search processes. Perhaps a better way to capture search process may involve extracting cognitive features, which is beyond the scope of our work. Finally, the set of features used for evaluation and prediction were based on known literature, past experiences with similar tasks, and the availability of the data. One could use more, different, and richer features allowing for possible improvements in the results.

Conclusion

The primary objective of the work reported here was to find a way to forecast a user’s performance in an exploratory search task independent of the search system being used. We presented a set of experiments to address search performance prediction problem in exploratory search tasks. This was done by using data sets collected through several user studies and showing how different features extracted from a search process could be used to predict a searcher’s performance several minutes ahead.

Our specific findings from the work reported here are listed below.

- Our experiments with single users as well as pairs of users demonstrated that we could make reasonably accurate predictions in the early stage of their search process regarding their future performance (addressing RQ1).
- It was observed that features such as number of snippets, query specific features like length of query, number of tokens in query, and query entropy, were not significantly contributing to discriminating between the classes in the classification models at the early stages ($t = 1, 2, 3$) of the search process. This is because there were very few snippets/queries, which were used early during the process when a user is uncertain about the task at hand. In other words, query-related features were found to be inadequate indicators of search performance in the beginning of an information-seeking process.
- We also found that later during the process most of the features were found to be contributing significantly to the classification model (addressing RQ2). This finding helps to provide a better structure for process-focused IR evaluations. Such evaluations could include those utilizing time-series attributes of a search process in order to assess the effectiveness of a search system or a user, as well as predicting how the search process might look in the future. For instance, one could devise an evaluation metric that is time-dependent in that it would measure and account for different forms of information objects and user actions for evaluating the *goodness* of the search process depending on the time (absolute and relative with the overall task).

Another potential application of the prediction-based method described here would be to make users aware of how their future performance is determined by their

current/past search process. Based on a user’s early performance the system could then respond accordingly by identifying what could be done to improve overall performance.

Acknowledgment

The work reported here is supported by Institute of Museum and Library Services (IMLS) Early Career Development Grant #RE-04-12-0105-12.

References

- Adar, E., Weld, D.S., Bershad, B.N., & Gribble, S.D. (2007). Why we search: visualizing and predicting user behavior. In Proceedings of the 16th International Conference on World Wide Web (WWW '07) (pp. 161–170). New York: ACM. doi: 10.1145/1242572.1242595. <http://doi.acm.org/10.1145/1242572.1242595>
- Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734–749.
- Agichtein, E., White, R.W., Dumais, S.T., & Benet, P.N. (2012). Search, interrupted: understanding and predicting search task continuation. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (pp. 315–324). New York: ACM. <http://dl.acm.org/citation.cfm?doi=2348283.2348328>
- Alonso, O., Gertz, M., & Baeza-Yates, R. (2007). On the value of temporal information in information retrieval. *SIGIR Forum*, 41(2 (December 2007)), 35–41. doi: 10.1145/1328964.1328968; <http://doi.acm.org/10.1145/1328964.1328968>
- Beheshti, J., Cole, C., Abuhimed, D., & Lamoureux, I. (2014). “Tracking middle school students” information behavior via Kuhlthau’s ISP Model: Temporality. *Journal of the Association for Information Science and Technology*. doi: 10.1002/asi.23230. <http://onlinelibrary.wiley.com/doi/10.1002/asi.23230/abstract>
- Borlund, P., & Dreier, S. (2014). An investigation of the search behavior associated with Ingwersen’s three types of information needs. *Information Processing & Management*, 50(4), 493–507. <http://dx.doi.org/10.1016/j.ipm.2014.03.001>; ISSN 0306-4573, July 2014.
- Buckland, M.K. (1991). Information as thing. *Journal of the American Society for Information Science*, 42(5), 351–360.
- Case, D.O. (2012). Looking for information: A survey of research on information seeking, needs and behavior. Bingley, UK: Emerald Group Publishing.
- Cole, C. (2011). A theory of information need for IR that connects information to knowledge. *Journal of the American Society for Information Science and Technology*, 62(7), 1216–1231.
- Cole, C. (2012). Information need: A theory connecting information search to knowledge formation. Medford, NJ: Information Today.
- Cronen-Townsend, S., Zhou, Y., & Croft, W.B. (2002). Predicting query performance. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 299–306). New York: ACM. <http://dl.acm.org/citation.cfm?id=564429>
- Dervin, B. (1998). Sense-making theory and practice: An overview of user interests in knowledge seeking and use. *Journal of Knowledge Management*, 2(2), 36–46. doi: 10.1108/13673279810249369
- Dignum, S., Kruschwitz, U., Fasli, M., Kim, Y., Song, D., Beresi, U.C., & De Roeck, A. (2010). Incorporating seasonality into search suggestions derived from intranet query logs. In 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (Vol. 1, pp. 425–430). IEEE: Toronto, ON.
- Diriye, A., White, R., Buscher, G., & Dumais, S. (2012, October). Leaving so soon?: understanding and predicting web search abandonment rationales. In Proceedings of the 21st ACM international conference on

- Information and knowledge management (pp. 1025–1034). New York: ACM. <http://dl.acm.org/citation.cfm?id=2398399>
- Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White, T. (2005). Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)*, 23(2), 147–168.
- Guo, Q., White, R.W., Dumais, S.T., Wang, J., & Anderson, B. (2010). Predicting query performance using query, result, and user interaction features. In *Adaptivity, Personalization and Fusion of Heterogeneous Information* (pp. 198–201). LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE. New York: ACM. <http://dl.acm.org/citation.cfm?id=1937104>
- Golbandi, N.G., Katzir, L.K., Koren, Y.K., & Lempel, R.L. (2013, February). Expediting search trend detection via prediction of query counts. In *Proceedings of the sixth ACM international conference on Web search and data mining* (pp. 295–304). New York: ACM. <http://dl.acm.org/citation.cfm?id=2433435>
- González-Ibáñez, R., Haseki, M., & Shah, C. (2012). Time and space in collaborative information seeking: The clash of effectiveness and uniqueness. *Proceedings of the American Society for Information Science and Technology*, 49(1), 1–10. Baltimore, MD. Published online by Wiley Online Library.
- González-Ibáñez, R., Shah, C., & White, R.W. (2012). Pseudo-collaboration as a method to perform selective algorithmic mediation in collaborative IR systems. *Proceedings of the 75th Annual Meeting of the American Society for Information Science and Technology*, 49(1), 1–4. Published online by Wiley Online Library. <http://onlinelibrary.wiley.com/doi/10.1002/meet.14504901296/full>
- Gwizdka, J. (2008). Cognitive load on Web search tasks. Workshop on Cognition and the Web, Information Processing, Comprehension, and Learning. Granada, Spain. Available from http://eprints.rclis.org/14162/1/GwizdkaJ_WCW2008_short_paper_final.pdf
- Hassan, A., Jones, R., & Klinkner, K. (2010). Beyond DCG: user behavior as a predictor of a successful search. In *Proceedings of the third ACM international conference on Web Search and Data Mining (WSDM '10)* (pp. 221–230). New York: ACM. doi: 10.1145/1718487.1718515. <http://doi.acm.org/10.1145/1718487.1718515>
- He, B., & Ounis, I. (2006). Query performance prediction. *Information Systems*, 31(7), 585–594. 10.1016/j.is.2005.11.003; ISSN 0306-4379; November 2006.
- Ingwersen, P. (1986). Cognitive analysis and the role of the intermediary in Information retrieval. In: Davies, R., (Ed.), *Intelligent Information Systems*, Chichester, West Sussex, England: Horwood: 206–237.
- Jansen, B.J., & Rieh, S.Y. (2010). The seventeen theoretical constructs of information search and information retrieval. *Journal of the American Society for Information Science and Technology*, 61(8), 1517–1534.
- Järvelin, K. (2012). IR research: systems, interaction, evaluation and theories. In *ACM SIGIR Forum* (Vol. 45, No. 2, pp. 17–31). New York: ACM. <http://dl.acm.org/citation.cfm?id=2093348>
- Kent, A., Berry, M.M., Luehrs, F.U., Jr., & Perry, W. (1955). Operational criteria for designing information retrieval systems. *American Documentation*, 6(2), 1936–6108.
- Kuhlthau, C.C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science and Technology*, 42(5), 361–371.
- Kules, B., & Capra, R. (2009). Designing exploratory search tasks for user studies of information seeking support systems. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)* (pp. 419–420). New York: ACM. <http://dl.acm.org/citation.cfm?id=1555492>
- Kules, B., Capra, R., Banta, M., & Sierra, T. (2009). What do exploratory searchers look at in a faceted search interface?. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)* (pp. 313–322). New York: ACM. <http://dl.acm.org/citation.cfm?id=1555452>
- Kyoung-jae, K. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, 55(1–2), 307–319. ISSN 0925-2312; September 2003, 10.1016/S0925-2312(03)00372-2.
- Liao, Z., Song, Y., He, L., & Huang, Y. (2012). Evaluating the effectiveness of search task trails. In *Proceedings of the 21st international conference on World Wide Web (WWW '12)* (pp. 489–498). New York: ACM. doi: 10.1145/2187836.2187903. <http://doi.acm.org/10.1145/2187836.2187903>
- Liu, C., Gwizdka, J., Liu, J., Xu, T., & Belkin, N.J. (2010a). Analysis and evaluation of query reformulations in different task types. *Proceedings of the American Society for Information Science and Technology*, 47(1), 1–9. Published online by Wiley Online Library. <http://onlinelibrary.wiley.com/doi/10.1002/meet.14504701214/full>
- Liu, J., Gwizdka, J., Liu, C., & Belkin, N.J. (2010b). Predicting task difficulty for different task types. *Proceedings of the American Society for Information Science and Technology*, 47(1), 1–10. Published online by Wiley Online Library. <http://onlinelibrary.wiley.com/doi/10.1002/meet.14504701173/full>
- Lv, Y., & Zhai, C. (2009). Adaptive relevance feedback in information retrieval. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management—CIKM '09* (p. 255). New York: ACM Press. doi: 10.1145/1645953.1645988
- MacKay, D.M. (1969). *Information, mechanism and meaning*. Boston, MA: MIT Press.
- Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4), 41–46.
- Ming-Chi, L. (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications*, 36(8), 10896–10904. 10.1016/j.eswa.2009.02.038; ISSN 0957-4174; October 2009.
- Ord, J., Hyndman, R., Koehler, A., & Snyder, R. (2008). *Forecasting with exponential smoothing (The state space approach)*. Berlin: Springer.
- Radinsky, K., Svore, K., Dumais, S.T., Teevan, J., Horvitz, E., & Bocharov, A. (2012). Modeling and predicting behavioral dynamics on the web. In *Proceedings of the 21st international conference on World Wide Web (WWW '12)* (pp. 599–608). New York: ACM. doi: 10.1145/2187836.2187918. <http://doi.acm.org/10.1145/2187836.2187918>
- Rao, R., & Card, S.K. (1994). The table lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (pp. 318–322). New York: ACM. <http://dl.acm.org/citation.cfm?id=191776>
- Ricci, F., Rokach, L., Shapira, B., & Kantor, P.B. (Eds.). (2011). *Recommender systems handbook* (p. 842). Berlin: Springer. Retrieved from <http://www.springer.com/computer/ai/book/978-0-387-85819-7>
- Savenkov, D., Lagun, D., & Liu, Q. (2013). Search engine switching detection based on user personal preferences and behavior patterns. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '13)*. Dublin, Ireland. July 28—August 01, 2013, pp.33–42. <http://dl.acm.org/citation.cfm?id=2484099>
- Saracevic, T. (1995). Evaluation of evaluation in information retrieval. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 138–146). New York: ACM. <http://dl.acm.org/citation.cfm?id=215351>
- Shah, C. (2010). Coagmento-a collaborative information seeking, synthesis and sense-making framework. Integrated demo in *Proceedings of the ACM International Conference on Computer Supported Cooperative Work (CSCW)*, 6–11. New York: ACM. http://research.microsoft.com/en-us/um/redmond/groups/connect/CSCW_10/docs/p527.pdf
- Shah, C., & Croft, W.B. (2004). Evaluating high accuracy retrieval techniques. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 2–9). New York: ACM. <http://dl.acm.org/citation.cfm?id=1008996>
- Shah, C., & González-Ibáñez, R. (2010). Exploring information seeking processes in collaborative search tasks. *Proceedings of the American Society for Information Science and Technology*, 47(1), 1–7. Published online by Wiley Online Library. <http://onlinelibrary.wiley.com/doi/10.1002/meet.14504701211/full>
- Shah, C., & González-Ibáñez, R. (2011). Evaluating the synergic effect of collaboration in information seeking. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 913–922). New York: ACM. <http://dl.acm.org/citation.cfm?id=2010038>

- Shah, C., & González-Ibáñez, R. (2012). Spatial context in collaborative information seeking. *Journal of Information Science*, 38(4), 333–349.
- Shah, C., & Marchionini, G. (2010). Awareness in collaborative information seeking. *Journal of American Society of Information Science and Technology (JASIST)*, 61(10), 1970–1986.
- Shah, C., Liu, J., González-Ibáñez, R., & Belkin, N. (2012). Exploration of dynamic query suggestions and dynamic search results for their effects on search behaviors. In *Proceedings of American Society of Information Science & Technology (ASIS&T) Annual Meeting*. Baltimore, MD, October 26–30, 2012. Published by Wiley Online Library. <http://onlinelibrary.wiley.com/doi/10.1002/meet.14504901135/abstract>
- Shannon, C.E., & Weaver, W. (1963). *Mathematical theory of communication*. Urbana, IL: University of Illinois Press.
- Spink, A. (2002). A user-centered approach to evaluating human interaction with Web search engines: An exploratory study. *Information Processing & Management*, 38(3), 401–426.
- Taylor, R.S. (1968). Question-negotiation and information seeking in libraries. *College & Research Libraries*, 29(3), 178–194.
- Tunkelang, D. (2009). Precision and recall. *IEEE Computer*, 42(3), 39. in press.
- Weiss, S.M., & Indurkha, N. (1998). *Predictive data mining: A practical guide*. Burlington, Massachusetts: Imprint: Morgan Kaufmann. ISBN: 9781558604032.
- White, R. W., & Huang, J. (2010). Assessing the scenic route: measuring the value of search trails in web logs. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 587–594). New York: ACM. <http://dl.acm.org/citation.cfm?id=1835548>
- White, R.W., & Roth, R.A. (2009). *Exploratory search: Beyond the query-response paradigm*. Synthesis lectures on information concepts, retrieval, and services #3. Morgan & Claypool.
- White, R.W., Kules, B., Drucker, S., & Schraefel, M.C. (2006a). Supporting exploratory search. *Communications of the ACM (CACM)*, Vol: 49, Issue: 4.
- White, R.W., Muresan, G., & Marchionini, G. (2006b). Report on ACM SIGIR 2006 workshop on evaluating exploratory search systems. *SIGIR Forum*, 40(2), 52–60. doi: 10.1145/1189702.1189711
- Wildemuth, B. M., & Freund, L. (2012). Assigning search tasks designed to elicit exploratory search behaviors. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval (HCIR)* (Article no. 4). New York: ACM. <http://dl.acm.org/citation.cfm?id=2391228>
- Wilson, T.D., Ford, N., Ellis, D., Foster, A., & Spink, A. (2002). Information seeking and mediated searching. Part 2. Uncertainty and its correlates. *Journal of the American Society for Information Science and Technology*, 53(9), 704–715.
- Zhang, Y., Chen, W., Wang, D., & Yang, Q. (2011, August). User-click modeling for understanding and predicting search-behavior. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1388–1396). New York: ACM. <http://dl.acm.org/citation.cfm?id=2020613>