Is Exploratory Search Different? A Comparison of Information Search Behavior for Exploratory and Lookup Tasks

Kumaripaba Athukorala, Dorota Głowacka, and Giulio Jacucci
Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Helsinki, Finland. E-mail: kumaripaba.athukorala@helsinki.fi; glowacka@cs.helsinki.fi; giulio.jacucci@helsinki.fi

Antti Oulasvirta
Department of Communications and Networking, Aalto University, School of Electrical Engineering, Espoo, Finland. E-mail: antti.oulasvirta@aalto.fi

Jilles Vreeken
Max Planck Institute for Informatics and Saarland University, Saarbrücken, Germany. E-mail: jilles@mpi-inf.mpg.de

Exploratory search is an increasingly important activity yet challenging for users. Although there exists an ample amount of research into understanding exploration, most of the major information retrieval (IR) systems do not provide tailored and adaptive support for such tasks. One reason is the lack of empirical knowledge on how to distinguish exploratory and lookup search behaviors in IR systems. The goal of this article is to investigate how to separate the 2 types of tasks in an IR system using easily measurable behaviors. In this article, we first review characteristics of exploratory search behavior. We then report on a controlled study of 6 search tasks with 3 exploratory—comparison, knowledge acquisition, planning—and 3 lookup tasks—fact-finding, navigational, question answering. The results are encouraging, showing that IR systems can distinguish the 2 search categories in the course of a search session. The most distinctive indicators that characterize exploratory search behaviors are query length, maximum scroll depth, and task completion time. However, 2 tasks are borderline and exhibit mixed characteristics. We assess the applicability of this finding by reporting on several classification experiments. Our results have valuable implications for designing tailored and adaptive IR systems.

Received March 27, 2015; revised July 14, 2015; accepted July 14, 2015

● Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23617

Introduction

Search activities are commonly divided into two broad categories: lookup and exploratory (Marchionini, 2006). Lookup search is by far the better understood and assumed to have precise search goals. The predominant design goal in information retrieval (IR) systems has been fast and accurate completion of lookup searches. Exploratory search is presently thought to center around the acquisition of new knowledge and considered to be challenging for the user (White & Roth, 2009). Although there has been a lot of research on understanding exploratory search, there are many open questions when it comes to the design of IR systems that provide tailored and adaptive support. One of the key problems is how we can make an IR system automatically distinguish the two categories of search in the course of a search session (Belkin, 2008). In this article, we look into if, and how well, we can tell apart lookup and exploratory search activities from properties that IR systems can easily observe.

It is difficult to separate exploratory and lookup search in IR systems. This is because currently there is a gap between our knowledge in exploratory search behaviors and requirements of IR system design. First, many studies compared the exploratory and lookup searches by cognitive strategies only (J. Kim, 2009; Thatcher, 2008). However, IR systems require more reliable quantitative behavioral indicators to be able to act on them. Second, studies that do empirically analyze implicit measures in exploratory and lookup searches focus only on the most obvious type of exploratory

TABLE 1. List of search tasks categorized under lookup and exploratory search categories by Marchionini (2006). The set of tasks that are included in our investigation are in the row marked "included." We operationalize tasks with clear exploratory or lookup characteristics as *core* tasks. Core tasks are highlighted with bold font.

		Exploratory	
	Lookup	Learning	Investigation
Included	**Navigational**	**Knowledge acquisition**	**Planning**
	Fact-finding/Informational	Comparison	
	Question answering		
Not included	Known item	Comprehension	Accretion
	Transactional	Aggregation	Analysis
	Verification	Socialize	Exclusion
			Evaluation
			Discovery
			Synthesis
			Transform

activity: learning or knowledge acquisition. However, it is held that exploratory search involves many subcategories of search activities (Marchionini, 2006; White & Roth, 2009). Third, many studies that attempt to distinguish between different task types only consider web search behaviors but not behaviors specific to IR system use (Liu et al., 2010b). There are marked differences between web searching and searching with IR systems (Jansen & Pooch, 2001). To this end, a thorough empirical analysis of exploratory and lookup activities within an IR environment is necessary. Moreover, to provide tailored and adaptive support, we should be able to predict the task type as early as possible. Hence, we need properties that we can measure from the first search engine results page (SERP) on. Finally, the information search behaviors considered in all the prior studies pay little attention to the interactions that are observable early on in the search process.

Our objective is to provide a systemic and rigorous analysis of exploratory and lookup information search behaviors across several search activities. Our definition of information search behaviors builds on the conceptualization of Li and Belkin (2010), who define information search behavior as interactions between users and IR systems. We are particularly interested in directly measurable behaviors, that can be leveraged to support IR systems in identifying the type of search activity as early as possible.

To subject exploratory search to empirical investigation, we first operationalize exploratory and lookup categories by reviewing prior studies. We then design a controlled study that allows us to clearly set the search tasks and control other variables that could affect search behavior, such as prior knowledge and task difficulty. As participants in our study, we consider users with a background in computer science who will search for scientific literature in the machine learning domain, with arXiv data set. In identifying representative tasks, we follow the framework of Marchionini (2006). This framework assigns the lower level search activities, such as, fact-finding, knowledge acquisition, into high-level categories—exploratory and lookup. For clarity, we will refer to such lower level search activities as "tasks." We give an overview of the taxonomy and the tasks we selected for our investigation in Table 1. Investigating the taxonomy, we find that some of these tasks not only have characteristics of the assigned category, but also of the other. We refer to these tasks as "borderline." Later in this article, we operationalize tasks that can be clearly assigned to either exploratory or lookup categories as *core* tasks and others as *borderline* tasks. To make an informed decision, we review information search behaviors identified in the literature and select a set of behaviors that are both expected to be informative as well as easy to measure by an IR system.

Our goal is to push forward the design of IR techniques and search user interfaces to better cover this important aspect of search behavior. Presently, research on search interfaces has proposed various techniques to support exploratory search tasks (Diriye, 2012; Kules et al., 2008). However, to adapt them to different subcategories of exploration, we need to better distinguish between search tasks (Cutrell & Guan, 2007). Moreover, most of the retrieval algorithms treat exploratory and lookup tasks in the same way. Knowledge of the task category can be used to improve the performance of IR algorithms and compute the implicit relevance feedback more accurately (Joachims, Granka, Pan, Hembrooke, & Gay, 2005). Thorough understanding of how users behave in exploratory search can also improve user modeling techniques and evaluation methods of IR systems (Athukorala, Oulasvirta, Głowacka, Vreeken, & Jacucci, 2014; Pirolli & Card, 1995).

Our main contribution is a systematic enumeration and quantification of behavioral variables that can be used to separate lookup versus exploratory tasks. These data suggest that core exploratory tasks can be distinguished from the core lookup tasks with a few simple indicators of information search behaviors. The most informative indicators are the first query length, maximum scroll depth, and task completion time. However, the two borderline tasks show mixed behaviors. To critically show that the outcome of this study is actionable we trained a machine learning classifier

TABLE 2. Summary of studies that investigate the effect of search *goal*, *difficulty*, *complexity*, and user *knowledge* on information search behavior.

Aims of the study	User tasks	Information search behaviors	References
Distinguish search goals in web search	factual, interpretive, exploratory (experts only)	qualitative analysis self-reports and screen recording	J. Kim (2009)
	navigational, informational, transactional (search engine logs)	query properties	Jansen et al. (2008); Rose and Levinson (2004); Broder (2002)
	fact findings tasks with specific, mixed and amorphous goals, and low and high objective complexity	task completion time, page visits, queries issued, unique search engines, and eye gaze data	Liu et al. (2010b)
Guidelines to support web search goals	navigational, informational (experts vs. novices)	qualitative analysis of cognitive style	Navarro-Prieto et al. (1999); Palmquist and Kim (2000)
Investigate how search goal and expertise affect web search	known-item, subject (experts vs. novices)	number of visited nodes, issued keyword searches, frequency of clicking back buttons, jump options, and Home button	K.-S. Kim (2001)
Investigate document relevance and search goal	parallel and dependent	dwell time	Liu and Belkin (2010a)
	navigational and informational	click data, gaze distribution	Joachims et al. (2005)
Task difficulty and search behaviors	easy and difficult closed informational tasks	query properties, task completion time, proportion of first result page browsing time	Aula et al. (2010)
	open, closed with different difficulty levels	dwell time	Liu et al. (2010c)
Subjective task complexity and qualitative reports	tasks are not controlled, participants self-categorized the tasks as automatic information processing, normal information processing, and decision	self-reports (diaries, interviews)	Byström (2002); Byström and Järvelin (1995)
	routine, normal, genuine	self-reports	Ingwersen and Järvelin (2006)
Analyzed objective task complexity	tasks with three levels of complexity	self-reports	Bell and Ruthven (2004)
Investigate domain knowledge and web expertise	informational	query properties, information-seeking steps	Hölscher and Strube (2000)
	fact-finding, exploratory	interviews, cognitive strategies	Navarro-Prieto et al. (1999)
	navigational, knowledge acquisition	task completion time, qualitative analysis of navigation path	Jenkins et al. (2003)
	general, specific	qualitative analysis of navigation path	Saito and Miwa (2001)
	open, closed (novices only)	query properties, task completion time	Marchionini (1989)

to predict the task type. According to the classifier, the core lookup tasks are separable from the core exploratory tasks with nearly 85% accuracy. Moreover, the findings confirm some, but not all, assumptions of the current understanding of these tasks. The in-depth characterization of behavior we provide will be of interest to both theorists and pragmatists struggling to infer search goals, and tasks from implicit sources of evidence (Belkin, 2008).

Background

This review will serve as our basis for operationalizing exploratory and lookup tasks as well as identifying variables to measure information search behaviors in our study. To this end we first investigate factors that influence information search behavior. Then, we cover how in prior work exploratory and lookup search tasks have been operationalized for those facets.

Factors Influencing Information Search Behaviors

Information search tasks can be classified using many factors that affect search behavior (Li & Belkin, 2008, 2010; Liu & Belkin, 2010a). The most salient factors include the search *goal*, objective and perceived *complexity* and *difficulty* of the task, and the *knowledge of the user*. Below, we review each aspect and Table 2 summarizes them.

Search goal is the primary reason for a user to interact with an information search system (J. Kim, 2009). Many studies manipulated the preciseness of the search goal definition and investigated how it affects user behavior. In an early study of encyclopedia use by novices, Marchionini (1989) introduced two types of tasks—"closed tasks" with precise search goals, and "open tasks" with fuzzy search goals and no definite boundary. According to the results, in open tasks, novices have difficulty in formulating search queries, spend more time and involve a higher number of query reformulations. K.-S. Kim (2001) investigated the navigational style of novice and expert web users with known-item search and subject search goals. Here, "subject search" is similar to open task and the results indicate that the number of visited nodes, issued keyword searches, and frequency of clicking different buttons are influenced by the search goal. In another study J. Kim (2009) qualitatively analyzed information-seeking strategies of web users with three search goals: *factual* or finding a definitive answer with a precise search goal, *interpretive* or configuring an answer with a less precise search goal, and *exploratory* or broadening knowledge with open-ended search goals. Results suggest that in exploratory tasks users spend considerable time reading a found page to determine its relevance. These studies indicate that when the search goal is less precise users behave differently. Task completion time, number of query formulations, click interactions, and reading time are useful metrics of information search behaviors.

There are other studies that categorize web search goals as informational, navigational, and transactional. Navarro-Prieto, Scaife, and Rogers (1999) and Palmquist and Kim (2000) investigated how *navigational* and *informational* search goals affect cognitive styles. Jansen, Booth, and Spink (2008), Rose and Levinson (2004), and Broder (2002) use external evaluators to manually classify search queries collected from search engine logs into these search goals and investigated how to distinguish them from query properties. These studies provide useful findings. However, the log data are assessed by external evaluators and their evaluation may be different from the intent of the user, which makes these evaluations rather unreliable (Rose & Levinson, 2004).

Difficulty and *complexity* are two other important factors that influence user behavior. Task difficulty is always considered as a subjective measure that depends on the user perception (Li & Belkin, 2008). Task complexity is measured with both objective and subjective approaches. It is difficult to distinguish between subjective task complexity and task difficulty because they are both assessed by the task doer with respect to their familiarity and degree of uncertainty within the task requirements (Bell & Ruthven, 2004; Byström, 2002; Vakkari, 2003). However, objective task complexity is different from difficulty, and it is commonly measured by the number of paths involved in the search process (Byström, 2002). Tasks with a single determinable path that could be easily automated are commonly referred to as simple tasks, whereas tasks where the results, process, and information requirements are indeterminable were categorized as complex tasks. Literature suggests exploratory search tasks to have high objective task complexity (White & Roth, 2009).

Several studies categorize tasks by considering the search goal and the complexity or difficulty. For example, Liu et al. (2010b) categorized web search tasks by considering the preciseness of the search goal, objective complexity, product (is the outcome factual or intellectual), and level (whether the document is judged as a whole or a segment). Although they do not explicitly compare their task classification with characteristics of exploratory and lookup tasks, their classification is intuitive and shows that there are tasks with mixed characteristics—such as specific search goals but high complexity. They show that web search behaviors, such as task completion time, number of different search engines used, eye movement behavior, and queries issued are all affected by the complexity and the preciseness of the search goal. In a similar study Liu, Liu, Gwizdka, and Belkin (2010c) analyzed how task difficulty and two types of search goals—open and closed, influence search behavior. Their results suggest that closed tasks and difficult tasks are associated with long dwell time, which measures the time spent on reading retrieved documents. Aula, Khan, and Guan (2010) explored how to detect task difficulty from information search behaviors by assigning easy and difficult closed informational tasks. They found that when tasks become more difficult, users issue numerous search queries, view many results, and spend more time on search results pages. Similar studies demonstrate that in exploratory search tasks users display similar behavior (Hassan, White, Dumais, & Wang, 2014; Marchionini, 2006; White & Chandrasekar, 2010). This work shows the importance of further investigations on disambiguating lookup and exploratory tasks, while fixating the task difficulty at a moderate level.

Knowledge of the user is another factor influencing information search behavior (Li & Belkin, 2008). Prior studies revealed that web experts heavily rely on query-formatting tools, whereas domain experts with low experience in Internet use heavily rely on terminology and avoid query formatting (Hölscher & Strube, 2000). There are several studies aiming to understand how cognitive strategies are influenced by the level of domain knowledge, web expertise and task type (Jenkins, Corritore, & Wiedenbeck, 2003; Navarro-Prieto et al., 1999; Saito & Miwa, 2001). Navarro-Prieto et al. (1999) compared fact-finding and exploratory tasks with dispersed structure and category structures. Their results provide qualitative evidence that web experts follow different cognitive strategies compared to novices in exploratory tasks.

In summary, previous studies point to differences in task completion time, number of queries issued, dwell time, etc., for different task types. However, they miss two important aspects with respect to the design of IR systems. First, they focus on web search rather than IR system use. Hence, many measures they use, such as the number of unique search engines used, are less informative. Furthermore, Jansen and Pooch (2001) suggest that there are marked differences between web search and IR system use, because IR systems create a special environment with a specific data set. Second, most of these studies examined search behaviors from the entire search session level, rather than the first query session level. To adapt IR systems to different task types, we need measures of search behaviors that allow us to predict the task type as early as possible.

Definitions of Lookup and Exploratory Search

Lookup is the most basic kind of search, which returns discrete and well-structured objects, such as specific websites or definitions (White & Roth, 2009). Most distinctive types of lookup tasks involve finding facts (also referred to as factual) to answer a specific question, for example, the amount of blood a human heart pumps in a minute (Aula & Nordhausen, 2006). Common characteristics of lookup tasks are precise search goals with simple search paths. The search process of the simplest lookup tasks can even be automated (Byström, 2002). There are also broader lookup tasks where the search goal is precise and the user could decide easily whether they found the answer, yet the search process is more complex and would involve several paths, for example, finding information on different antivirus software and their prices (Aula & Nordhausen, 2006). In similar studies, lookup tasks that involve thinking or understanding rather than simply locating an item are referred to as interpretive tasks (J. Kim, 2009). These kinds of lookup tasks are more focused and goal-oriented than exploratory tasks, yet, they may involve locating several results to configure an answer.

Exploratory search is naturally multifaceted, so there is a wide variety of qualitative definitions (Wildemuth & Freund, 2012). Marchionini (2006) illustrated exploratory and lookup tasks as an overlapping cloud and suggested that lookup tasks are embedded in exploratory tasks and vice versa. The problem context that motivates the search and the search process are two primary attributes considered in the definitions of exploratory search (White & Roth, 2009). Imprecise task requirements or open-ended search goals are attributes commonly used in the literature to define exploratory search with respect to the problem context (J. Kim, 2009). The exploratory search process is considered to be cognitively complex with the information seeker being uncertain about the search process (White & Roth, 2009). These attributes of exploratory search influence search behavior, such as the number of search queries issued and links clicked, and the duration of the search task (Marchionini, 2006). However, there are exploratory tasks with borderline characteristics. For example, Navarro-Prieto et al. (1999) defined two types of exploratory tasks, (a) dispersed structure and (b) category structure. Exploratory tasks with dispersed structure have open-ended search goals as well as complex search paths. There are borderline exploratory tasks with open-ended search goals but low complexity in the search process—for example—find all information about the 1997 Nobel Prize in Literature (Navarro-Prieto et al., 1999). These characteristics of lookup and exploratory tasks make it difficult to clearly separate the two categories.

Approach

The goal of this article is to explore if, and how well, we can distinguish exploratory search tasks from lookup tasks in an IR system using only search behavior information that is easily measurable. We start by seeking a conceptualization of exploratory and lookup tasks. We then identify the most appropriate tasks and corresponding information search behaviors. We finally design an experiment controlling the task type and external factors.

Operationalizing Exploratory and Lookup Tasks

As prior work shows, there are numerous facets by which search tasks can be categorized. As our goal is to distinguish tasks, we should only consider those facets that characterize exploratory and lookup tasks. Following insights from prior work we provided a demarcation by using two primary facets—preciseness of the problem context or the search *goal*, and the *objective complexity* of the search process (Liu et al., 2010b). We keep constant the values of two subjective measures—user knowledge, and subjective or perceived task difficulty because they are not necessary characteristics of either exploratory or lookup tasks, rather, both lookup and exploratory tasks are likely to be conducted in familiar and unfamiliar domains as well as perceived to be either easy or difficult (Hassan et al., 2014). For example, a user with no background in human biology could look for a very specific fact, such as the amount of blood a human heart pumps (Aula & Nordhausen, 2006). A fact-finding task like this cannot be categorized as an exploratory search task just because the user has no background in the search topic (Marchionini, 2006). In this operationalization exploratory and lookup tasks have the following characteristics:

Goal. In *exploratory* tasks, the search goal is imprecise and open-ended. That is, there does not exist a *single* answer that accomplishes the user's information needs and no clear criterion on when to end the search. Hence, the assessment of the relevance of results is not discrete. In *lookup* tasks, there does exist a precise search goal. The search goal is reached by retrieving a finite set of relevant results, and the relevance of results can be assessed discretely.

Complexity. Objective complexity of a search task is commonly defined by the number of paths involved in the search process (Byström, 2002). This objective is intuitive and used in many studies (Li & Belkin, 2008; Liu et al., 2010b). Clearly, for *exploratory* tasks we cannot identify a single and direct path that leads to the desired results. Therefore, exploratory tasks have high complexity (White & Roth, 2009). In lookup tasks, the search process is more straightforward and involves only a few steps—lookup tasks are typically of much lower complexity than exploratory search tasks.

Table 3 illustrates the primary categorization of tasks according to this conceptualization. We use the terms "core lookup" and "core exploratory" to refer to tasks that clearly fit the aforementioned characteristics. Core exploratory tasks have both high complexity for the search process and imprecise search goals, whereas the core lookup tasks have

TABLE 3. Our categorization of exploratory and lookup tasks according to their primary facets.

	Low complexity	High complexity
Precise goals	Core lookup	Borderline lookup
Open-ended goals	Borderline exploratory	Core exploratory

low complexity and precise search goals. However, there are tasks with mixed characteristics. There are lookup tasks with precise search goals, yet the search process is not straightforward—we referred to them as "borderline lookup" tasks (Aula & Nordhausen, 2006; Liu et al., 2010b). We refer to the other category of tasks with open-ended search goals but low complexity as "borderline exploratory" (Navarro-Prieto et al., 1999). In this study we explore how well we can distinguish both core and borderline tasks.

Experimental Approach

As all methods, experimental approaches have drawbacks—for example, users who are not truly motivated to perform the tasks. In our setting, however, the alternative approach of collecting data from search engine logs would provide little information on the actual task that was performed (Rose & Levinson, 2004), let alone about task success (Aula & Nordhausen, 2006). Additionally, information search is affected by many other factors (Li & Belkin, 2008). A well-designed experiment including realistic tasks, questionnaires, and follow-up interviews allows us to obtain a rich data set while controlling other factors that affect search behavior.

We control three external factors that could affect search behaviors—domain knowledge, search expertise, and perceived task difficulty—while altering the task complexity and preciseness of search goals—factors defining exploratory and lookup tasks. In our setting, participants performed both exploratory and lookup tasks in moderately familiar domains with expert search skills. We select expert web users because they adapt their search behavior according to the task type (Saito & Miwa, 2001). We define "expert web users" as those who search for information daily as part of their work tasks (Jenkins et al., 2003).

We use a representative version of the most commonly used interface for literature search, Google Scholar (Athukorala, Hoggan, Lehtiö, Ruotsalo, & Jacucci, 2013). As we want to allow for experimental features, like logging, we cannot use it directly but instead design a very similar interface. Moreover, as Google Scholar does not allow access to its data set, we instead use a free digital library, arXiv, as our data source. arXiv is one of the most popular open access digital libraries in mathematics and computer science domains. In all other aspects, the interface is very close to Scholar; each result snippet contained the article title, authors, publication forum and year, and part of the abstract. Users can click any result item and further investigate the articles if needed. In exploratory tasks users are expected to explore more results (White & Roth, 2009), yet many users do not move beyond the first SERP as a habit even if they are interested in exploring more results (Jansen et al., 2008). Hence, to investigate this behavior without having users to click through to "second page results," we display more results than traditional interfaces do. To determine how many items to display, we consult the literature. Athukorala et al. (2014) showed that in exploratory tasks users are interested in scanning at least 33 items. We round this number up and display 40 items per SERP. Seven items are visible on the screen without the need to scroll down. Figure 1 illustrates the interface.

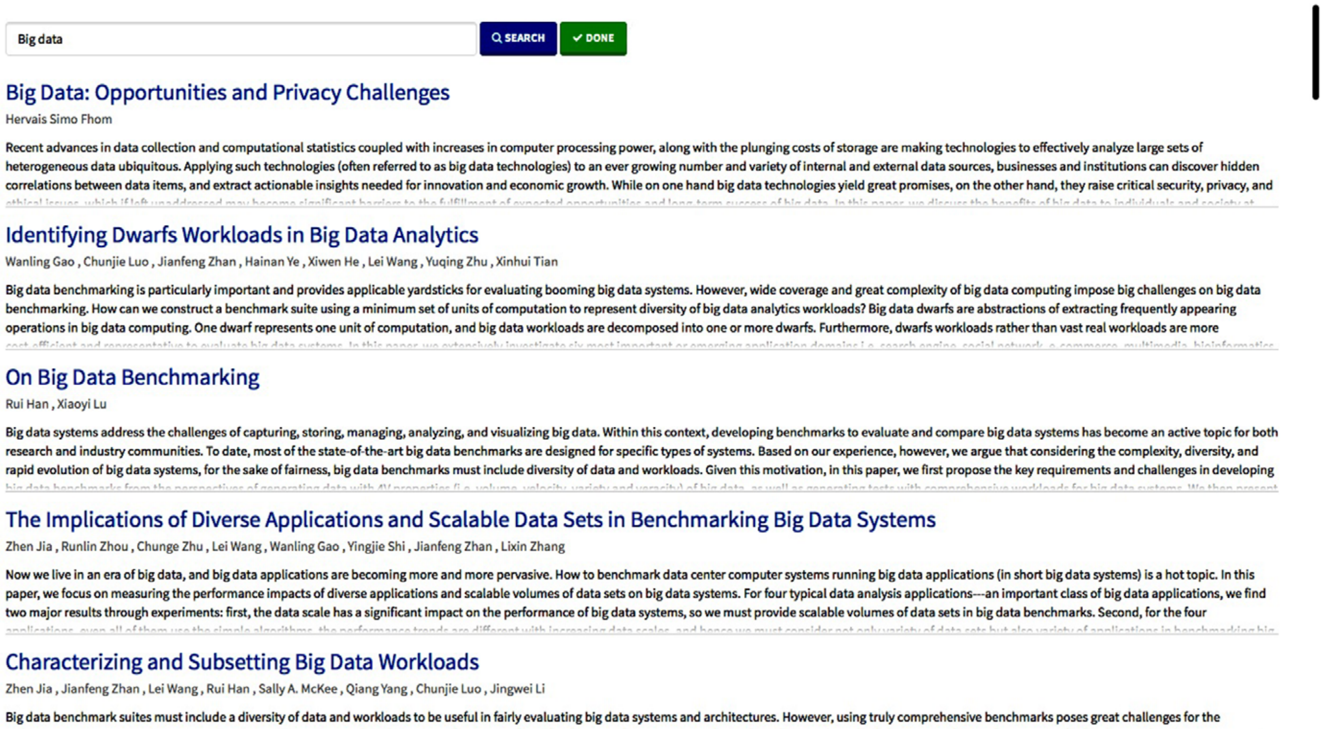
We set the tasks in an academic information search scenario because a main goal of exploratory search is to acquire new knowledge, which is particularly important within an academic context (Wildemuth & Freund, 2012). Further, user behaviors in different search tasks in the scientific domain are less well-studied (White & Roth, 2009). Other advantages of using the scientific domain include the availability of free data sets. We selected the machine learning domain to create all the tasks because there is a good coverage of machine learning courses at the university, which makes it easier to recruit participants with same familiarity of the topics. Additionally, a large number of machine learning articles are freely available in our data set, arXiv.

Task Selection

As there exists a large variation in how exploratory and lookup tasks are defined, we need a systematic approach to select a set of representative tasks. According to our operationalization, there exist both *core* and *borderline* tasks for each task category. We choose the framework of Marchionini (2006) to select key tasks from each category. Then, using both our operationalization and prior literature, we label the selected task as either *core* or *borderline*.

We select three tasks from the lookup category and the exploratory category—resulting in six task altogether (Table 1). Out of the six tasks Marchionini's framework identifies as lookup, we select three—*navigational search*, *fact-finding*, and *question answering*. According to the literature, of these three both fact-finding and navigational tasks display the *core* lookup characteristics, whereas the question answering task is identified by borderline characteristics (Aula & Nordhausen, 2006). We consider these tasks, and not all six, as the remaining three (known-item, transactional, and verification search) are not relevant to our setting. Although beyond the scope of this work, it would make for interesting future research to investigate how well these tasks can be identified by an IR system.

Marchionini's framework includes many tasks under the exploratory search category, however, there is little information about the differences between them (White & Roth, 2009). That is to say, it is unclear how to create distinct search tasks for each type. Therefore, we focus on representative three exploratory search tasks—*knowledge*



Big data

Big Data: Opportunities and Privacy Challenges
Hervais Simo Fhom
Recent advances in data collection and computational statistics coupled with increases in computer processing power, along with the plunging costs of storage are making technologies to effectively analyze large sets of heterogeneous data ubiquitous. Applying such technologies (often referred to as big data technologies) to an ever growing number and variety of internal and external data sources, businesses and institutions can discover hidden correlations between data items, and extract actionable insights needed for innovation and economic growth. While on one hand big data technologies yield great promises, on the other hand, they raise critical security, privacy, and ethical issues, which if left unaddressed may become significant barriers to the fulfillment of expected opportunities and long-term success of big data. In this paper, we discuss the benefits of big data to individuals and society at

Identifying Dwarfs Workloads in Big Data Analytics
Wanling Gao , Chunjie Luo , Jianfeng Zhan , Hainan Ye , Xiwen He , Lei Wang , Yuqing Zhu , Xinhui Tian
Big data benchmarking is particularly important and provides applicable yardsticks for evaluating booming big data systems. However, wide coverage and great complexity of big data computing impose big challenges on big data benchmarking. How can we construct a benchmark suite using a minimum set of units of computation to represent diversity of big data analytics workloads? Big data dwarfs are abstractions of extracting frequently appearing operations in big data computing. One dwarf represents one unit of computation, and big data workloads are decomposed into one or more dwarfs. Furthermore, dwarfs workloads rather than vast real workloads are more cost-efficient and representative to evaluate big data systems. In this paper, we extensively investigate six most important or emerging application domains i.e. search engine, social network, e-commerce, multimedia, bioinformatics

On Big Data Benchmarking
Rui Han , Xiaoyi Lu
Big data systems address the challenges of capturing, storing, managing, analyzing, and visualizing big data. Within this context, developing benchmarks to evaluate and compare big data systems has become an active topic for both research and industry communities. To date, most of the state-of-the-art big data benchmarks are designed for specific types of systems. Based on our experience, however, we argue that considering the complexity, diversity, and rapid evolution of big data systems, for the sake of fairness, big data benchmarks must include diversity of data and workloads. Given this motivation, in this paper, we first propose the key requirements and challenges in developing big data benchmarks from the perspectives of generating data with 4V properties (i.e. volume, velocity, variety and veracity) of big data, as well as generating tests with comprehensive workloads for big data systems. We then present

The Implications of Diverse Applications and Scalable Data Sets in Benchmarking Big Data Systems
Zhen Jia , Runlin Zhou , Chunge Zhu , Lei Wang , Wanling Gao , Yingjie Shi , Jianfeng Zhan , Lixin Zhang
Now we live in an era of big data, and big data applications are becoming more and more pervasive. How to benchmark data center computer systems running big data applications (in short big data systems) is a hot topic. In this paper, we focus on measuring the performance impacts of diverse applications and scalable volumes of data sets on big data systems. For four typical data analysis applications---an important class of big data applications, we find two major results through experiments: first, the data scale has a significant impact on the performance of big data systems, so we must provide scalable volumes of data sets in big data benchmarks. Second, for the four applications, even all of them use the simple algorithms, the performance trends are different with increasing data scales, and hence we must consider not only variety of data sets but also variety of applications in benchmarking big

Characterizing and Subsetting Big Data Workloads
Zhen Jia , Jianfeng Zhan , Lei Wang , Rui Han , Sally A. McKee , Qiang Yang , Chunjie Luo , Jingwei Li
Big data benchmark suites must include a diversity of data and workloads to be useful in fairly evaluating big data systems and architectures. However, using truly comprehensive benchmarks poses great challenges for the

FIG. 1. Screen shot of the interface. Note that this image shows only part of the 40 items that are displayed per page. Users can scroll down to explore more items. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

acquisition, *planning*, and *comparison*—that are the most suitable for scientific search and are commonly used in other studies of exploratory search (White & Roth, 2009; J. Kim, 2009). Among these, knowledge acquisition and planning exhibit core exploratory characteristics (Brand-Gruwel, Wopereis, & Vermetten, 2005; Navarro-Prieto et al., 1999). We define comparison tasks in such a way that they have borderline characteristics.

Indicators of Information Search Behaviors

We analyze prior work to identify the most suitable indicators of information search behaviors. We focus on behaviors that could be captured quickly within the first query iteration. First query iteration refers to all the interactions between the user and the IR system pertaining to the first query (i.e. the first SERP) up until another query is entered or the session ends. The information search behaviors we selected allow us to predict the task type without any post-processing, while the user is actively engaged in the task and before leaving the first SERP. Therefore, we did not include all the qualitative and postquery analysis behaviors from prior work, such as number of changes in querying approach or cognitive search strategies. These behaviors allow IR systems to adapt their support as early as possible.

Query-related behaviors are the most common, from which we selected two measures:

Query length: Total number of terms in the first query. A term is defined as "a string of characters separated by some delimiter such as a space, a colon, or a period" (Jansen & Pooch, 2001, p. 244). We select query length because the literature suggests that task type affects the query length (Aula & Nordhausen, 2006).
Query duration: The duration of the first query iteration. In exploratory tasks users need more time to get familiarized with the topic to formulate new queries (White & Roth, 2009). We included first query duration to capture this behavior.

Query complexity is mostly used in older studies where logical operators were common in queries. Nowadays logical operators are rarely used (Jansen, Spink, & Saracevic, 2000) and hence we exclude query complexity.

We also select two behaviors related to the interactions with the first SERP:

Maximum scroll depth: The maximum number of results exposed by scrolling. In exploratory search tasks users are expected to explore more items (Marchionini, 2006) but the scroll behavior is never examined. We log the scroll time and position to calculate the maximum number of items in the SERP (out of 40) the user was exposed to by scrolling during the first query iteration.
Cumulative clicks: The total number of links in the SERP the user clicked. In exploratory search tasks users tend to click more items (White & Roth, 2009). We include this indicator to evaluate whether this behavior holds for all the different types of exploratory tasks.

DOI: 10.1002/asi

We select three behaviors related to time:

Proportion of browsing: Proportion of first query duration spent browsing the results in the first SERP. This behavior is used in identifying successful search sessions when the search process is complex (Aula et al., 2010). It is suitable for capturing the objective complexity in exploratory tasks.

Duration dwelling: The time users spend on investigating clicked documents. Dwell time is used as a predictor of document relevance when task type and user information are known in advanced (Liu & Belkin, 2010a). The task type affects the duration dwelling when other parameters are controlled.

Task completion time: Total time users spend from the moment they issues the first query until the task done button is clicked. Exploratory tasks tend to last longer (Marchionini, 2006). Even though we cannot capture this behavior within the first query iteration, we included it to assess the validity of this claim for borderline and core tasks.

In addition, we select a feature that requires an external sensor: eye tracking. Eye trackers are becoming more common. We include a behavior that can be easily captured through eye trackers.

Gaze distribution: The percentage of gaze points on each item of the result list calculated from the moment the user issues the first query until the first click on a result item. Prior eye tracking studies suggest that different presentations of results affect the performance differently depending on task type (Cutrell & Guan, 2007). Other studies suggest that users examine more results in exploratory tasks (White & Chandrasekar, 2010). Gaze distribution helps us to get a clearer understanding of how task types affect browsing behavior.

Finally, we use *self-reporting* to explain our findings.

Method

The purpose of this study is to collect information on search behaviors from lookup and exploratory search tasks to investigate how well we can tell apart the two task types. To this end we need a controlled experimental study. This section provides a detailed description of the experiment design.

Participants

To recruit participants we posted advertisements in the Computer Science (CS) department mailing list of the local universities. We selected researchers from the CS domain because they are the most active users of electronic bibliographic tools and web search is a major part of their daily work (Athukorala et al., 2013). Thirty-two CS researchers participated in the study. Six of them (19%) were female and 26 were male, which reflects the 20% gender distribution in the CS department of the universities we considered.

To ensure that the domain knowledge within participants is at a moderate level, we only selected researchers with some background in but who were not overly familiar with the topic of the search tasks. We provided a questionnaire to subjectively rate the familiarity with the topics of the search tasks. We selected only those who were neither actively working on any machine learning related research topic nor belong to any research group related to this area, but who have taken the introduction to machine learning course offered by the department (or an equivalent course). The median familiarity with the search domain was 2, whereas the lower and upper boundaries were 1.5 and 2.2, respectively (ratings are given in a 5-point Likert scale as 1 [*not at all familiar*] to 5 [*very familiar*]). We recruited participants at different academic levels in order to randomize the effect of research experience: 2 of the participants were in the process of writing their bachelor's thesis, 18 were MSc students, 8 were PhD students, 4 were postdoctoral researchers. The mean age of the participants was 28 years (min. age = 21 and max. = 45 years). We provided a prestudy questionnaire to assess how long they have been conducting research (*Median* = 2 years, *min* = 1.7, *max* = 4.5). Google Scholar is the primary literature search tool of 26 participants, while 4 participants use the Association for Computing Machinery (ACM) digital library and 2 use a combination of tools, including Google Scholar, ACM and the Institute of Electrical and Electronics Engineers (IEEE) digital libraries and arXiv. All the participants were experienced users of scientific literature search tools.

Task Design

We created three tasks under the exploratory category and three tasks under the lookup category. The exploratory tasks have different attributes than the lookup tasks in terms of preciseness of the search goal and objective complexity. Table 4 provides all tasks used in this study. Task definitions are given below.

- *Knowledge acquisition* tasks have *open-ended search goals*, because learning tasks have no clear criteria on when to end the search. The information-seeker could continue such a task until a subjective satisfaction level is reached (Wu, Kelly, Edwards, & Arguello, 2012). In this task, the search process has a high *objective complexity* because there is neither a definitive path to obtain the required information nor a boundary on the number of documents to be consulted. Following the characteristics of complex tasks defined by Bell and Ruthven (2004), in this task it is difficult to understand how to begin the search and interpret the relevance of the results. Li and Belkin (2010) defined similar tasks in their study under the category of intellectual work tasks with high objective complexity. This task belongs to the *core exploratory* category.
- *Planning* tasks involve gathering overviews of a new area in preparation for a future activity. Wu et al. (2012) defined such exploratory tasks as putting together elements to construct a coherent structure through planning. Planning tasks also follow a very similar pattern to knowledge acquisition tasks, yet they involve obtaining a general overview of a topic. We followed the tasks defined in similar studies to create the planning tasks (Wildemuth & Freund, 2012). This task also has a *high complexity* because many documents need to be

TABLE 4. Tasks used in the study. We created two tasks per type, that is 12 in total, and randomly assigned one task per type to every participant; every participant attempted six tasks in total (within subject design). The topic of the second task in the same type is given in brackets.

Task (Abbrev.)	Tasks
Knowledge acquisition (Know)	You are going to start a new research project on the topic *Reinforcement learning* (or *Active learning*). You would like to learn as much information as possible about this topic, e.g. applications, problems, specific algorithms
Planning (Plan)	You are planning to give a talk on the topic *Deep neural networks* (or *Clustering techniques*). Plan the structure of your presentation, including short titles of the headings of your slides and using bullet points describe the content
Comparison (Comp)	Collect literature to write a short essay describing similarities and differences between *Supervised learning and Unsupervised learning* (or *Transfer learning and Multitask learning*).
Fact-finding (Fact)	Define the term SVM (or UCB) as in the first article that proposed it.
Navigation (Navi)	Navigate to the article that presents the most commonly used topic model–latent Dirichlet allocation–for the first time (or Navigate to the article that solves the—multi-armed bandit—problem for the first time.)
Question answering (Question)	What are the most common sampling methods used in machine learning? (List three) (or What are the kernels used in machine learning? (List three))

consulted and how to find them is not straightforward. Search goals are *open-ended* because there is no clear criteria on what to find and when to end the search—belongs to the *core exploratory* category.

- *Comparison* tasks involve gathering information about two or more topics to analyze similarities and differences between them. In prior studies, similar tasks were referred to as parallel tasks or exploratory tasks with category structure (Liu & Belkin, 2010a; Navarro-Prieto et al., 1999). In this task the search goals are *open-ended* as in exploratory tasks, because there is no specific criteria on when to end the task. Yet, the task *complexity is low* compared to the other exploratory tasks, because there is a structure to the task and the search process. With respect to exploratory tasks, this task is a *borderline exploratory* task.
- *Fact-finding* tasks involve finding a specific answer to a straightforward question. We followed the structure of the closed informational tasks defined in prior studies where the information-seeker could easily decide when they found the relevant information (Aula et al., 2010). Here, the search process is *less complex* because only one fact needs to be found. The search goal is *precise* because there is a clear target and the information-seeker can judge the relevance of the results. This task belongs to the *core lookup* category.
- *Navigational* tasks involve locating a particular website or document. In navigational searching the information-seeker may just "think" a particular website/document exists and look for it (Jansen et al., 2008). As in lookup tasks, the goal is *precise* with a specific target and the search process is straightforward making the task *less complex*—it belongs to the *core lookup* category.
- *Question answering* tasks involve finding a correct set of answers where a clear list of relevant answers exists. This task is similar to the fact-finding task, yet a number of documents or sources of information need to be consulted. In accordance with lookup characteristics, this task has a *precise goal* with the ending criteria; yet, it is broader than a general lookup tasks because several documents need to be consulted and the search process is not straightforward. Aula et al. (2010) referred to tasks with similar structure as broader closed informational tasks. This task is a *borderline lookup* task.

We carefully controlled the other attributes that could affect the information search behavior: subjective task difficulty, user knowledge, and success. We set the task difficulty at a moderate level because tasks that are too easy may result in too few interactions, while tasks that are too difficult may lower the user commitment to the experiment (Bell & Ruthven, 2004). To set the subjective task difficulty at moderate level, we followed two measures. First, the task designers performed the tasks themselves and conducted a preliminary assessment of retrievability and availability of the relevant information. Second, we conducted a series of five pilot studies with 2 new participants in each study followed by task modification until all the tasks are at a moderate level of difficulty. In the pilot studies, participants rated the difficulty of tasks (on a 5-point Likert scale) with detailed explanations for the reason behind the rating. Then, we interviewed them and modified the tasks taking into consideration their explanations and run another pilot. We repeated this process until the mean difficulty rating of all the tasks fell approximately on 3.

Expert researchers in the machine learning domain designed the task. We created two tasks from each task type in order to improve the generalizability and randomly assigned one from every task type to each participant. Every participant covered all six types of search tasks—within subject design. We used the Latin square method to counter balance the order of tasks.

As the search strategies are different between successful and unsuccessful performers (Aula & Nordhausen, 2006), we decided to include only the data collected from successful performers. To this end, prior to data analysis, two expert researchers from the machine learning domain evaluated the performance of the participants. First, the experts categorized every task as a success or a failure by considering the answer. Then, they further rated the answers of the successful tasks on a 5-point Likert scale by considering the quality of the final answer and relevance of visited articles for the task.

Measures

For every result item clicked, we logged time, title, position of the article in the SERP, and the time spent on reading it. We also logged the task start and end time. The task end time was logged when the participant clicks the "done" button next to the query-typing box. We also logged the scroll data and every issued query with time. We used Tobii X2-60 Compact eye tracker to log the gaze data.

Procedure

We conducted all the studies in a controlled laboratory, with a desktop computer and 27-inch display. First, the conductor explained the purpose and the procedure of the study to the participants. We informed the participants that the purpose was "to understand the normal scientific information-seeking behaviors." Therefore, we instructed the participants to perform the search tasks as they would normally do using the search tool we provide. Further, we explained that the search tool we provided is linked to a database containing all the literature required to performing the given tasks and it has the same features as their most familiar literature search tool, Google Scholar. Next, we provided the participants with a trial task to familiarize them with the setting.

Before each task, we calibrated the eye tracker and at the end of every search task we saved the eye tracker data in a separate file. We did not restrict when and how the participants formulate search queries or links following each query. We first presented a written task description to participants to read thoroughly until they understand it. Once they were ready to start the tasks, they clicked the "start task" button to allow the system to log the start time and proceed to the search interface. We allowed the participants to take notes on an article or note pad application if necessary. They could also download or bookmark any article and browse through links as they normally do. We instructed the participants to inform us when they had completed each task; however, each task was limited to 15 minutes maximum. To keep the search process natural, we did not ask the participants to think aloud. When the participants decided that they had collected enough information for the task, they clicked the "done" button to allow the system to log the task end time. We also kept track of the time and informed them when the 15 minutes were over. While the participants were performing the task, we unobtrusively observed their search behavior and made notes of special behaviors, which we discussed with them during the interview. Then, the participants wrote their answers in a web form that we created for logging their answers. For the knowledge acquisition and comparison tasks the participants wrote an abstract of their essay.

At the end of each task we conducted a semistructured interview about their search behaviors. We compensated each participant with two movie tickets. Each study lasted approximately 90 minutes.

Results

All the participants successfully completed all tasks. Their success rates ranged from moderate to highly successful according to the expert ratings. Cohen kappa test indicated a substantial interannotator agreement between the two experts who rated the task success, Kappa = 0.72, $p < 0.01$. We excluded the data of 2 participants who scored 2 or less for more than one task. All the others received a task success score greater than or equal to 3 (out of 5) for all the six tasks. For all the six tasks the median score is 4 and the lower and upper quartile bounds are 3.0 and 4.0, respectively.

TABLE 5. Predictive Power per Feature per Task Combination. We report how significantly the data over seven features differ between every combination of Exploratory (Knowledge Acquisition, Planning, Comparison) and Lookup (Fact-Finding [Fact], Navigation [Nav], and Question Answering [Q/A]) tasks. We used a Wilcoxon signed rank test. Entries written with *s are significant after Bonferroni correction, with * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$.

Exploratory:	Knowledge Acq.			Planning			Comparison		
Lookup:	Fact	Nav	Q/A	Fact	Nav	Q/A	Fact	Nav	Q/A
				Query length					
p	**	***	.22	*	***	.53	***	.10	***
Z	2.97	4.4	1.22	1.72	4.07	.61	3.23	1.63	3.58
				Maximum scroll depth					
p	*	*	.83	**	***	.14	***	***	*
Z	2.21	1.9	.21	2.38	3.06	1.46	3.56	3.37	1.99
				Query duration					
p	***	***	***	.37	.97	.76	.14	.77	.91
Z	4.45	3.58	3.36	.89	.03	.30	1.48	.29	.12
				Proportion of browsing					
p	*	*	*	.39	.26	.57	.71	.79	.91
Z	−1.71	−1.82	−2.19	.85	1.12	.57	.37	.26	.11
				Duration dwelling					
p	***	***	*	.47	.54	.82	.19	.54	.84
Z	3.80	3.42	2.89	.71	.60	−.21	1.30	.61	−.2
				Task completion time					
p	***	***	.12	***	***	.29	***	***	***
Z	4.3	3.8	1.5	4.4	3.8	1.06	4.5	4.2	3.4
				Cumulative clicks					
p	*	*	*	.61	.56	.69	.22	.19	.59
Z	2.57	2.39	2.34	.52	.58	.40	1.23	1.30	.54

Before we pooled the two tasks of each type for the final analysis, we statistically analyzed whether there were any significant differences between the tasks of the same type. We performed Mann-Whitney U tests on the two groups for all the seven information search behaviors, and the task success scores. There was no significant difference between the two tasks for any of the analyses. This suggests that the tasks are indeed representative of their types.

Next we pooled the two tasks of each type to analyze the exploratory and lookup tasks. For statistical testing, we conducted nonparametric Friedman test on each search variable followed by pairwise comparisons between tasks using Wilcoxon signed rank tests. We performed nonparametric analysis because the data are not normally distributed. We used all data without removing any outliers to keep the prediction task realistic. All the *p* values were Bonferroni corrected. Table 5 shows the results of the pairwise comparison between each exploratory task and the lookup tasks.

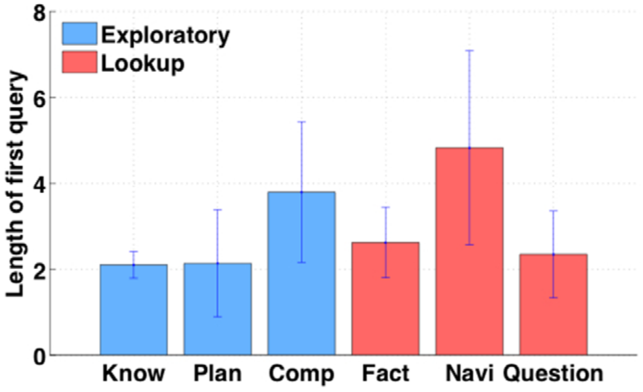


FIG. 2. First query length: Mean and standard deviation for each task type. Notice that navigational tasks have the longest queries and the two core exploratory tasks—knowledge acquisition and planning—have the shortest first queries. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

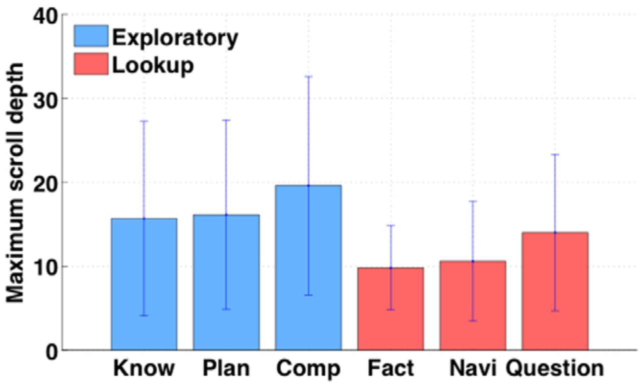


FIG. 3. Maximum scroll depth: Mean and standard deviation of maximum scroll depth in the first query. Notice that there is greater depth of scrolling in the exploratory tasks. Standard deviation is high because the data are not normally distributed—we used nonparametric analysis. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Query Length: Longer Queries in Navigational Tasks

Figure 2 shows the mean and standard deviation of query length per task. Navigational task from the lookup category has the longest first queries. The core exploratory tasks—knowledge acquisition (*Know* for short) and planning (*Plan* for short)—have the shortest first queries. The first row of Table 5 shows the significance of the difference in query lengths between each task in the exploratory category with each task in the lookup category. The queries in the core lookup tasks—fact-finding (*Fact* for short) and navigational (*Navi* for short)—are significantly longer than the queries in the core exploratory tasks. The queries in the borderline exploratory task—comparison (*Comp* for short)—are longer than that in the other two exploratory tasks. Even though the question answering (*Question* for short) task belongs to the lookup category, the queries in this task are shorter than in other lookup tasks. In summary, we can suggest that the core lookup tasks are distinguishable from the core exploratory tasks by using query lengths. The two borderline tasks from each category show a mixed behavior.

To further understand the reason for these differences, we qualitatively analyzed the first queries. In the knowledge acquisition task on the topic of reinforcement learning, the most common first query is indeed “reinforcement learning.” Similarly, in the planning tasks about the topic of deep neural networks, the most common first queries are “neural networks” or “deep neural networks” and “introduction neural networks.” This behavior is expected because during the first iterations of typical exploratory search tasks, users generally formulate vague search queries using the key terms related to the task. We see mixed behavior in the comparison task. Some participants attempted this task as the knowledge acquisition task by first trying to learn about each topic to be compared separately—the most common first query for these users for the task about comparing supervised and unsupervised learning is “supervised learning.” Yet, many other participants attempted to solve this as a typical lookup task by directly querying for similarities and differences—sample query “difference between supervised and unsupervised learning.” Typical queries used for the fact-finding and navigational tasks are: “first article defining SVM,” “what’s the first article on latent Dirichlet allocation,” respectively. Although in Figure 2 the overall mean and standard deviation of query length between fact-finding and the core exploratory tasks do not appear significant to the human eye, the difference is significant with nonparametric Wilcoxon signed rank tests, because there is a difference within each individual user. In the question-answering task about sampling methods, many users tried to first learn about the topic rather than finding the answer. This is evident from the most common first query “sampling methods.” This behavior is expected, as the domain knowledge in the topics of both exploratory and lookup tasks is at the same level. We suggest that when question answering type of lookup tasks are conducted in unfamiliar domains, even though the search goals are clear, users still follow exploratory query formulation strategies.

Maximum Scroll Depth: More Scrolling in Exploratory Tasks

Figure 3 shows the means and standard deviation for the scroll depth. In exploratory tasks users tend to scroll a lot more than in lookup tasks. There is a statistically significant difference in the scroll depth between all the exploratory tasks and the core lookup tasks—navigation and fact-finding (second row, Table 5). The borderline lookup tasks—question answering—could be distinguished only from the comparison task with a statistical significance. We can conclude that in general in all the exploratory tasks, users tend to examine more results by scrolling more than in lookup tasks.

TABLE 6. The mean duration of the first query and dwelling. The first three tasks that belong to the exploratory category have the highest mean dwelling times and query duration.

Task type	First query duration (s)	Dwelling duration (s)
Knowledge Acquisition	492.9	306.9
Planning	240.6	132.5
Comparison	219.4	121.1
Fact retrieval	116.4	50.9
Navigational	190.8	90.1
Question Answering	196.6	122.4

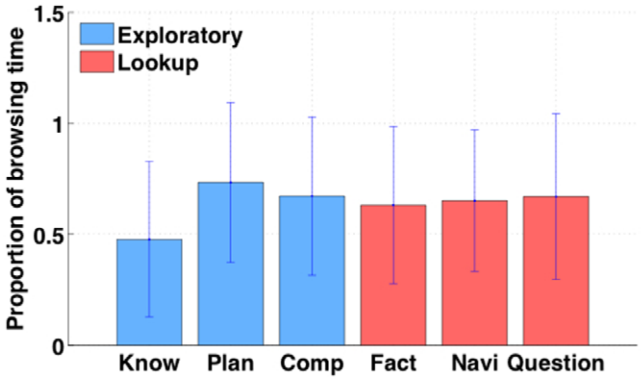


FIG. 4. Proportion of Browsing: Mean and standard deviation of the proportion of the first query duration spent on browsing the SERP. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

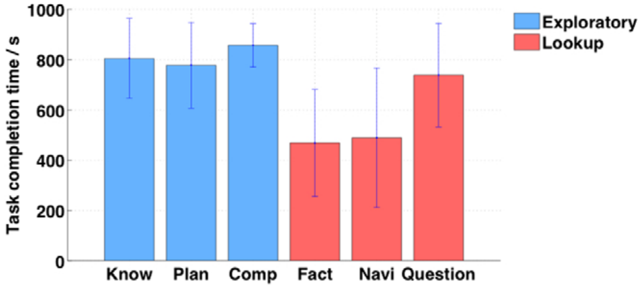


FIG. 5. Task completion times: Mean and the standard deviation. The maximum time allowed per task is 900 seconds. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

We further analyzed the scroll behavior before the first click. We did not observe any statistically significant difference in the depth of scrolling before the first click between the six search tasks, according to Friedman test $\chi^2 = 8.3$, $p = 0.08$. In both exploratory and lookup tasks users mostly focus on the top-most results prior to their first click, but after the first click in exploratory tasks they scroll deeper.

Behaviors Related to Query Time

Longest first query iteration duration in knowledge acquisition tasks (mean = 493 seconds in Table 6). The fact-finding task has the shortest query durations (mean = 116). According to the statistical analysis (third row, Table 5), we can distinguish only the knowledge acquisition task from all the lookup tasks using query duration. In planning and comparison tasks users have a higher mean query duration than in all the lookup tasks, yet this difference is not statistically significant. We suggest that in exploratory tasks in general users spend more time in the first query iteration than in lookup tasks. This behavior is most significant in the knowledge acquisition task.

Shortest proportion of browsing in knowledge acquisition tasks—Figure 4, and fourth row, Table 5. The difference is not significant between any other tasks.

Longer dwelling duration in the core exploratory tasks—knowledge acquisition and planning—than in the core lookup tasks—fact retrieval, and navigation (Table 6). According to the statistical analysis (fifth row, Table 5), only the knowledge acquisition task is significantly different from the lookup tasks. We can conclude that in exploratory tasks users spend more time examining the clicked documents than browsing search engine results pages. This behavior is prominent in the knowledge acquisition type of exploratory tasks.

Task Completion Time: Exploratory Tasks Take Longer to Complete

Figure 5 shows that the three exploratory search tasks take much longer to complete than the core lookup tasks. However, the question answering task, classified as borderline lookup task, lasts longer than the core lookup tasks. Statistical analysis confirms (sixth row, Table 5) that users take significantly longer to complete the exploratory tasks and question answering tasks than fact-finding and navigational.

In summary, task completion time is a good indicator to discriminate between exploratory and lookup search tasks. The comparison task can be easily distinguished from all the lookup search tasks. The question answering task is an outlier in the lookup category. One reason for this is that in question answering tasks, users spend more time verifying their answers even after finding the correct one. This behavior is visible from the titles of the clicked articles, issued queries, and interviews.

Cumulative Clicks: More Clicks in Knowledge Acquisition

All the exploratory search tasks have a higher mean of cumulative clicks (over 1) than the lookup tasks (Table 7). Among them, the knowledge acquisition task has the highest mean cumulative clicks. The difference is statistically significant between the knowledge acquisition task and all the lookup tasks (seventh row, Table 5). The differences between the other exploratory tasks and the lookup tasks are not statistically significant. We can conclude that cumulative

TABLE 7. The mean and maximum cumulative clicks in the first query iteration.

	Cumulative clicks	
Task type	mean	max
Knowledge Acquisition	1.8	7
Planning	1.1	5
Comparison	1.2	6
Fact retrieval	0.8	3
Navigation	0.9	4
Question Answering	0.9	3

clicks are useful to distinguish the knowledge acquisition task from the rest of the lookup tasks.

Gaze Distribution: No Differences Among The Six Tasks

During the study, we did not restrict the participants' head position or movements and tried to capture the gaze data without interfering with their natural postures. As a result, we could capture more than 70% of gaze data only for 15 participants and we excluded all the other participants from this analysis. Figure 7 shows the mean percentage of gaze points on the first 10 items on the result list before the first click.

Overall, we did not find large differences between the tasks, however, we observed subtle differences, such as in the planning tasks there is a higher percentage of gaze points on the third item. This effect can be explained by the fact that there happened to be a survey article in the third position of the results list for one planning task. Overall, the amount of attention decreases over the result list (Figures 6 and 7).

We conducted Friedman tests on the percentage of gaze points on each position on the rank list between all the tasks to confirm whether users gaze more at the first items in lookup tasks, however, the tests suggest no significant differences, $p > 0.05$.

In conclusion, the task type has no effect on the gaze distribution over the list of results until the first click.

Self-Reports

Interviews explained our findings. In the knowledge acquisition tasks users seem to click more results and spend longer time dwelling them. As 29 participants explained, they follow this strategy to get an idea about the topic: "I repetitively clicked [and read] every article that seem[ed] relevant, to get an idea" [participant 5]. Participants also explained that in exploratory tasks they did not know how to reformulate queries and continue to the second query iteration: "At the beginning I did not know how to [reformulate a] query. I just read the documents" [participant 15]. For the same reason users scrolled deeper into the first results list: "kept on scrolling and clicking anything that looks relevant and reading until I get some idea about the topic" [participant 9].

In lookup tasks there were fewer clicks, greater proportion of browsing, shorter first query iteration duration compared to the exploratory tasks. According to 23 participants, the reason is that in the core lookup tasks they can judge the relevance from the titles and reformulate queries more easily: "I know from the title [if it is correct]. So I carefully browsed the results [SERPs] reading abstracts and titles" [Participant 14], "I only [click to] read the most relevant document." [participant 30], "If the top results don't have it, I change the query" [participant 11].

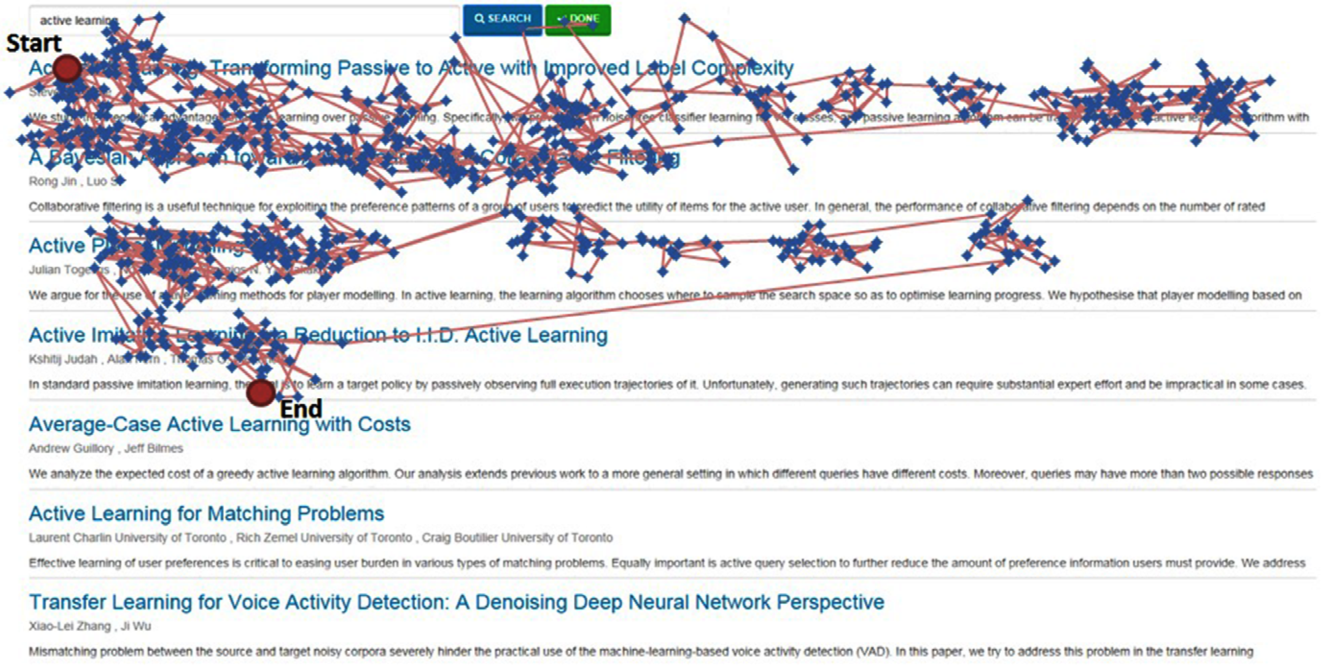
In many instances, the question answering task was very similar to comparison and planning search tasks. According to 18 participants in the question answering tasks they repeatedly searched for information to verify their answers: "I tried to confirm it is correct" [participant 13]. We further investigated their search queries to confirm that this is true. We found a repetitive use of the same query. For example, in one of the question answering about finding sampling methods, one participant issued the following queries—[q1] sampling methods, [q2] uniform sampling, [q3] Gibbs sampling, [q4] uniform sampling, [q5] sampling. This behavior shows that the question answering task involves two of the search tasks given in Marchionini's framework: question answering and verification.

Validation Through Classification

To evaluate whether our results are applicable for distinguishing exploratory from lookup tasks in a real IR system, we performed classification experiments using state-of-the-art machine-learning methods. We excluded the gaze data because there is no significant difference in gaze distribution between the tasks and, gaze-tracking data are not commonly available to IR systems. We excluded task completion time, as clearly it will not be available to the system while the user is searching. Otherwise, we included all the other behaviors we measured. To evaluate how well the classification results generalize, we use 10-fold cross-validation to perform all classification experiments. We ran all experiments using WEKA (Witten & Frank, 2005), and report the average scores over 10 independent runs. As the classifier we used Random Forests—a powerful technique providing state-of-the-art classification accuracies.

We first investigated whether we can predict the task category, exploratory or lookup, given the core exploratory tasks—knowledge acquisition, and planning—and the core lookup tasks—fact-finding, and navigational. We found that the task types can be predicted with 85% accuracy, obtaining an AUC (Area-Under-the-ROC-Curve) of 0.859 when using only the core tasks—we beat the baseline, 50% and 0.5, respectively, by a clear margin.

Next, we considered all the six tasks to predict the task category. We now obtained a 60.3% accuracy and AUC of 0.658. Compared to the previous task the prediction



Ac[illegible] Transforming Passive to Active with Improved Label Complexity

Steve[illegible]

We stu[illegible] advantage [illegible] learning over [illegible]. Specifically[illegible] noise-free classifier learning [illegible] classes, [illegible] passive learning [illegible] can be [illegible] active [illegible] algorithm with

A Bayesian Approach to[illegible] Filtering

Rong Jin , Luo S[illegible]

Collaborative filtering is a useful technique for exploiting the preference patterns of a group of users to predict the utility of items for the active user. In general, the performance of collaborative filtering depends on the number of rated

Active P[illegible]

Julian Toge[illegible]

We argue for the use of a[illegible] methods for player modelling. In active learning, the learning algorithm chooses where to sample the search space so as to optimise learning progress. We hypothesise that player modelling based on

Active Imit[illegible] Reduction to I.I.D. Active Learning

Kshitij Judah , Ala[illegible] , Thomas [illegible]

In standard passive imitation learning, the [illegible] to learn a target policy by passively observing full execution trajectories of it. Unfortunately, generating such trajectories can require substantial expert effort and be impractical in some cases.

Average-Case Active Learning with Costs

Andrew Guillory , Jeff Bilmes

We analyze the expected cost of a greedy active learning algorithm. Our analysis extends previous work to a more general setting in which different queries have different costs. Moreover, queries may have more than two possible responses

Active Learning for Matching Problems

Laurent Charlin University of Toronto , Rich Zemel University of Toronto , Craig Boutilier University of Toronto

Effective learning of user preferences is critical to easing user burden in various types of matching problems. Equally important is active query selection to further reduce the amount of preference information users must provide. We address

Transfer Learning for Voice Activity Detection: A Denoising Deep Neural Network Perspective

Xiao-Lei Zhang , Ji Wu

Mismatching problem between the source and target noisy corpora severely hinder the practical use of the machine-learning-based voice activity detection (VAD). In this paper, we try to address this problem in the transfer learning

FIG. 6. One instance of a gaze scan path. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

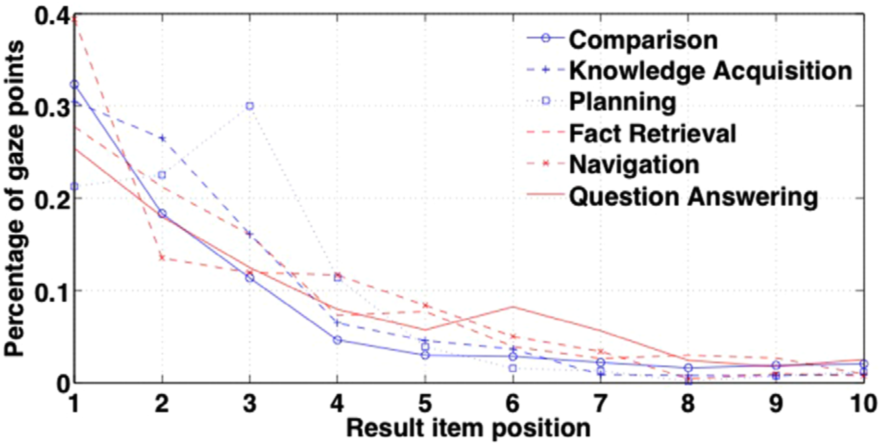


FIG. 7. Mean percentage of gaze points on the first 10 documents before the first click. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

accuracy has dropped from 85% to 60.3% after the two borderline tasks, comparison, and question answering, are included. It suggests that an IR system can easily predict core exploratory tasks, however, when a task with borderline characteristics is included it becomes difficult for the system to predict it.

We also analysed whether the borderline tasks are more difficult to predict. We evaluated how well the specific task type could be predicted out of all six tasks. We obtained a 34.6% accuracy, which, given the baseline of 16.7%, is quite high. For the core exploratory tasks—knowledge acquisition (AUC = 0.793), planning (AUC = 0.626)—and the core lookup tasks—fact-finding (AUC = 0.722) and navigational (AUC = 0.745)—we obtained rather high AUC values. For the borderline exploratory and lookup tasks, however, the AUC values are considerably lower with 0.528 for comparison and 0.573 for question answering. These results confirm that user behavior for the core tasks is much easier to keep apart than for the borderline tasks.

Next, we investigated to which of the two main categories the user behavior for the borderline tasks comes closer. That is, can we predict the task category more easily by *swapping* the labels of the borderline task types from "lookup" to "exploratory?" We swapped the labels of both borderline tasks—question answering we labeled as exploratory and comparison as lookup. This increased the accuracy from 60.3% to 72.4%, and the AUC from .658 to .741. This leads to the question, what task is actually reducing the classifier accuracy. We conducted the same analysis but swap only one label. First, we swapped the label of comparison task from "exploratory" to "lookup." We trained a classifier over all six tasks, and obtained an accuracy of 69.2% with an AUC of 0.681. Although not as good as swapping both labels, this is a good improvement over the original label. This 9% increase in accuracy suggests that the comparison task has some characteristics of the lookup class.

Finally, we swapped the label of the question answering task from "lookup" to "exploratory." We obtained a 75.6% accuracy and AUC of 0.777. This is a considerable increase over the 60.3% accuracy obtained over six tasks with the original labels, and an improvement over the case where we swapped both labels. By swapping the labels the accuracy of the classifier comes surprisingly close to the 85% obtained on just the *four* core tasks—this is particularly impressive noting that these results are cross-validated; we consider how well the effect of swapping the labels generalize. Although this result does not conclusively show that question answering should be considered an exploratory rather than a lookup task, it shows that users, in our setting, exhibit behavior that comes closer to that which they exhibit on the core exploratory tasks.

Discussion and Conclusions

This article contributes by characterizing user behavior in exploratory search tasks in the widely used conceptual framework of Marchionini (2006). Taking insights from prior work (Liu et al., 2010b), we first operationalized exploratory and lookup categories using two facets: preciseness of the search goal and objective complexity. Core exploratory tasks have open-ended search goals and high objective complexity, and core lookup tasks have precise search goals and low objective complexity. We empirically validated that IR systems can distinguish exploratory tasks within the first query session by various information search behaviors, including length of the first query, scroll depth, first query iteration duration, proportion of browsing, dwell time, and task completion time. Here, we synthesize how facets that operationalize the tasks (Table 3) relate to these information search behaviors (Table 5).

The length of the first query shows that in core lookup tasks users issue longer queries than in the core exploratory tasks. According to existing literature, when the information need is specific, search queries become longer (Phan, Bailey, & Wilkinson, 2007). In the core lookup tasks, information need is very specific because of precise search goals. On the other hand, the borderline exploratory task—comparison—has longer search queries than the core exploratory tasks. This relates to the low complexity of this task, which gives a structure to the search process. Navarro-Prieto et al. (1999) referred to similar tasks as exploratory tasks with category structure. Their analysis explains that when there is some structure to exploratory tasks, users follow mixed strategies of searching for specific information and information general to the topic. Similarly, in the borderline lookup task—question answering—which has high objective complexity, users issue shorter queries because they need to find several documents. Hence, they issue shorter general queries that retrieve many related documents.

Scroll depth analysis suggests that in all three exploratory tasks users scroll significantly more than in lookup tasks. This behavior relates to the open-ended search goals associated with exploratory tasks making it difficult to judge the relevance of the document. J. Kim (2009) reported that in exploratory web search tasks users prefer web pages with lots of links. Self-reports further confirm that as the search goals are open-ended users first attempt to gain an overview of the topics of the task by examining as many items as possible. For the same reason, all three exploratory tasks involve a higher number of clicks than the lookup tasks. However, in the borderline lookup tasks users also scroll more than in core lookup tasks. This is due to the high objective complexity—users need to find many documents and hence they scroll more.

We measured three indicators related to the query duration: first query duration, proportion of browsing, and dwelling duration. The knowledge acquisition task has the longest first query iteration duration because it has the least descriptive search goals: The user is asked to learn about a topic without being given any criteria what to search for. Similar behavior is reported in other studies (White & Roth, 2009). In the core exploratory tasks, users spend more time dwelling or reading clicked documents. But in the borderline tasks we see a mixed behavior. Variation in objective complexity is the reason for these differences. In the comparison task, the information-seeker can quickly skim through the documents and find similarities and differences. But in the learning-oriented exploratory tasks they need to read the documents more thoroughly to get an understanding, which results in longer dwelling time.

As we expected, exploratory tasks took longer to complete (Marchionini, 2006; White & Roth, 2009). However, in the borderline lookup tasks users also spent more time than in other lookup tasks. This relates to the high complexity of this task. As the query analysis suggested, users spent more time finding answers and verifying them. White and Drucker (2007) also showed that over 80% of web search tasks indicate similar borderline behavior because they involve some degree of exploration.

Regardless of the search task, users mostly gaze at articles at the top of the results list before their first click. Prior work on lookup tasks indicates that users are biased towards results ranked higher in a list, even when their relevance is low (Joachims et al., 2005). We confirm that this happens also in exploratory search tasks, even when users are doubtful about their own query. On the other hand, users scroll deeper into the result list in exploratory tasks after the first click. This suggests that users need time to read and analyze the links at the top to realize that they cannot depend only on the top ranked results.

To sum up, the data elaborate on the original conceptual classification proposed by Marchionini (2006), by proposing information search behaviors that could help to detect at a fine granularity when an exploratory task might take lookup behavioral aspects and vice versa. The information search behaviors we propose provide guidelines on how to distinguish search tasks while the user is still performing the search. This allows the search systems to predict the task type early and adapt its support (Shah, Hendahewa, & González-Ibáñez, 2015).

Implications for IR Systems

Our findings have important implications for the design of IR systems. Our classification analysis shows that the outcome is actionable. Here, we propose how the task type prediction performed by a classifier using the information search behaviors we analyzed is used for tailoring and adapting IR systems. We inform three aspects of IR systems that can be tailored: interface design, retrieval algorithm design, and user model design.

Adjusting the number of result items shown per SERP. Many search engines today provide a constant number of results per page—typically 10 items. We confirm that in all exploratory tasks users scroll deeper into the results lists than in lookup tasks. This result shows the importance of tailoring the length of the result list for exploratory tasks. Prior studies also suggest that in exploratory tasks users prefer to examine more items (Athukorala et al., 2014; J. Kim, 2009). We propose a future search interface that shows a longer list of items in the first SERP. This allows the system to use the maximum scroll depth as a behavioral measure. Once we predict the task type, from the second iteration onwards the interface adapts to show fewer items if the task is lookup and increase the number of displayed documents if the task is exploratory.

Adjusting the length of results snippet according to task type. An open problem relating to snippet length is the trade-off between showing long informative summaries and minimizing screen space allocated per result item. Research suggests that the ideal snippet length depends on the task type (Cutrell & Guan, 2007). In tasks where the search goals are imprecise and oriented towards learning, longer snippets improve performance, whereas in navigational tasks longer snippets degrade the performance. According to our results, in core lookup tasks users spend a longer time dwelling or reading content than browsing SERPs. We propose a search interface that automatically increases the snippet length for exploratory search tasks. Paek, Dumais, and Logan (2004) also showed that users prefer search interfaces that dynamically increase the snippet length. Further research is needed to identify the ideal snippet length for different tasks.

Adapting implicit relevance feedback techniques according to task type. Implicit relevance feedback algorithms use information derived as a byproduct of information search behaviors, such as query suggestions, to provide more support and automatically retrieve new documents (Agichtein, Brill, & Dumais, 2006). However, the task type has a significant influence on search behaviors, such as dwell time (Kelly & Belkin, 2004; White & Kelly, 2006). Information about the task type can greatly improve the accuracy of implicit relevance feedback (Joachims et al., 2005). The classification investigation we perform shows how an IR system could predict the task type, while the user is still at the first SERP. This allows the implicit relevance feedback techniques to adapt to the task type and provide more informative results in the second SERP.

Adjusting exploration rate according to task type. Making a trade-off between exploration—making the results more diverse by including alternative topics—and exploitation—making the results narrower by including very specific sub-topics—is a known problem in machine learning. Studies show that in exploratory search tasks, users benefit from exploring more diverse topics than exploiting narrower sub-topics (Athukorala et al., 2014; Głowacka et al., 2013). It would be useful for the user if IR algorithms could adapt the parameters that decide the right balance between exploration and exploitation. Our findings have useful implications for such algorithms. They can predict the task type from the proposed information search behaviors. If the task is exploratory, these algorithms can increase the level of exploration to retrieve more diverse topics and if the task is lookup, they can increase the level of exploitation to retrieve narrower results.

Acknowledgments

Jilles Vreeken is supported by the Cluster of Excellence "Multimodal Computing and Interaction" within the Excellence Initiative of the German Federal Government. The work of Antti Oulasvirta is funded by the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 637991) and the Academy of Finland project COMPUTED. Dorota Głowacka is supported by The Finnish Funding Agency for Innovation (under projects Re:Know and D2I) and by the Academy of Finland (under the Finnish Centre of Excellence in Computational Inference). This work has been partly supported by MindSee (FP7 – ICT; Grant Agreement # 611570). We thank Dr. Antti Salovaara for reading the draft and providing insightful comments, and Kalle Ilves and Imtiaj Ahmed for helping to set up the user study.

References

Agichtein, E., Brill, E., & Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. In *SIGIR* (pp. 19–26).

Athukorala, K., Hoggan, E., Lehtiö, A., Ruotsalo, T., & Jacucci, G. (2013). Information-seeking behaviors of computer scientists: Challenges for electronic literature search tools. In *ASIST* (pp. 1–11).

Athukorala, K., Oulasvirta, A., Głowacka, D., Vreeken, J., & Jacucci, G. (2014). Narrow or broad? Estimating subjective specificity in exploratory search. In *CIKM* (pp. 819–828).

Aula, A., & Nordhausen, K. (2006). Modeling successful performance in web searching. JASIST, 57(12), 1678–1693.

Aula, A., Khan, R.M., & Guan, Z. (2010). How does search behavior change as search becomes more difficult? In *CHI* (pp. 35–45).

Belkin, N.J. (2008). Some (what) grand challenges for information retrieval. ACM SIGIR Forum, 42(1), 47–54.

Bell, D.J., & Ruthven, I. (2004). Searchers assessments of task complexity for web searching. In *ECIR* (pp. 57–71).

Brand-Gruwel, S., Wopereis, I., & Vermetten, Y. (2005). Information problem solving by experts and novices: Analysis of a complex cognitive skill. Computers in Human Behavior, 21(3), 487–508.

Broder, A. (2002). A taxonomy of web search. ACM SIGIR Forum, 36(1), 3–10.

Byström, K. (2002). Information and information sources in tasks of varying complexity. JASIST, 53(7), 581–591.

Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. Information Processing & Management, 31(2), 191–213.

Cutrell, E., & Guan, Z. (2007). What are you looking for?: An eye-tracking study of information usage in web search. In *CHI* (pp. 407–416).

Diriye, A.M. (2012). Search interfaces for known-item and exploratory search tasks Unpublished doctoral dissertation. UCL (University College London).

Głowacka, D., Ruotsalo, T., Konyushkova, K., Athukorala, K., Kaski, S., & Jacucci, G. (2013). Directing exploratory search: Reinforcement learning from user interactions with keywords. In *IUI* (p. 117–128).

Hassan, A., White, R.W., Dumais, S.T., & Wang, Y. (2014). Struggling or exploring?: Disambiguating long search sessions. In *WSDM* (pp. 53–62).

Hölscher, C., & Strube, G. (2000). Web search behavior of internet experts and newbies. Computer Networks, 33(1), 337–346.

Ingwersen, P., & Järvelin, K. (2006). The Turn: Integration of Information Seeking and Retrieval in Context (Vol. 18). Dordrecht, Netherland: Springer.

Jansen, B.J., & Pooch, U. (2001). A review of web searching studies and a framework for future research. Journal of the Association for Information Science and Technology, 52(3), 235–246.

Jansen, B.J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. Information Processing & Management, 36(2), 207–227.

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

Jenkins, C., Corritore, C.L., & Wiedenbeck, S. (2003). Patterns of information seeking on the web: A qualitative study of domain expertise and web expertise. IT & Society, 1(3), 64–89.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *SIGIR* (pp. 154–161).

Kelly, D., & Belkin, N.J. (2004). Display time as implicit feedback: Understanding task effects. In *SIGIR* (pp. 377–384).

Kim, J. (2009). Describing and predicting information-seeking behavior on the web. JASIST, 60(4), 679–693.

Kim, K.-S. (2001). Information-seeking on the web: Effects of user and task variables. Library & Information Science Research, 23(3), 233–255.

Kules, B., Wilson, M., Schraefel, M.C., Shneiderman, B. (2008). *From keyword search to exploration: How result visualization aids discovery on the web* (Tech. Rep. No. HCIL-2008-06). University of Maryland.

Li, Y., & Belkin, N.J. (2008). A faceted approach to conceptualizing tasks in information seeking. Information Processing & Management, 44(6), 1822–1837.

Li, Y., & Belkin, N.J. (2010). An exploration of the relationships between work task and interactive information search behavior. JASIST, 61(9), 1771–1789.

Liu, J., & Belkin, N.J. (2010a). Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. In *SIGIR* (pp. 26–33).

Liu, J., Cole, M.J., Liu, C., Bierig, R., Gwizdka, J., Belkin, N.J., & Zhang, X. (2010b). Search behaviors in different task types. In *JCDL* (pp. 69–78).

Liu, J., Liu, C., Gwizdka, J., & Belkin, N.J. (2010c). Can search systems detect users' task difficulty?: Some behavioral signals. In *SIGIR* (pp. 845–846).

Marchionini, G. (1989). Information-seeking strategies of novices using a full-text electronic encyclopedia. JASIST, 40(1), 54–66.

Marchionini, G. (2006). Exploratory search: From finding to understanding. Com. ACM, 49(4), 41–46.

Navarro-Prieto, R., Scaife, M., & Rogers, Y. (1999). Cognitive strategies in web searching. In *Human Fact. & the Web* (pp. 43–56).

Paek, T., Dumais, S., & Logan, R. (2004). Wavelens: A new view onto internet search results. In *CHI* (pp. 727–734).

Palmquist, R.A., & Kim, K.-S. (2000). Cognitive style and on-line database search experience as predictors of web search performance. JASIST, 51(6), 558–566.

Phan, N., Bailey, P., & Wilkinson, R. (2007). Understanding the relationship of information need specificity to search query length. In *SIGIR* (pp. 709–710).

Pirolli, P., & Card, S. (1995). Information foraging in information access environments. In *CHI* (pp. 51–58).

Rose, D.E., & Levinson, D. (2004). Understanding user goals in web search. In *WWW* (pp. 13–19).

Saito, H., & Miwa, K. (2001). A cognitive study of information seeking processes in the WWW: The effects of searcher's knowledge and experience. In *WISE* (Vol. 1, pp. 321–327).

Shah, C., Hendahewa, C., & González-Ibáñez, R. (2015). Rain or shine? forecasting search process performance in exploratory search tasks. Journal of the Association for Information Science and Technology. doi:10.1002/asi.23484.

Thatcher, A. (2008). Web search strategies: The influence of web experience and task type. Information Processing & Management, 44(3), 1308–1329.

Vakkari, P. (2003). Task-based information searching. Annual Review of Information Science and Technology, 37(1), 413–464.

White, R.W., & Chandrasekar, R. (2010). Exploring the use of labels to shortcut search trails. In *SIGIR* (pp. 811–812).

White, R.W., & Drucker, S.M. (2007). Investigating behavioral variability in web search. In *WWW* (pp. 21–30).

White, R.W., & Kelly, D. (2006). A study on the effects of personalization and task information on implicit feedback performance. In *CIKM* (pp. 297–306).

White, R.W., & Roth, R.A. (2009). Exploratory Search: Beyond the Query-response Paradigm. San Rafael, CA: Morgan & Claypool.

Wildemuth, B.M., & Freund, L. (2012). Assigning search tasks designed to elicit exploratory search behaviors. In *HCIR* (pp. 1–10).

Witten, I.H., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann.

Wu, W.-C., Kelly, D., Edwards, A., & Arguello, J. (2012). Grannies, tanning beds, tattoos and nascar: Evaluation of search tasks with varying levels of cognitive complexity. In *IIiX* (pp. 254–257).