User Activity Patterns During Information Search

MICHAEL J. COLE, CHATHRA HENDAHEWA, NICHOLAS J. BELKIN,
and CHIRAG SHAH, Rutgers University

Personalization of support for information seeking depends crucially on the information retrieval system's knowledge of the task that led the person to engage in information seeking. Users work during information search sessions to satisfy their task goals, and their activity is not random. To what degree are there patterns in the user activity during information search sessions? Do activity patterns reflect the user's situation as the user moves through the search task under the influence of his or her task goal? Do these patterns reflect aspects of different types of information-seeking tasks? Could such activity patterns identify contexts within which information seeking takes place? To investigate these questions, we model sequences of user behaviors in two independent user studies of information search sessions (N = 32 users, 128 sessions, and N = 40 users, 160 sessions). Two representations of user activity patterns are used. One is based on the sequences of page use; the other is based on a cognitive representation of information acquisition derived from eye movement patterns in service of the reading process. One of the user studies considered journalism work tasks; the other concerned background research in genomics using search tasks taken from the TREC Genomics Track. The search tasks differed in basic dimensions of complexity, specificity, and the type of information product (intellectual or factual) needed to achieve the overall task goal. The results show that similar patterns of user activity are observed at both the cognitive and page use levels. The activity patterns at both representation layers are able to distinguish between task types in similar ways and, to some degree, between tasks of different levels of difficulty. We explore relationships between the results and task difficulty and discuss the use of activity patterns to explore events within a search session. User activity patterns can be at least partially observed in server-side search logs. A focus on patterns of user activity sequences may contribute to the development of information systems that better personalize the user's search experience.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*

General Terms: Performance, Design, Human Factors

Additional Key Words and Phrases: Task, cognitive effort, user study, personalization, information search behavior

ACM Reference Format:
Michael J. Cole, Chathra Hendahewa, Nicholas J. Belkin, and Chirag Shah. 2015. User activity patterns during information search. ACM Trans. Inf. Syst. 33, 1, Article 1 (March 2015), 39 pages.
DOI: http://dx.doi.org/10.1145/2699656

This work was supported by a Google Faculty Research Award to N.J. Belkin and C. Shah and by IMLS grant LM-06-07-0105-07. Parts of this article were presented at the 2014 ACM SIGIR Conference.

Authors' addresses: M. J. Cole, N. J. Belkin, and C. Shah, Department of Library and Information Science, Rutgers, The State University of New Jersey, 4 Huntington Street, New Brunswick, NJ 08901-1071, USA; emails: {m.cole, belkin, chirags}@rutgers.edu; C. Hendahewa, Department of Computer Science, Rutgers, The State University of New Jersey, 110 Frelinghuysen Road, Piscataway, NJ 08854-8019, USA; email: chathrah@cs.rutgers.edu.

DOI: http://dx.doi.org/10.1145/2699656

1. INTRODUCTION

Interactive information search is rooted in complex user behaviors derived from highly conditionalized decisions. The user's type of motivating task (the task that led the user to engage in information search) affects his or her information behaviors during search. Decisions about acquiring and using information to make progress toward the user's task goal depends in part on complex mental states, including the user's level of knowledge about the task both in the sense of his or her knowledge about how to do the task and his or her knowledge of the domain, the nature of the task, and the task goal.

Information behaviors are expressed in sequences of user actions. They reflect the user's search intent and the plan to carry out that intent. The sequence of user intention–plan behavior pairs for a task session constitutes a record of the user's attempts to achieve the task goal. It is intuitive that properties of the task will influence the user's behaviors to solve the task. User studies show that the nature of a user's task affects his or her information-seeking behavior. Specific task properties that have been shown to affect search behavior include complexity and difficulty, and there are also task stage effects on user usefulness and relevance judgments [Byström and Järvelin 1995; Byström 2002]. Taking account of task type has also been found to improve implicit relevance feedback based on a user's dwell time on documents [White and Kelly 2006].

By activity, we mean the immediate flow of users' actions as they work to achieve their overall task goal. A user's moment-to-moment observable activity can be characterized as patterns of actions while interacting with the information objects. These actions usually track allocation of attention, and this level of representation relates to the flow of user experience during the task session.

A search behavior is a class made of instances of action sequences that in some related manner are used to make progress toward a search task goal. Observations of search interaction can be made at multiple levels. Information science research has often looked at page interactions. It is common to consider the page type or content, dwell time, and actions such as scrolling or click-through to another document. Importantly, this level of observation is available to some degree in server-side search logs or can be reported to a search engine by suitably instrumented web browsers. Such modeling of search logs is a conventional approach to server-side personalization.

Human search is fundamentally a cognitive activity, and attempts to capture such cognitive activity of a searcher can provide another level of representation for user search behavior. One observable cognitive activity is eye movement patterns. Sequences of eye fixations provide a low-level behavioral observation of interactive tasks. For example, Triesch et al. [2003] showed that people fixate briefly on objects needed in the next step of a task. Reading eye movements corresponds to an essential step in the (textual) information acquisition process [Rayner and Pollasek 1989; Findlay and Gilchrist 1998]. Eye fixation and movement observations are well suited to represent the textual information acquisition process during search because reading requires fixation on word glyphs and eyes remain fixated until the meaning of the word is acquired [Rayner 1998].

Inferring intent by eye movement has been used for HCI [Salvucci 1998]. More generally, visual search is mediated by action intentions [Bekkering and Neggers 2002].

One of our research goals is to learn users' task types using observations of their information behaviors during a search task session. We hypothesize that a system can exploit such knowledge to interpret observable implicit indicators of usefulness, relevance, and so forth. For example, implicit relevance feedback using query expansion or reranking usually depends on calculating the similarity of the query content to documents. Personalization could also take account of the user's current information-seeking goal and the goal that drives his or her whole session. The user's sequence of behaviors during a task session is one place to look for evidence of his or her current intent and overall task goal.

Association of task type with sequences of observed behaviors could allow a system to handle the document collection differently in query expansion and document reranking. The system could take account of the expected usefulness of the document for the task type, for instance, in the difference between high-recall tasks and aspectual-retrieval tasks [Dumais and Belkin 2005], or factual and process-oriented tasks [Murdock et al. 2007]. Personalization based on this level of representing user information search strikes in a new direction for system user support, because it is grounded in user actions and intent rather than document similarities.

The goal of an information system is to help users achieve their task goals. The focus of information system design is to make information resources, including document collections, available to users and to provide support during the search so that the users will be able to find the information they need to achieve their motivating task goals. A basic challenge for study of user actions in information search tasks is that essential parts of the task session are not observable. User learning and decision making occur during most search sessions and can be essential to solve the user's task challenge. Such mental events are a crucial piece of the search interaction picture. Further, in order to build systems that better support the user, it is important to acknowledge that information retrieval happens in the larger context of the user's final task, rather than just his or her search task [Belkin 2008]. This goal informs and influences user decision making and search intentions, and all of this essential apparatus of search is not observable directly.

To further the general research goal of improving personalization based on expression of user search behaviors, we focus on the following research questions in this article.

—RQ1: Can we identify aspects of a user's task type by observing his or her search interaction behaviors/activities?
—RQ2: Is observation of user activity/behavior patterns effective in distinguishing between tasks and characteristics of tasks?
—RQ3: Does the approach generalize across multiple datasets to allow differentiation among task types and task characteristics in different contexts?

In order to address the aforementioned research questions, we present a technique to represent user activity patterns and show relationships between properties of the activity patterns, task types, and one task characteristic, difficulty. The main contributions of this work are as follows:

—Further development of a user-centred representation of information search interaction based on the activity patterns of the user during search.
—Presentation of a technique to explore relationships between aspects of user information states that are not (easily) observed and intentional user activities of page selection and use that have traditionally been used to characterize search actions.
—Presentation of results showing that task types and task difficulty can be distinguished by properties of user activity patterns, in particular by the complexity of the patterns and the distribution of activity states.

These contributions suggest concrete ideas for algorithms that accept sequences of user actions and apply learned models to infer aspects of the user's guiding task. One possibility is that a system could take action to help the user based on some understanding of the guiding task and improve the performance of the system in terms of both the user's time on task and search quality. Another is that such models could serve as user-behavior models for simulation of search sessions for different task types.

The rest of the article is organized as follows. Section 2 presents the background and related work with respect to task categorization, eye movements, time-based sequence analysis, and Markov chains and graph properties. Then we explain the important details about the user studies conducted to capture user search activities in Section 3. Section 4 provides a detailed explanation of the methodology used for analysis, including high-level (page sequences) analysis and low-level (eye movements) analysis. The results obtained by conducting the experiments on our user study datasets are presented in Sections 5 and 6. Section 7 discusses the findings and how they address the research questions being raised. We also discuss our methodology and its benefits, drawbacks, and potential to advance personalization of search, especially in the context of extended or exploratory search. Finally, we conclude the article in Section 8.

2. BACKGROUND AND RELATED WORK

2.1. Task Categorization

Task classification and the effects of task on user search behaviors have been examined in many studies. Marchionini [1989] found that time on task and numbers of actions were positively correlated with open-ended tasks versus tasks that had only one correct answer. Qiu [1993] found that users adopted less structured search patterns in general tasks. Kim and Allen [2002] found that individual cognitive differences interacted with task variables to affect search behaviors but not search performance. Gwizdka and Spence [2006] showed that subjective task difficulty was influenced by the number of web pages seen, the dwell time on a page, and the complexity of the navigation path, and that the relative importance of those factors depended on objective task complexity. Kellar et al. [2007] showed that information gathering had the most complex interactions of four types of tasks: fact finding, information gathering, browsing, and transactions. Li [2008] found that intellectual tasks involved more effort, measured by number of systems used and pages consulted and query length, as compared to decision/solution tasks. These various results are difficult to interpret because of the variation in definitions of task-type classification. Also, such user behavior research has evaluated whole-session task-type influences rather than considering them by task session subunits, for example, sequences of page interactions. In practical applications, it is desirable to detect the user's current behavior in order to personalize the ongoing interaction.

Li [2009] provides a classification system, with 15 facets or subfacets, for tasks that identifies and integrates the various aspects of task. This allows construction of instances of task types for user studies that can be, in principle, distinguishable by facet values. User studies can be designed to observe associations between dependent behavioral variables and a small set of topic-independent task variables.

2.2. Eye Movements, Reading, and Information Acquisition

Eye movement is a sequence of *fixations* at a location with very little variance followed by a rapid movement, a *saccade*, to the next fixation. Little, if any, signal is acquired by the person during a saccade, and one's conscious perception of the world is constructed from the data acquired during fixations.

In information search tasks, it is typical for users to address their needs via textual information acquisition, that is, the reading process. The cognitive activity in reading revolves around processes of word meaning access, sentence and phrase parsing, and text comprehension. Reading eye movement patterns have long been studied [Rayner 1998]. Research has established relationships between eye movements and eye fixation properties and semantic and cognitive processing states. It is known that eye movements are cognitively controlled [Findlay 2003]. Competing models of reading eye movement patterns disagree mostly about the object of the cognitive control. Morrison [1984] proposed that words were both the perceptual unit for cognitive processing and, as orthographic objects, the targets for the next saccade. This approach to handling the saccade programming problem is the source of a widely accepted class of models, especially as developed by Reichle et al. [2004]. These models have been able to explain observed fixation duration and word-skipping behaviors.

Cognitive capabilities of individuals appear to have some effect on reading behaviors. Hyönä [2002] measured working memory and had participants read and summarize a text passage. Four clusters of eye fixation patterns were found to be associated with reading the text. Distinguishing cluster features included the probability of refixing on a previous sentence versus look-ahead to an upcoming sentence, and attention to structural text, such as headlines and subject headings. Those with high working memory had a bias to process structural text and produced the best summaries. This work both shows individual differences and provides evidence for the influence of cognitive strategies on reading patterns to acquire textual information.

The potential to apply reading models to personalize information search systems has been examined. Buscher et al. [2008] developed an eye-tracking-based model of information acquisition processing by distinguishing "reading" versus "skimming" based on the separation of succeeding eye fixations. Using the labeled regions of text, they found better implicit query expansion performance when using words from "reading" fixation sequences. Cole et al. found relationships between eye movement patterns and high-level task features, such as number of documents examined and document use, when entire task sessions were considered [2010, 2011a], and with individual differences in task domain knowledge [2012]. Jiang et al. [2014] show that eye fixations, scan paths, query characteristics, and clicking actions as behaviors change over the course of complex search sessions and that the relationships are affected by task.

Cole et al. [2011a] implemented a line-oriented reading eye movement pattern classification system based on Reichle [2004] to calculate cognitive processing features of user reading eye movement patterns during information search tasks. The same eye movement analysis system is used in the present work and explained in greater detail later.

2.3. Time-Based Sequence Analysis

Some research has exploited temporal information embedded in documents for presentation, clustering, and exploration of search results [Alonso 2007]. Yue et al. [2012] and Han et al. [2013] used session-based sequence and temporal measures to classify user action sequences using Markov chains and a Hidden Markov Model (HMM), in order to infer user latent tactics as described in Marchionini's information-seeking process model [Marchionini 1995].

Hendahewa and Shah [2013] show that analyzing the search process based on sequences of actions performed by a user during an online search task can be used to identify different stages of the search process. They looked at user actions of content page visits, querying, bookmarking, and snipping and constructed user action sequences to learn activity type clusters across the search process that are linked to different task stages. In this article, we use the same basic technique to make and analyze cluster sequences to distinguish among different task types because we hypothesize that task-type differences and task difficulty influence the structure of user activities during the task session.

2.4. Task Difficulty

The topic of assessing information task difficulty has received widespread attention [Byström and Järvelin 1995; Byström 2002; Kim and Rieh 2005; Gwizdka and Spence 2006; Kim 2006; Toms et al. 2007; Gwizdka 2008; Liu et al. 2010; Aula et al. 2010]. Task difficulty can be assessed by objective measurements, such as time on task and number of queries or pages used, and by subjective measurements, for example, user self-assessments of perceived difficulty.

Task difficulty has been found to be correlated with multiple aspects of user task performance. These include search effort and efficiency [Gwizdka and Spence 2006; Li and Belkin 2010]. Task difficulty is also correlated with user characteristics, such as cognitive style and ability [Kim and Allen 2002; Kim 2006; Gwizdka 2008; Aula et al. 2010].

Another research focus in this work is the effect of task difficulty on search behavior. Objective measures of task events and actions correlate with difficulty. Difficult tasks make it more likely that users will access more web pages [Kim 2006; Gwizdka 2006], issue more queries [Kim 2006; Aula et al. 2010], and dwell longer on search engine result pages (SERPs) [Aula et al. 2010].

Relationships between behaviors and self-assessed task difficulty have also been explored. Gwizdka [2008] showed that subjective task difficultly was correlated with the number of result pages, the number of individual documents visited, the number of documents marked as relevant, and individual cognitive differences. Liu et al. [2010] showed that there are task-type variances across behavioral predictors of task difficulty and concluded that task type is an important factor for useful task difficulty models.

2.5. Task Session Representation

Interaction can be represented as a sequence of user actions. It is reasonable to suppose that the next action of a user is influenced by the results of previous actions. In fact, this has long been a basic assumption for research in the area. To understand user-centered search dynamics, one must work out the details of the linkage between user actions. How strong is the influence from one step to the next? In what ways does previous interaction affect the user's next intentional action? What long-range (i.e. structural) influences exist on the next action? What are the appropriate groupings of interaction sequences to usefully represent task session units and the task session as a whole?

There is a long history of research in information science in identifying the various search behaviors, and their sequences, that people engage in during information seeking in general, as well as in information retrieval systems. For instance, Bates [1979a, 1979b] identified both tactics and strategies that information seekers engage in during information-seeking sessions in information systems. She also characterized a quite different model of information-seeking behavior, "berrypicking," in Bates [1989]. Ellis et al. [1993] and Ellis and Haugan [1997] have also described a set of information-seeking behaviors and their sequences. The use of patterns of search tactics has also been characterized [Joo and Xie 2012].

More recent work has explored modeling of search tactics using HMMs and Markov chains of classified user actions to identify behaviors and changing tactics in search [Yue et al. 2012]. Joo and Xie [2012] explored transitions in search tactics and applied a Markov chain (with history = 5) analysis to identify the most frequently used tactics and tactic transitions by search session stage.

This empirical work shows that information seeking is a process. Recent research has begun to address model development to identify search session behaviors. These modeling efforts are still exploratory and depend on supervised learning of user studies. It is not yet clear how they may be extended to use in operational systems.

The Cranfield search system evaluation model reduces the user's role to query production and scoring the results returned by the system. The shortcomings of this general approach are recognized for system evaluation when complex search tasks are addressed. Complex sessions have multiple queries, and those queries have relationships. If only the immediate query is used by the system to decide how to make the document collection available to the user, a lot of information is ignored that might allow for a better understanding of the user's search intent. One does not experience search as a succession of independent units of interactions. Calculating system performance by simply summing the performance for each query unit is clearly inadequate for evaluation of a system's capacity to support complex search.

Recent work to bridge the gap between search interaction seen in Cranfield-style evaluation based on simple query-results units and whole-session measurements has turned to focus on query segments [c.f. Fuhr 2008; White et al. 2010; Liu et al. 2012; Araujo et al. 2012]. Query segments are the task units of interaction delineated by a user query followed by action sequences. The immediate search intent is expressed somewhat in the query, and the following actions with the search results are presumed to be shaped by the search intent in service of the user's task goal(s).

Several approaches to considering influences in a succession of query–response units have been used. Fuhr [2008] proposes a framework for interactive information retrieval that models user moves between situations where the system provides a set of choices. Every choice is associated with costs. When a choice has been made, the user is in a new situation. An optimal choice ordering can be achieved within an appropriate set of parameter constraints. Another idea is to attribute a context to a query and the succession of user interactions, measured through a semantic representation, such as a language model from queries and selected document content, and user actions during the query segment. White et al. [2010] described an approach that develops user interest models. These models are used with context derived from the previous query segment to define a user's intent for the current query. They evaluated a variety of context signals and found good potential to improve performance.

The TREC Session Track [Kanoulas et al. 2013] has also addressed this general problem of using information from previous queries and user actions with the search results to modify the results returned in response to the current query or to modify that query itself. Liu et al. [2012] used recursive partitioning models learned from user studies involving SERP content and dwell time on pages and time between queries learned from user studies to predict the usefulness of a document before returning a ranked list to the user. Araujo et al. [2012] developed a system that incorporates a dynamic interaction model that observes the user's actions and updates its model of what is relevant for the user at that time. This framework accepts other model outputs as inputs, including user models, click models, and task models.

These examples show a turn toward dynamic systems that are driven by the observation of user actions with the idea that these actions express current search intents and task needs. The query–action sequence unit structure as a task session unit is accepted as a building block to understand task sessions and as a plausible unit input to learn models capable of predicting a user's next move or aspects of the user's search intentions. Chunking the raw sequences of interactions during a task session into search intention–action sequence units allows one to explore properties that may provide important information for modeling, for example, semantic relationships between the query and the document contents. In principle, query–action sequence boundaries in a task session can be learned directly from interaction sequences. There are, however, significant challenges to reliable identification of query segments for a single user and task session in search engine logs.

The TREC Session Track and other research expand on the Cranfield *query-search results* unit analysis to focus on representation and modeling the task unit of *query plus results interaction sequence*. This can be seen to be part of a bottom-up research approach to understanding the whole of the task session. In the introduction to this article, we emphasized the importance of a whole-task approach to understanding the user's task session and suggested that significant personalization benefit could accrue from modeling whole sessions. Both Fuhr [2008] and Araujo et al. [2012] situate the query segment interactions within a whole-session framework. For Fuhr, the task session plays out with query segment interactions guided by a whole-session view of a probability ranking principle (PRP) [Robertson 1977]. For Araujo et al., the task session is represented in the successive partitioning of the system's document collection space by relevance as the user drives the query and search results evaluation process.

A central theme of these approaches is to focus on the session as a sequence of query interactions. The potential practical value of such an approach is evident. However, such approaches may miss influences on local query segments due to global structure associated with the task properties, such as task type, or with the user properties, including knowledge of the domain or of the task. There is evidence for global structure in queries. Qiu [1993] found that users tended to adopt more structured and analytic search patterns when engaging in specific tasks than in general tasks. General tasks also caused more use of browsing features in the system. Wildemuth [2004] found that domain knowledge affects search tactics use and described common patterns of successive search queries, for example, concept specification followed by terms to elaborate the concept and then narrowing within the retrieved document set by more specific queries.

2.6. Markov Chains and Graph Properties

We consider the search task session as a sequence of interactions. A simple representation of user actions is to hypothesize that such sequences can be treated as Markov chains, in which the interactions are linked and dependency is limited to the previous interaction state. Markov chains to represent searching have been used by a number of researchers (c.f. Ageev et al. [2011] and Yue et al. [2012]). Graph representations of such sequences can be constructed for further analysis by assigning a user interaction state to a vertex and using directed edges to represent the interactions as state transitions. Ageev et al. [2011] analyzed user action sequences observed in a web-based question-answering game to characterize search behaviors and measure query efficiency. They looked at web page state transition patterns and built a Markov model and a Conditional Random Fields (CRF) model. They found that shorter paths resulted for easy questions and that the search actions and timing were the most important variables. Ageev et al. represented each task session as a single sequence of web page interactions. Our work in this article is based on a representation of the task session that is different from that used by Ageev et al. We take each session's interaction sequence but look at the common behavior sequences in the task session. This is accomplished by calculating patterns of actions in a short time window that slides over the task session. Our analysis and results relate to these patterns of behaviors, which we call "activity patterns." The procedure is described in detail later.

As with the systems of Fuhr [2008] and Araujo et al. [2012], a simple Markov chain for interaction takes account of long-range influences only to the degree that such influences are propagated from step to step. Fuhr emphasizes changes in the user "situation" for each query segment. Araujo et al. can be seen to have a similar idea of user situation expressed in the parameters for the inputs to the process that repartitions the document relevance space. Other session search modeling approaches involve inferred user states using Markov models, for example, user intentions using an HMM [Yue et al. 2012; Han et al. 2013], and user belief states using a Partially Observable Markov Decision Process (POMDP) [Luo et al. 2014].

We adopt a cognitive perspective in representing activity patterns and focus on the user's involvement in the flow of the session rather than learning user mental states to predict next action biases. Interpreting task session interaction as an activity Markov chain provides a view of how the user changes states step by step during the session. The interaction chain covers the entire session and describes the user's task session interaction as a flow from state to state without chunking into the query segments associated with distinct user "situations."

One connection between these interaction representations based on a user's "situation" and a user's immediate state is to understand the user as having a representational state that is a substrate upon which the next representational state is produced. This is one way to account for the propagation of long-range search task interactions. The flow from situation to situation versus flow from state to state is a matter of granularity of representation. For example, at the boundary of a query segment, the state of the user matches the situation of the user. In this work, we make a representation of user states that is limited to the direct observables of eye movements and page interactions. We hypothesize that the transitions between these states express the latent conditionalization of the interaction process by the user's situation.

Part of the motivation to develop a representation of user activity comes from the desire to find interaction sequence properties that generalize across users and task topics. Our work explores interaction patterns in collections of task session user action sequences. A sufficient statistic for both infinite and finite exchangeable Markov chains is the state transitions and transition counts between the states represented in the chain [Diaconis 1988]. For the general case of finite state Markov chains, only the initial state and the frequency of subchain occurrences are needed [Wolfe and Chang 1993]. So the distribution of states in a session, properties of encoding sequences of interactions, and properties of the Markov chain graphs each provide a way to characterize properties of the complete task session interaction sequence.

3. USER STUDIES

The data for the analysis presented in this article was obtained in two independent user studies. One study, "Task Cognitive Effects" (TCE) (N = 32), involved university journalism students carrying out realistic journalism tasks. It was designed to explore the effects of task type on information search behaviors. The other study, "Domain Knowledge Effects" (DKE) (N = 40), involved undergraduate and graduate students in biology-related disciplines gathering articles for search tasks in the genomics domain. The goal of that study was to investigate the effects of different levels of domain knowledge on search behaviors.

In the following sections, we describe each study and the details of the tasks. Both studies used a custom interaction logging system [Bierig et al. 2009] that automated the experiment procedure and gathered a wide range of user behaviors with an array of heterogeneous logging tools. During the experiment, keyboard and mouse activity was logged using RUI (http://ritter.ist.psu.edu/projects/RUI/), web traffic using UsaProxy (http://fnuked.de/usaproxy/) and Morae (http://www.techsmith.com), and eye movements using a Tobii T60 eyetracker with Tobii Studio (http://www.tobii.com). The system applied questionnaires prior to search to gather background information and participant expectations about their tasks, such as anticipated difficulty. After each search task, a questionnaire asked about actual difficulty, whether users had enough time to accomplish the task, and self-assessment of success. For the journalism experiment, two cognitive assessments were conducted before the start of the experiment. For both studies, there were four search tasks and a training task. Users accessed the experiment through an interface that provided instructions and presented the tasks in rotated order using a Latin Square Design.

3.1. Task-Type Classification

The task-type classification framework [Li 2009] that was used to construct the four tasks for the TCE study is presented in Table I. The specific intention in task construction was to design motivating tasks that differed systematically on several of the

Table I. Facets of Task That Were Varied in This Study (After Li [2009], Modified)

Facets	Subfacets	Values	Operational Definitions/Rules
Product		Physical	A task that produces a physical product
		Intellectual	A task that produces new ideas or findings
		Decision (solution)	A task that makes a decision or solves a problem
		Factual information	A task locating facts, data, or other similar items in information systems
		Image	A task locating image(s) in information systems
		Mixed product	A task locating different types of items in information systems
Goal	Quality	Specific goal	A task with a goal that is explicit and measurable
		Amorphous goal	A task with a goal that cannot be measurable
		Combined goal	A task with both concrete and amorphous goals
Task characteristics	Objective task complexity	High complexity	A work task involving at least five activities during engaging in the task; a search task involving searching at least three types of information sources
		Moderate	A work task involving three or four activities during engaging in the task; a search task involving searching two types of information sources
		Low complexity	A work task involving one or two activities during engaging in the task; a search task involving searching one type of information source
	Level	Document	A task for which a document as a whole is judged
		Segment	A task for which a part or parts of a document are judged

facets of task that were shown by Li to affect search behavior. Specifically, search tasks were varied according to Product (Intellectual/Factual), Goal (Specific/Amorphous), Complexity (High/Low), and Level (Segment/Document).

The five tasks for the Genomics (DKE) study were drawn from the 2004 TREC Genomics Track [Hersh et al. 2005] and selected to vary by difficulty. These tasks were designed as information tasks a research professional might conduct.

3.2. Journalism Study (TCE)

3.2.1. Study Overview. The journalism user study was designed to explore task-type influences on information search behaviors. There were 32 participants, each completing four tasks. The study logged both high-level activity, including document interaction and use, and the low-level information acquisition process as revealed in eye movement patterns.

3.2.2. Tasks. The work domain of journalism was used for reasons of both validity and convenience. Journalism has a relatively small number of work task types, but, because of its nature, the tasks can be associated with any topic. This allowed us to choose any topic for a task, which enhances the realism of the work task, thus enhancing validity. Since journalists will work on the same task type with different topics, it is not unreasonable to think their search behaviors will reflect the task type and not be highly topic dependent. The university journalism department afforded us access to experts to help develop realistic work tasks as well as a participant pool trained for such professional journalism tasks.

Table II. TCE (Journalism Study) Tasks

Task	Product	Level	Goal Quality	Objective Complexity
BIC (Background)	Mixed	Document	Specific	High
CPE (Copy editing)	Factual	Segment	Specific	Low
INT (Interview prep)	Mixed	Document	Mixed	Low
OBI (Advance obituary)	Mixed	Document	Amorphous	High

Table III. TCE Task Pair Facet Value Differences

Tasks	Sum Facet Differences
BIC, OBI	2
BIC, INT	2
INT, OBI	2
CPE, INT	3
BIC, CPE	3
CPE, OBI	5

Tasks were identified through interviews with journalism faculty, including practicing journalists, about typical journalism work, focused on information search tasks that are part of the training for professional journalists. Task descriptions were then formalized. Four of the work/search tasks were selected, which could be varied according to facet values in the task classification system. We focused on facets that have been shown to affect search behavior, for example, the objective complexity of a task and its specificity. The result of this procedure was four different tasks, presented in Table II, which varied according to values of the facets in Table I.

The tasks as presented use Borlund's [2003] normal scenario practice. In this case, we used the language of journalism work and the participants were given an *assignment* and an associated *task* to complete. The four work assignments and associated search tasks used in the study are presented in Appendix A.

A categorical distance, that is, the Manhattan distance, can be calculated for each of the study task pairs using the task facet values. For example, the facet value distance between BIC (Background) and OBI (Advance obituary) with respect to Goal Quality is 2 because Goal Quality can be Specific, Mixed, or Amorphous and Mixed lies between the other values. The facet value difference between two tasks is the absolute sum of the differences in each facet. Table III shows the differences between each pair of tasks. By this measure, BIC and OBI, and BIC and INT (Interview prep) are nearest to one another and CPE (Copy editing) and OBI are furthest apart.

3.2.3. Participants. The participants were upper-division undergraduate journalism students recruited from our university. They had completed at least one journalism writing or reporting class and so they had familiarity with the types of tasks used in the study. The age range was 18 to 27 years old. They were all native English speakers (73%) or had a high degree of English knowledge. They had an average of 8.5 years' experience using web browsers and strong web search experience. To motivate them to take their tasks seriously, they were informed that compensation would be either $20.00 or $40.00, depending on their performance as judged by experts.

3.2.4. Procedure. Participants were given a tutorial in the form of a practice task to familiarize them with the specifics of the data collection instruments and procedures and with the interface they would be using during their searches. They then performed the four web search tasks described in Appendix A and discussed in Section 3.2.2. Participants could search freely on the web using Internet Explorer (v6). The participants were restricted to this browser because it was integrated with our custom action logging system. They were asked to search until they had gathered enough information

Table IV. Search Topics/Tasks

TREC Task	MeSH Category	Topic Keywords
2	Genetic structure	Generating transgenic mice
7	Genetic processes	DNA repair and oxidative stress
42	Genetic phenomena	Genes altered by chromosome translocations
45	Genetic phenomena	Mental Health Wellness-1
49	Genetic structure	Glyphosate-tolerant gene sequences

to accomplish the task. After participants decided that they found and saved enough information for the purposes of their task, their search was replayed, and they were asked to evaluate the usefulness of the information objects they saved or saved and then deleted. An online questionnaire then asked about their searching experience, including subjective evaluation of their performance and reasons for that evaluation. Finally, an exit questionnaire was administered to elicit perceptions of the search experiences, differences in the tasks, and their ability to perform the tasks and overall search experiences.

3.3. Genomics Study (DKE Tasks)

3.3.1. Study Overview. This study investigated the relationship of task difficulty to search behavior. It used the Indri search engine from the Lemur toolkit (http://lemurproject.org) to support a custom search interface. Users could choose between a simple and familiar interface with a single text box and a search interface presenting possible search terms in different categories. We used documents (n = 1.85 million) from 2000 to 2004 in the TREC Genomics collection, a 40-year, 4.5 million document subset of the MEDLINE bibliographic database. Five search tasks were fashioned from the TREC Genomics retrieval topics. A gold standard for user performance in selecting relevant documents was provided by the TREC-expert-graded relevance judgments for each topic.

3.3.2. Tasks and Procedure. The purpose of the DKE study was to investigate the effect of domain knowledge on search behaviors. The tasks were ad hoc retrieval tasks from 50 topics with five general topic types developed for the TREC Genomics Track [Roberts et al. 2009]. Five tasks were selected that varied by search difficulty as determined by the overall performance of the retrieval systems determined by judging on the search result sets by expert assessors. The tasks selected as hard were ones where the Genomics search topic had a small number of documents judged as relevant in the search collection. The tasks selected as easy used topics where a large number of relevant documents were available in the collection. The selected tasks varied by difficulty from easy to hard, but these types of tasks are in general difficult, even for trained medical librarians [Liu and Wacholder 2008].

Considered as a task type in Li's system, all the DKE study tasks were of one type, that is, a Factual information product, with a Specific goal, of Low complexity, at the Document level. Compared to the TCE tasks, the DKE tasks were most similar to CPE (Copy editing), differing only in the Level value. CPE has the Level value of Segment. The facet value difference between the DKE tasks and CPE is 1.

Participants were asked to find and save as many documents as possible that were useful for answering the topic. The topics were presented unchanged from the TREC Genomics Track descriptions. The task labels and their corresponding MeSH categories are listed in Table IV. The tasks as presented to the participants are given in Appendix B.

Each participant performed four search tasks, which were regularly rotated using a blocked and counterbalanced Latin Square design. Everyone completed tasks 2, 7, and 45. The experiment was designed to elicit domain knowledge effects on information behaviors, and halfway through the study, we found that task 42 was too easy. We therefore switched from using task 42 to using task 49, so there were 20 participants for those tasks. The substitution of tasks is a limitation of the experiment as it reduces the number of task sessions for tasks 42 and 49 and also limits the ability to compare task pairs involving those tasks because of the effects of individual differences.

3.3.3. Participants. The participants in this study were undergraduate and graduate students and postdoctoral researchers (n = 40) in biology-related schools and departments, including, among others, biology, pharmacy, animal science, biochemistry, and public health. Their ages ranged from 18 to 32. Usable eye-tracking data was obtained from 38 participants (12 male, 26 female). The numbers of undergraduates roughly balanced the numbers of graduate students and researchers. Their genomics knowledge was elicited by asking them to self-assess their knowledge of 409 MeSH genomics concepts. Three domain knowledge groups were identified. There were high-domain-knowledge (n = 6) and low-domain-knowledge (n = 8) groups, with the majority of participants falling into an intermediate-domain-knowledge (n = 24) group. Further details regarding the procedure used to rate their knowledge are provided in Cole et al. [2012]. Participant search system experience was quite similar and they all considered themselves expert. The DKE experiment used a custom-designed search interface so that general search system expertise did not introduce behavior biases due to search system familiarity.

4. METHODOLOGY

4.1. Overview of Analysis Procedure

Different levels of user actions during information search can be observed. For example, one can observe the sequence of pages processed, a search results page followed by a linked document page, and so on. Another level is that of cognitive processing of the information made available by the system in the pages, in particular by observing the process of reading on the pages seen in the task session. The strategy of analysis in this article is to take activity representations at a high level (page sequences) and low level (eye movements) and compare statistics for each level of representation and the degree to which they distinguish between tasks. This focuses on patterns within the sequences and measures of complexity of the sequences.

We explore the statistics of patterns within the sequence representations by looking at repeated actions, that is, an action of the same class, and by making Markov models. This is accomplished by calculating compression of the task session sequences using run length encoding (RLE). We also calculate measures of complexity of the sequences by analyzing the size of the RLE codebooks and some graph properties of the Markov models.

The sequences of interactions by page type were collected from the user study logs. The cognitive effort vectors (i.e., eye fixation features) were calculated for the time period covering each web page. The sequences of observations of the succeeding page types and the page-level cognitive effort class were fashioned for each task session. Then we applied a procedure to make a representation of the user's changing activity during the session, as described in the following sections.

4.2. Eye Movement Analysis

A model to process reading eye movements was developed based on the E-Z Reader model [Reichle et al. 2006]. The E-Z Reader model hypothesizes that word identification, visual processing, attention, and control of the oculomotor system are joint determinants of eye movement in the reading process. The E-Z Reader model is a processing model that takes reading to be a cognitively controlled process where the saccade to the next word is programmed during the time the person is processing the text in the currently attended fixation. The saccade programming has a labile stage. If the next word is recognized during this labile stage, the programmed saccade is canceled and a saccade to the next word is programmed.

Text is processed during fixations in several stages. Orthographic recognition of glyphs happens in about 40ms, followed by phonological recognition (∼60ms). Minimal time for lexical processing is usually about 150ms [Reingold and Rayner 2006]. The labile lexical processing period is from 113ms to 168ms, during which the next saccade can be reprogrammed with a new saccade target. When the labile period is completed, the pending saccade will be executed after the cognitive processing is completed. This is one way in which observations of eye movements can be connected with the semantics of information processing. Eyes remain fixated during the lexical processing period independently of the stimuli, for example, even if the word is removed [Findlay and Gilchrist 2003]. The next saccade takes place only after this cognitive processing is completed. It has long been known that familiarity and conceptual complexity of the text processed is positively correlated with the fixation duration (e.g., Rayner and Duffy [1986]).

The basic E-Z Reader model does not account for higher-order cognitive processes, for example, those involving language comprehension and conceptual processing. While this is a limitation of the model, it is claimed that the model provides an explanation of the moment-to-moment reading process when linguistic processing is running smoothly [Reichle et al. 2004].

The implementation details of our method are provided in Cole et al. [2011a]. The duration of a fixation plays a key role because it is indicative of the cognitive processing required to establish the meaning of the word or phrase. The typical time range to acquire word meanings is 150ms to 300ms [Reingold and Rayner 2006]. The minimum fixation duration required for lexical access is somewhat less than that. We used a threshold of 113ms. Fixations that likely resulted in word meaning acquisition on the basis of this threshold were identified and grouped into reading sequences. Several cognitive effort measurements were calculated on each reading sequence. Simple counts of number of fixations and their durations were made. The total area covered by the reading sequence (the reading length) and the reading speed was calculated. For reading sequences containing four or more fixations, two derived cognitive effort measures were calculated:

—Perceptual span: The amount of text taken in one time. Perceptual span is operationalized as the average separation of fixations in display pixels. Studies of reading in different orthographic systems provide evidence that perceptual span reflects a concept throughput bottleneck [Inhoff and Liu 1998; Tsai and McConkie 1995; Pollatsek et al. 1986].

—Regression fixations: This is a fixation that returns to a portion of the text already processed in the reading sequence. The number of regressions and the fixation duration of the regression fixation have been associated with the difficulty of reading passages, resolution of ambiguous (sense) words, conceptual complexity of text, parsing difficulties, and the reading goal [Rayner 1989; Rayner et al. 2006].

The cognitive-level analysis in this article is based on representing each task session as a contiguous collection of reading sequences, representing the user's experience of information acquisition due to reading.

4.3. Page Sequence Analysis

Activity sequences for all the users across all tasks for each user study separately (N = 109 usable task sessions for the TCE data and N = 148 usable task sessions for the DKE data) were pooled and analyzed based on the time-series-based sequences using the page-type encoding.

The page-type codebook for the TCE study used the scheme of Gwizdka [2011]. There are four page types, distinguished by their format and by whether the user has already seen the page. The codes were query search engine results page (a query SERP) (QS); content page (C), including those accessed from a SERP and those reached by click-though from another document; Returns to a SERP (L); and revisits to a previously examined (in the query sequence) content page (M).

For the DKE data, this coding scheme was extended to take in an additional page type. During the task session, participants could click through to a page that showed all of the documents they had saved to that point in the search. We coded moves to the saved documents page as Z.

Following the approach in Hendahewa and Shah [2013], activity sequences of page types were constructed but without normalizing the session durations in order to maintain the variability in time spent by different users when conducting the tasks. Activity sequences composed of the coded page types were constructed over time for each user. Each activity sequence was constructed at the level of seconds, and if a certain activity was conducted for n seconds, then the page type corresponding to that activity was repeated n times along the activity sequence. Then the overall activity sequence for each user was divided into smaller subsequences with overlap using a sliding window approach. For such subsequence construction, a sliding window of 60 seconds with a step of 30 seconds was used. Each window was represented as the observed page type at 1-second intervals and then the Hamming distance between succeeding windows (subsequences) was calculated. An eight-cluster fit using Hierarchical Agglomerative Clustering (HAC) was used to find the clusters that each of the subsequences belongs to over the total pooled task sessions. The selection of an eight-cluster fit was to match the learned cognitive effort states, as explained later. At the end of this process, the time-based user activity sequences are represented as sequences of clusters, which are then compared across different task types using sequence complexity and Markov chain properties.

4.4. Low-Level Information Acquisition Activity

The eye fixation logs were processed to identify each reading eye movement sequence and to calculate the cognitive effort vectors. The details of the eye movement data processing and calculations are given in Cole et al. [2011b]. These cognitive effort vectors were grouped into sequences at the level of pages and a variety of statistics calculated, resulting in the 26 features listed in Table V.

These observations were then clustered using k-means and cross-checked using HAC. An eight-cluster fit was determined to give good separation and stability in the k-means clustering. The clusters were cross-checked for stability by making an HAC eight-cluster fit and comparing membership. The sequences of cognitive effort cluster classes were calculated for each task session and then used to label the cognitive effort vector for each page.

The level of the features in Table V relates to the relationship between the feature and the page. Page-level features are those that collect all of the cognitive effort data for the page without filtering. For example, the total number of fixations gets all of the fixations on the page. Reading sequences get all of the eye movement sequences classified as reading sequences. Likewise, the mean of the reading length is a statistic that uses all of the reading sequences on the page. In contrast, the Subset-level features apply a filter to select only some of the eye fixation interactions on the page. For example, it is reasonable to think that a user making more long reading sequences on a page is in a different information processing state as compared to scanning or briefly sampling

Table V. Page-Level Reading Features

Level	Code	Description
Page	nFix	Number of fixations on the page
Page	sumFix	Total fixation duration on the page
Page	nSS	Number of single-fixation (“scanning”) sequences
Page	durSS	Duration of the scanning sequences
Page	nFixRS	Number of fixations in the reading sequence
Page	durRS	Duration of the reading sequences
Page	medianNRS	Median number of fixations in the reading sequences
Subset	nLongRS	Number of “long” (number fixations >4) reading sequences
Subset	durLongRS	Total duration of the “long” reading sequences
Subset	medianNFixLongRS	Median number of fixations in “long” reading sequences
Page	nRRS	Number of regressions in reading sequences
Subset	nRLongRS	Number of regressions in the long reading sequences
Page	meanPS	Mean perceptual span in the reading sequences
Page	RL	Pixels covered by reading sequences (reading length)
Page	meanRL	Mean of the reading length for the reading sequences
Page	RSpeed	Reading length/total fixation duration (reading speed)
Page	medianFixLADE	Median of fixation duration’s lexical portion (LADE)
Subset	medianMaxFixDur	Median of the longest fixations in the reading sequences
Subset	medianLADELongRS	Median LADE for the long reading sequences
Local	nFix_LRS	Number of fixations in the longest reading sequence
Local	durFix_LRS	Total fixation duration for the longest reading sequence
Local	medianFixDur_LRS	Median fixation duration in longest reading sequence
Local	maxFixDur_LRS	Max fixation duration in longest reading sequence
Local	RL_LRS	Reading length of the longest reading sequence
Local	RSpeed_LRS	Reading speed for the longest reading sequence
Local	PS_LRS	Perceptual span for the longest reading sequence

disparate parts of the page. A feature that calculates a statistic about the long reading sequences would capture some of this aspect of user attention and cognitive activity. The Local level is a selection of a particular reading sequence, for example, just the longest reading sequence on the page.

After creating the page cognitive effort cluster-labeled data, we then made the activity representation of the cognitive effort sequences using the same technique as for the high-level page-type sequence activity representations.

4.5. Clustered Page Types

It is useful to note that the clustered cognitive effort feature vectors can be analyzed to understand how the classes differ from one another. The features in Table V were grouped by level of attention allocation on the page. It is interesting to ask which of these groups of features is most influential in determining the membership of each class. Cluster class properties might be useful in analyzing user activity patterns. For example, activity patterns might identify segments during a task session where one might want to investigate changes in cognitive effort.

Table VI shows the top four influential features in the clustered classes of pages. Each cluster class is understood as a textual information-processing type. The feature importance was calculated for each cluster by making a multiclass Random Forests model [Breiman 2001] for each of the learned TCE clusters. After creating and tuning the Random Forests model, the cluster feature importance was calculated using the Gini coefficient [Strobl et al. 2007]. For a couple of clusters, there were fewer than four features that met an “importance” threshold.

Table VI. Important Features for the Cognitive Effort States (TCE Study)

Cluster	Feature	Feature	Feature	Feature
1	medianFixDur_LRS (local)	medianLADELongRS (subset)	medianFixLADE (page)	medianMaxFixDur (subset)
2	meanRL (page)	nFixRS (page)	nLongRS (subset)	NA
3	durFix_LRS (local)	RL_LRS (local)	MeanPerceptualSpan (page)	meanRL (page)
4	RSpeed_LRS (local)	PS_LRS (local)	meanPS (page)	maxFixDur_LRS (local)
5	maxFixDur_LRS (local)	nFix_LRS (local)	RL_LRS (local)	durLongRS (subset)
6	nFix (page)	nRS (page)	NA	NA
7	medianFixDur_LRS (local)	durLongRS (subset)	nFix_LRS (local)	durRS (page)
8	medianFixDur_LRS (local)	maxFixDur_LRS (local)	medianLADELongRS (subset)	meanRS (page)

4.6. Complexity of Sequences

One measure of activity is the sequence complexity of the actions. To gain an idea of the complexity of the sequences, we opt for a simple measure of compression based on RLE. RLE is a widely used simple lossless data compression method in which a sequence is represented as a sequence value followed by how many times it occurred repeatedly in the sequence. It is a simple way of compressing the representation of the user cluster sequences for each task level and provides a rough measurement of the complexity of the action sequences by task. The compression rate for each user in a task is simply the RLE encoded length divided by the length of the sequence.

An approximate measure of the complexity of the RLE is to count the length of the codebook for the RLE. We calculate this value for all of the user task sessions and compare the distribution by task. The RLE compression for each user task session was calculated and aggregated by task and the distributions compared. We also look at the length of the task session RLE codebooks as another measure of task session action sequence complexity and compare the results by task.

4.7. Markov State Models and Graph Properties

Markov chain representations of the task session sequences at each representation level were created. We also collected the session sequences by task and then constructed an overall state model for each task. Several properties of the task graphs, including state transition probabilities, number of edges, and maximum clique size, were calculated and compared across tasks.

5. RESULTS

5.1. Subjective Difficulty

Each of the user studies involves tasks that varied by difficulty. For DKE, the genomics research tasks (Table VII) were selected by their retrieval difficulty. In TCE, the tasks were varied in facets, some of which were expected to contribute to overall task complexity. One of the facets was explicitly a measure of task complexity as the number of essential goals to achieve, for example, checking the truth of three assertions.

In both studies, participants were asked to assess the expected difficulty of the task before the tasks, and the experienced difficulty immediately after completing the tasks. Figures 1 and 2 show the results for each study.

For the TCE study (Figure 1(b)), one can see that Copy Editing (CPE) and Advance Obituary (OBI) were judged, respectively, as the easiest and most difficult tasks at the posttask assessment. These are also the most dissimilar tasks as measured by

Table VII. Designed and User-Rated Task Difficulty Ratings for DKE Study

Task	Topic Title	Designed Difficulty Level	User-Rated Difficulty Level	Mean User Difficulty Rating
2	Generating transgenic mice	Hard	Difficult	4.53
7	DNA repair and oxidative stress	Easy	Easy	3.83
42	Genes altered by chromosome translocations	Easy	Easy	4.32
45	Mental Health Wellness-1	Hard	Difficult	4.88
49	Glyphosate-tolerant gene sequence	Easy	Difficult	5.24

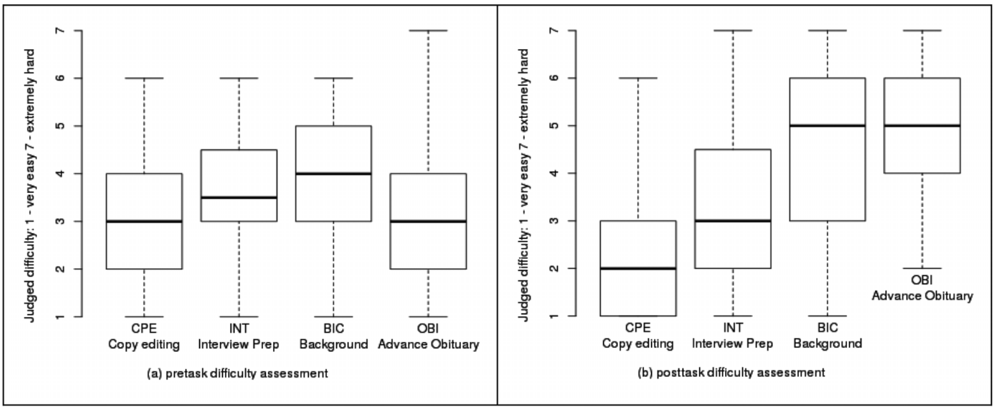


Fig. 1. TCE study: Participant assessed task difficulty: (a) pretask; (b) posttask.

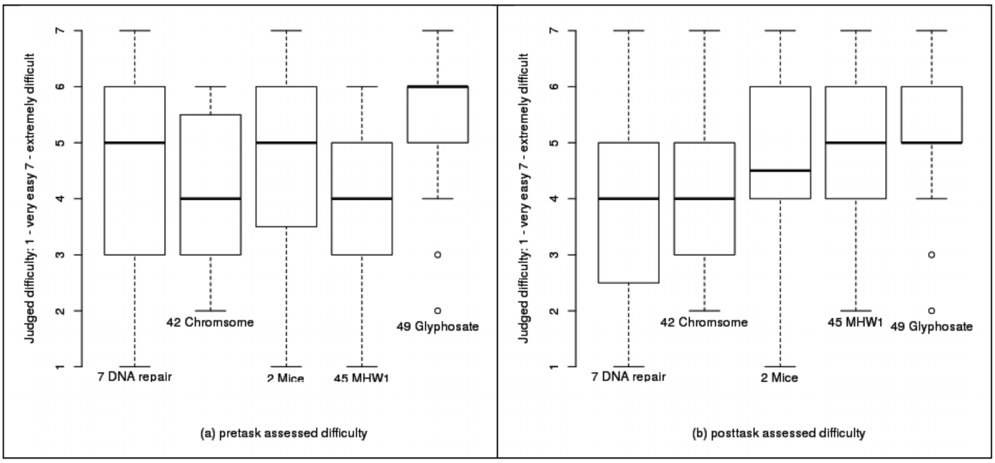


Fig. 2. DKE study: Participant assessed task difficulty: (a) pretask; (b) posttask.

task facet differences in the classification system. Further, the retrospective subjective difficulty rankings match the expected contribution of facets to task difficulty. For example, one expects amorphous tasks such as Advance Obituary (OBI) and Background Information (BIC) to be more difficult than specific tasks, such as Copy Editing (CPE). Recall that Interview Preparation (INT) can be seen as a two-part task where the first part is somewhat open-ended in finding experts germane to the assignment while the second part is specific (get the expert's contact information).

Table VIII. Distribution of Subsequence Clusters Across Task Types

		Page-Type Level				Cognitive Effort Level			
Tasks	Facet Diff	Value	Rank	χ^2	p	Value	Rank	χ^2	p
BIC, OBI	2	0.05	1	0.4667	0.4945	0.31	2	6.3623	0.0116*
BIC, INT	2	0.10	2	0.2765	0.5990	0.21	1	0.0111	0.9162
INT, OBI	2	0.11	3	0.0441	0.8336	0.32	3	6.3623	0.0116*
CPE, INT	3	0.17	4	0.8947	0.3442	0.44	4	0.0442	0.8334
BIC, CPE	3	0.23	5	1.8667	0.1719	0.47	5	0.0442	0.8334
CPE, OBI	5	0.24	6	0.9983	0.3177	0.60	6	6.3623	0.0116*

(Statistical Significance at 95% Is Denoted by *).

For DKE (Figure 2(a)), one can see that participants anticipated most tasks to be at least somewhat difficult. Tasks 2, 7, and 49 were expected to be difficult. After the tasks were completed (Figure 2(b)), the participants found task 7 to be easier than expected and task 45 harder than expected. The retrospective difficulty rankings as compared to the designed task difficulty ratings are presented in Table VII.

5.2. TCE Study Results

5.2.1. Task-Based Activity Analysis. One way to investigate task-type differences is to ask if there are differences between tasks in the distribution of classes of observations we learned by clustering. Table VIII shows the similarity of the task pairs. The *Facet difference* value is the Manhattan distance between the values of the respective task facets (see Table III). *Value* is the sum of the absolute difference of each class frequency at each level:

$$sim(Task_A, Task_B) = \sum_{n=1}^{N} \left| Task_A(C_n) - Task_B(C_n) \right|,$$

where N is the number of classes and $Task_X(C_n)$ is the normalized frequency of the class observed in the instances of $Task_X$. The differences were tested for significance using the Kruskal-Wallis test.

Eye-tracking data is subject to dropouts, for example, when participants look away from the screen, and other factors that sometimes result in gaps in the eye fixation data logs. For this article, the page-level cognitive effort representations were learned using only the search sessions where there was a complete eye-tracking record (N = 47) covering the page dwell times. The task session distribution was BIC 12, CPE 13, INT 14, OBI 10 (see Appendix A for descriptions of each of these tasks). This relatively even distribution of the complete task sessions is evidence for no systematic bias by task in the dataset. However, some bias against particularly long task sessions is expected because a longer run has a greater probability of encountering a dropout event. In practical settings, one could fill in bad data with a "not available" state, impute the missing data using the observed distributions of cognitive effort classes on pages, or use the session state transition probabilities. We have compared several techniques to impute data and find that use of the observed session transition probabilities is promising.

CPE is dissimilar to the other tasks, and most dissimilar to OBI at both activity levels. The pair rankings at the two levels only disagree on the two most similar task pairs, (BIC, INT) and (BIC, OBI). This is not too surprising as one might expect a measurement of user activity to be least effective when the tasks are most similar. Only OBI could be distinguished from the other tasks in the cognitive effort activity representation with high statistical significance.

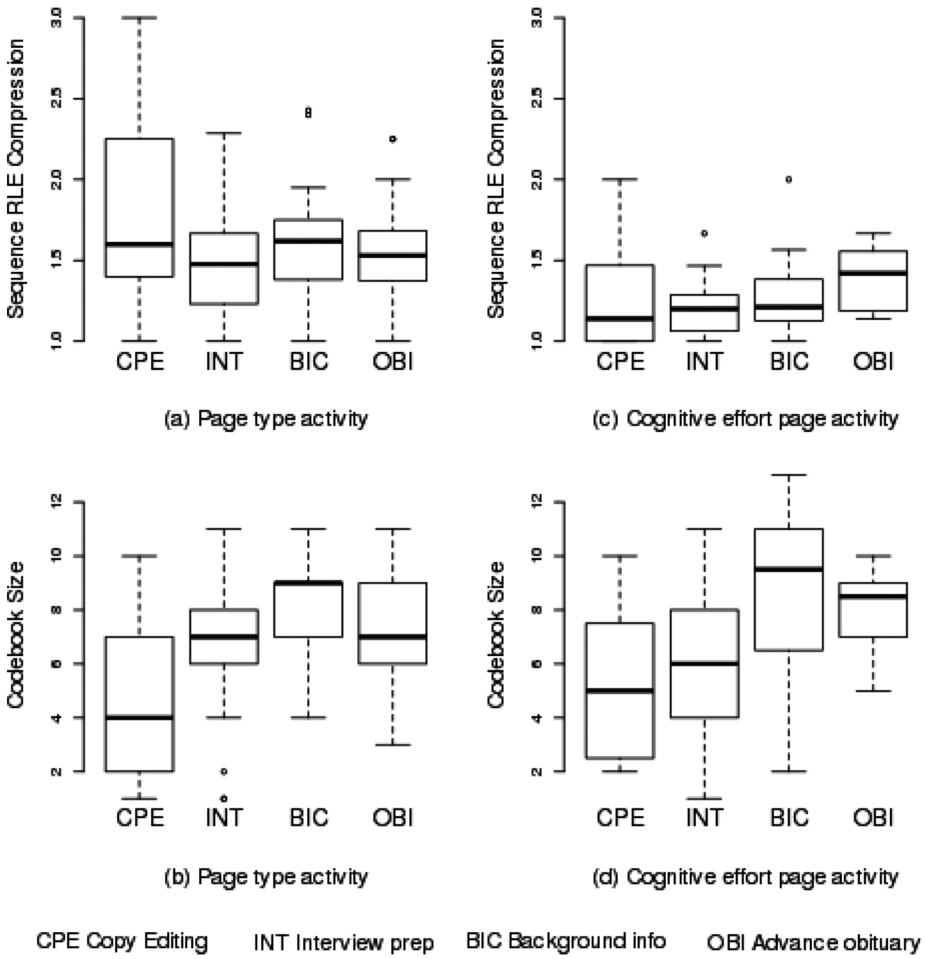


Fig. 3. TCE study: RLE compression and codebook size ordered by posttask difficulty assessment.

5.2.2. Run Length Encoding Analysis. RLE operates on patterns of user activity represented in the subsequence cluster classes to make a compressed representation of the user activity. If a type of activity observation is repeated in the sequence, RLE will provide greater compression.

Figures 3(a) and 3(c) show there is little overall difference in RLE compression rates for the page-type subsequence activity. RLE compression is likewise similar for all but OBI, which is greater. There is greater variability in the compression for CPE sessions for both the high-level page subsequence activity and the cognitive-level representations. Overall, there is greater compression achieved in the page subsequence activity representation as compared to the cognitive level, except for OBI, where the compression is nearly the same. The OBI, CPE task pair is the only pair with a statistically significant difference (Kruskal-Wallis chi-square = 4.017, p = 0.045) at the page subsequence level, while only the INT, OBI pair is statistically significant at the cognitive effort level (Kruskal-Wallis chi-square = 3.8296, p = 0.050) at a 95% significance level.

Another way to investigate patterns and complexity using RLE is to look at the codebook dimensionality, that is, the size of the alphabet needed to describe the sequence using the RLE algorithm. Figures 3(b) and 3(d) show that RLE codebook sizes vary

Table IX. Markov Models of Activity: Task Transition Matrix Similarities

		Page Type Level				Cognitive Effort Level			
Tasks	Facet Diff	Value	Rank	$M\chi^2$	p	Value	Rank	$M\chi^2$	p
BIC, OBI	2	0.79	1	5.58	1	5.15	2	54.17	0.8046
INT, OBI	2	1.08	2	6.09	1	5.92	3	81.95	0.0647
BIC, INT	2	1.35	3	11.83	1	4.45	1	17.62	1
BIC, CPE	3	2.00	4	7.24	1	6.29	4	85.86	0.0355*
CPE, INT	3	2.19	5	389.80	0*	6.92	6	166.47	<<0.0001*
CPE, OBI	5	2.25	6	773.46	0*	6.89	5	93.06	0.0103*

(Statistical Significance at 95% Is Denoted by *).

consistently by task type for the page-type subsequence activity and for the cognitive effort subsequence activity representations. There is a strong similarity in both the relative differences between tasks and in the absolute values. CPE is least complex by the RLE codebook dimensionality measure and BIC is the most complex.

There are some differences between the RLE codebooks for page subsequence activity and cognitive effort page activity. In the page subsequence activity, the RLE codebook size for BIC is largest and INT and OBI are nearly the same. Significant differences exist for four task pairs: (BIC, CPE) (Kruskal-Wallis chi-square = 20.4462, p = 6.13e-06); (BIC, INT) (Kruskal-Wallis chi-square = 5.7017, p = 0.017); (INT, CPE) (Kruskal-Wallis chi-square = 9.3322, p = 0.0022); and (CPE, OBI) (Kruskal-Wallis chi-square = 13.3371, p = 0.0003). In contrast, for the cognitive effort page activity, the OBI and BIC codebooks are nearly the same size and both are larger than the INT codebook. Significant differences exist for two task pairs: (BIC, CPE) (Kruskal-Wallis chi-square = 5.827, p = 0.0158) and (CPE, OBI) (Kruskal-Wallis chi-square = 4.5869, p = 0.0322). Two other task pairs are nearly at the significance threshold: (BIC, INT) (Kruskal-Wallis chi-square = 3.4708, p = 0.0624) and (CPE, OBI) (Kruskal-Wallis chi-square = 3.0949, p = 0.0785).

5.2.3. Markov Models of the Tasks. Markov models were constructed from the page subsequence and cognitive effort activity states. One way to explore patterns of user activity is to investigate the properties of the graphs made from the sequence of user activities in a task session represented as a Markov chain. For each task, the Markov chains were combined to make a model and several graph properties were analyzed. The state transition matrices are the basic representation for the chains and the graph properties are derivative. We begin by looking at the differences in the state transition matrices, including statistical significance analysis, and then investigate some of the task graph properties to illustrate aspects of the differences in the state transition structures.

5.2.3.1. State Transition Probabilities. Table IX shows the sum of the absolute differences in the state transition matrices for Markov models of each pair of tasks. Generally, CPE is dissimilar to BIC and OBI. CPE is most similar to INT at the high-level page subsequence representation and least similar at the low-level cognitive effort representation.

We tested the statistical significance of the similarity of Markov chain distributions for task pairs using the Markov chi-square test presented by Bickenback and Bode [2001] based on Anderson et al. [1957]. This test assumes comparison of stationary distributions, so if a state transition is observed in one transition matrix but not the other, the two distributions must have different generators. Since we make a finite number of observations, it is possible that a missing state transition will appear in the next observation. To handle this problem, we calculated a PAC (Probably Approximately Correct) upper bound for missing state transitions using De Morgan's rule and modified the observed distributions appropriately to test for statistical significance. CPE

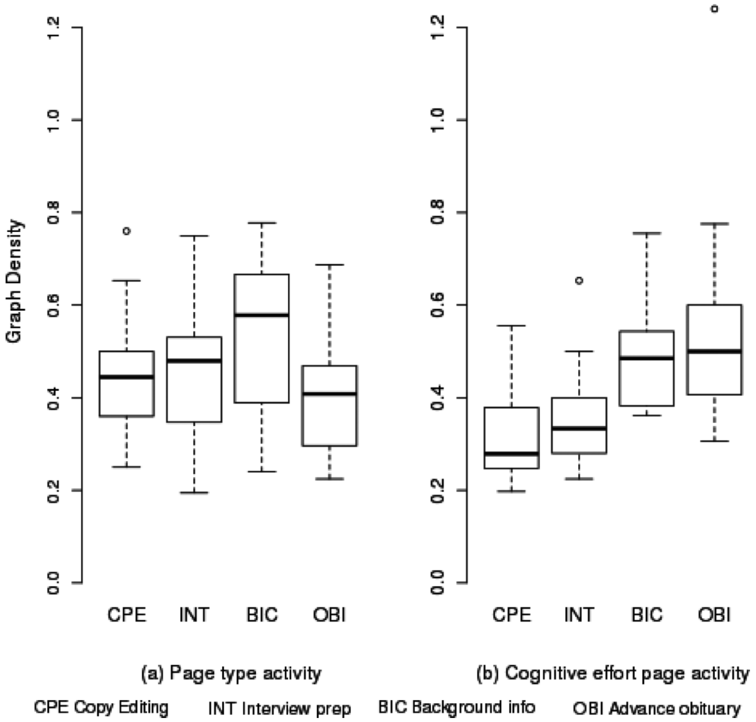


Fig. 4. TCE study: graph density by task ordered by posttask difficulty assessment.

can be distinguished with high significance from OBI and INT at the page subsequence level, and at the cognitive effort level CPE can be distinguished with high statistical significance from each of the other tasks. If the missing state transition fix value is reduced by an order of magnitude, then all of the task pairs are distinguishable with $p << 0.05$.

5.2.3.2. Graph Density by Task. Graph density is the ratio of the number of directed edges to the possible number of edges in the graph, assuming bidirectional edges are possible between any two vertices. For the search session activity graphs, density measures the sparseness of connections between user activity states represented as cluster classes.

Figure 4 shows a divergence between the two levels in distinguishing tasks. The cognitive effort subsequence activity shows distinctions between, at least, CPE and INT versus BIC and OBI. For the page subsequence activity, BIC is distinguished from the other tasks, but there is not much difference between the other tasks. The main difference between the two levels is the graph density of the OBI task sessions.

5.2.3.3. Graph Edges. One measure of complexity is to count the number of directed edges for the graph of user activity in each task session. Generally, if there are more edges in the graph, the user has engaged more patterns of activity in the sense of transitions between one activity state and another. Figure 5 shows the number of edges recorded in the task sessions for each task type at the two representation levels.

At both representation levels, CPE is distinguished from the other tasks with the fewest graph edges. Further, one can see that both the page subsequence level and the cognitive effort level are in close agreement on the graph complexities for each of the tasks.

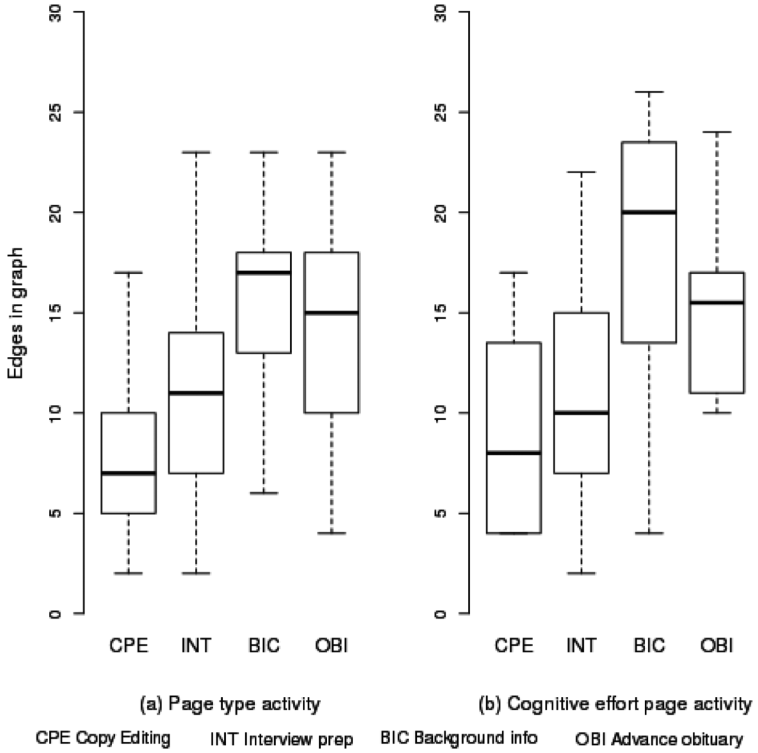


Fig. 5. TCE study: Markov model graph properties for activity patterns: number of graph edges ordered by posttask difficulty assessment.

5.2.4. Maximum Cliques. The density of connections in a graph is an important structural property. One pattern of user activity is free movement within a collection of states as contrasted with patterns that require a transition through one state to reach another state. One way of distinguishing activity graphs is to calculate the largest subgraph where the nodes are all connected to every other node in the subgraph. This is the maximum clique. Figure 6 shows the maximum clique size distributions in the activity graphs of the task sessions. The distribution of maximum clique sizes discriminates between the tasks for both the high-level page subsequence activity and the cognitive effort level activity.

The BIC graphs frequently have the highest maximum clique values, followed by OBI. CPE and INT have lower maximum clique values. The only significant difference between the two levels is for CPE, where the median value is lower at the cognitive effort level. These results are in line with the other results on differentiating tasks and are suggestive of the variety of activity patterns one might observe for different tasks. A CPE session is likely to have fewer variations on activity patterns as compared to BIC.

6. DKE STUDY TASK RESULTS

We followed the same procedures as for the data from the TCE study to fashion a page-type activity representation and a page cognitive effort activity representation. The page-type activity representation used 148 sessions and the page cognitive effort activity representation used the 100 (out of 152) sessions that were not affected by eye-tracking dropouts. The distribution of good sessions was:

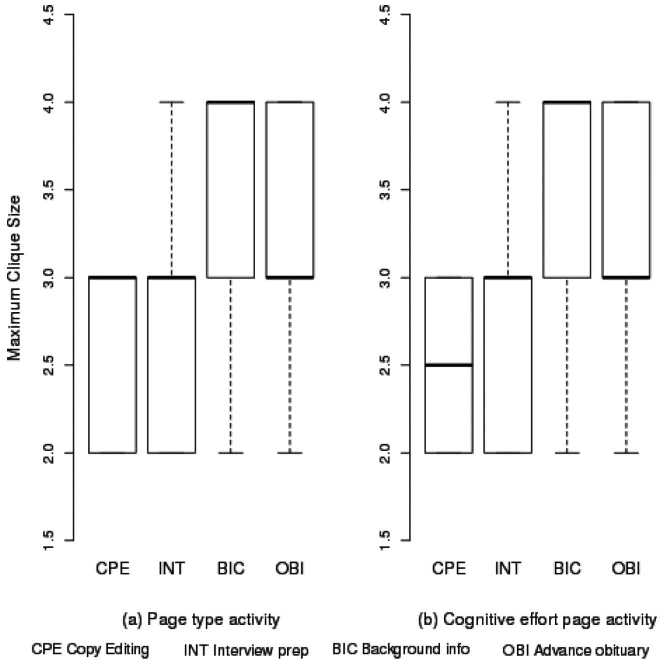


Fig. 6. TCE study: Markov model graphs: maximum clique size ordered by posttask difficulty assessment.

Task 2 (Mice) – 24; Task 7 (DNA repair) – 31; Task 42 (Chromosome) – 15; Task 45 (MHW1) – 23; Task 49 (Glyphosate) – 7

Recall that task 42 and task 49 were each used for half of the participants and so the effect of dropped sessions is more obvious. For those tasks, the results are subject to greater variance. Note that in notation of difficulty for each task, "D" and "E" represent that the user-rated task difficulty level was considered as "Difficult" and "Easy," respectively (as shown in Table VII).

6.1 Distribution Differences

Table X shows the absolute sum of the differences between the normalized cluster distributions for each pair of DKE tasks. At the page-type level, the most similar tasks are 42 (Chromosome) and 45 (MHW1) and the least similar pair is task 2 (Mice) and task 49 (Glyphosate). Two of the task pair differences are significant and two nearly significant, and those pair difficulties do not show a task discrimination bias (DD = 1 (of 3), ED = 2 (of 6), EE = 1 (of 1)). For the cognitive effort level, the most similar tasks are task 2 and task 45, while the least similar pair is task 2 and task 49. There are more significant pair differences (seven out of 10) and again, the ability to significantly distinguish task pairs covers the range of assessed difficulties (DD = 2 (of 3), ED = 4 (of 6), EE = 1 (of 1)).

For both page type and cognitive effort, the distribution differences generally fail to distinguish the task pairs by the assessed task difficulty by ranking, although they agree in a few cases. For example, the only easy tasks, 7 (DNA repair) and 42 (Chromosome), are ranked second in distribution similarity and are statistically significant for both levels. In the triad of difficult tasks, 2 (Mice), 45 (MHW1), 49 (Glyphosate), only the pair (2, 45) cannot be significantly distinguished.

Table X. Distribution of Subsequence Clusters Across Tasks

Tasks	Page-Type Level				Cognitive Effort Level			
(Difficulty)	Value	Rank	χ^2	p	Value	Rank	χ^2	p
42, 45 (ED)	0.07	1	6.8934	0.0087*	0.25	9	8.6471	0.0033*
7, 42 (EE)	0.09	2	3.5735	0.0588	0.13	2	9.9265	0.0016*
45, 49 (DD)	0.11	3	12.055	0.0170*	0.23	7	11.3108	0.0008*
7, 45 (ED)	0.12	4	2.4816	0.1152	0.14	3	0.2237	0.6363
2,42 (DE)	0.12	5	3.7804	0.0518	0.21	6	5.1059	0.0239*
2,7 (DE)	0.13	6	1.1029	0.2936	0.18	4	0.0028	0.9581
42, 49 (ED)	0.17	7	0.4688	0.4936	0.20	5	6.109	0.0135*
2, 45 (DD)	0.18	8	0.798	0.3717	0.13	1	0.0441	0.8336
7, 49 (ED)	0.20	9	1.7284	0.1886	0.23	8	11.3108	0.0008*
2, 49 (DD)	0.29	10	3.0116	0.0827	0.29	10	10.2904	0.0013*

(Statistical Significance at 95% Is Denoted by *).
Tasks: 2: Mice, 7: DNA repair, 42: Chromosome, 45: MHW1, 49: Glyphosate.

The rankings for the task pairs at the two levels do not correlate well. Overall, the absolute sum of distribution differences is not successful in ranking the similarity of all task pairs in the task difficulty dimension. However, it has success in distinguishing between individual task pairs.

6.2. Run Length Encoding Analysis

The RLE compression rates and codebook sizes for the page type and cognitive effort page activity representations are presented in Figure 7.

The compression measurements at the page subsequence level identify task 7 (DNA repair) as the least compressible and task 49 (Glyphosate) as the most compressible. Apart from 42 (Chromosome), where there is a disagreement between levels, the ordering of compressibility corresponds with the posttask difficulty assessments. When the task pairs are considered, only (7, 49 (ED)) (Kruskal-Wallis chi-square = 6.123, p = 0.013) and (7, 45 (ED)) (Kruskal-Wallis chi-square = 10.677, p = 0.001) could be distinguished as statistically significant.

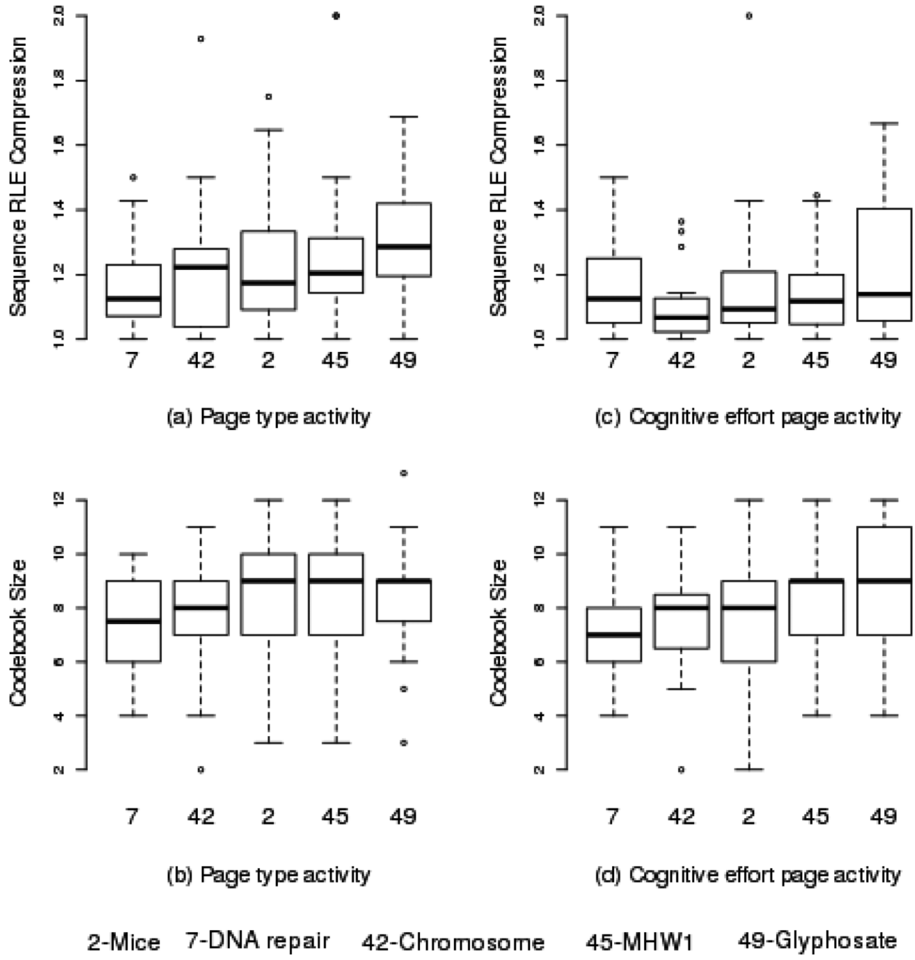
RLE compression on the cognitive effort activity finds task 42 (Chromosome) (E) as the least compressible and task 49 (Glyphosate) (D) as the most compressible. None of the task pair differences had statistical significance (Kruskal-Wallis).

6.3. Codebook Complexity

The codebook complexity shows good agreement in ordering of means between the two representation levels and with the posttask difficulty assessments. The page-type activity means for all of the difficult tasks (7 (DNA repair), 45 (MHW1), 49 (Glyphosate)) were very nearly the same. At the cognitive level, there is better discrimination of means. Considering codebook differences for the task pairs, only 7 (E) and 45 (D) are significantly distinguished: page subsequence level (Kruskal-Wallis chi-square = 4.0217, p = 0.0449) and cognitive effort level (Kruskal-Wallis chi-square = 5.985, p = 0.0144).

6.4. State Transition Probabilities

Table XI shows the absolute sum of differences in the normalized state transition probabilities for each of the DKE task pairs. The results parallel those for the cluster distributions. It is interesting to see that task 2 (Mice) (D) and task 49 (Glyphosate) (D) are the least similar, with strong statistical significance, at both levels. Tasks 7 (DNA repair) (E) and 42 (Chromosome) (E) were ranked as similar at both levels (with no statistically significant difference between them). Otherwise, there are some agreements but also strong disagreements between the two levels. As with the raw distribution results, the cognitive-effort-level task pairs are not significantly distinguishable apart from (2, 49



(a) Page type activity

(c) Cognitive effort page activity

(b) Page type activity

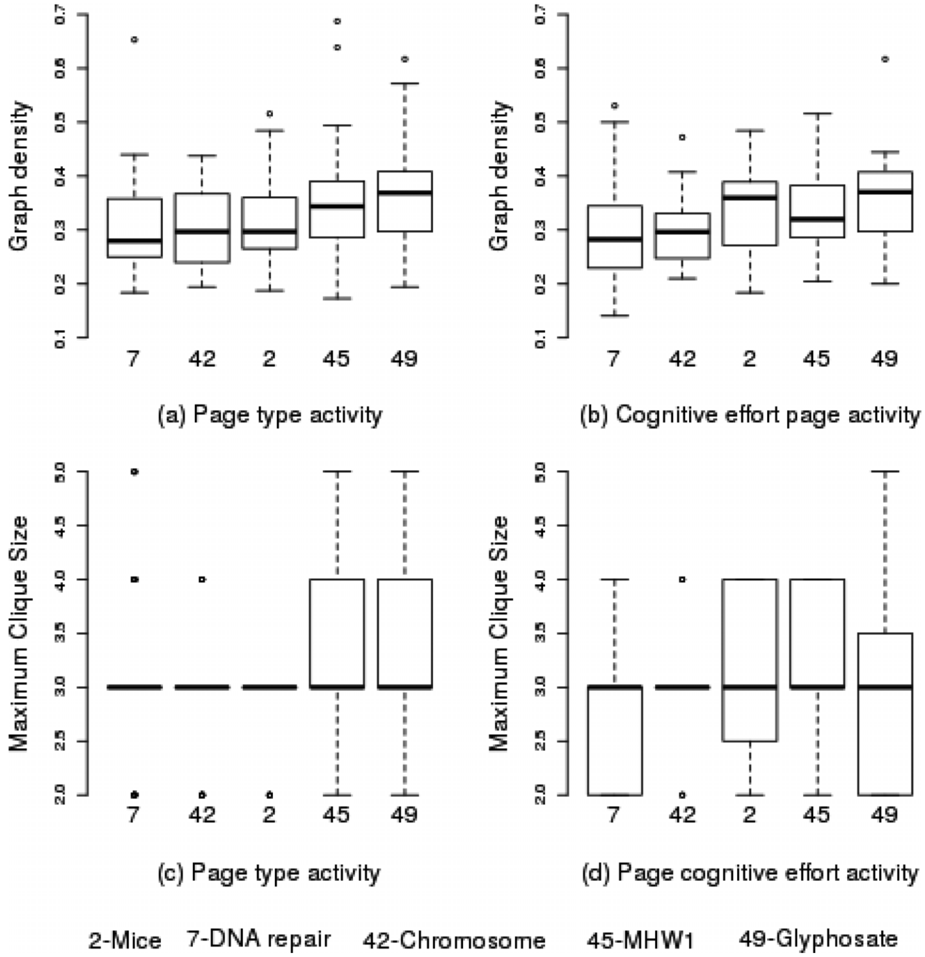
(d) Cognitive effort page activity

Fig. 7. DKE study: RLE compression and codebook sizes ordered by posttask difficulty assessment.

Table XI. Markov State Transition Probability Sum of Differences Between DKE Tasks

	Page-Type Level				Cognitive Effort Level			
Tasks	Value	Rank	$M\chi^2$	p	Value	Rank	$M\chi^2$	p
7, 42 (EE)	1.39	1	20.04	1	3.64	3	58.32	0.6767
7, 45 (EH)	1.40	2	4.01	1	3.03	1	21.12	0.9999
45, 49 (HH)	1.65	3	25.91	1	4.43	8	68.37	0.3312
42, 45 (EH)	1.88	4	7.68	1	4.27	6	78.85	0.1001
2,7 (HE)	2.36	5	11.23	1	3.82	4	26.03	0.9999
2,42 (HE)	2.40	6	11.06	1	4.79	9	107.54	0.0005*
7, 49 (EH)	2.75	7	167.33	<<0.0001*	3.83	5	78.20	0.1091
42, 49 (EH)	2.80	8	163.90	<<0.0001*	4.27	6	35.93	0.9982
2, 45 (HH)	3.42	9	38.26	0.9956	3.48	2	27.81	0.9999
2, 49 (HH)	4.90	10	1221.30	0*	5.23	10	195.84	<<0.0001*

(Statistical Significance at 95% Is Denoted by *).
Tasks: 2: Mice, 7: DNA repair, 42: Chromosome, 45: MHW1, 49: Glyphosate.



(a) Page type activity

(b) Cognitive effort page activity

(c) Page type activity

(d) Page cognitive effort activity

Fig. 8. DKE study: Graph density and maximum clique size ordered by posttask difficulty assessment.

(DD)), while at the page subsequence level, a couple more pairs are distinguishable. Again, the Markov chi-square test was used on the transition probability matrices modified using De Morgan's rule to fix a PAC upper bound on the probability of unobserved state transitions. So our significance estimates on the task pair differentiation are conservative and the two levels may be more similar than shown by this comparison.

6.5. Graph Properties

After making the Markov chains for each task session, the graph properties were calculated. The graph density at the page-type activity level (Figure 8(a)) and the cognitive effort page level (Figure 8(b)) show mean values that, except for task 2 (Mice) (D), correspond well with the order of the DKE participant posttask difficulty assessments. At the page-type activity level, tasks 45 (MHW1) (D) and 49 (Glyphosate) (D) had higher graph densities than the easy tasks, 7 (DNA repair) and 42 (Chromosome). However, task 2 is not distinguishable from task 7 or task 42. At the cognitive effort level, tasks 2 and 49 have the highest mean value and are well distinguished from the easy tasks, 7 and 42. In contrast to the TCE results, for DKE, the maximum clique size (Figures 8(c) and 8(d)) does not discriminate task difficulty at either level.

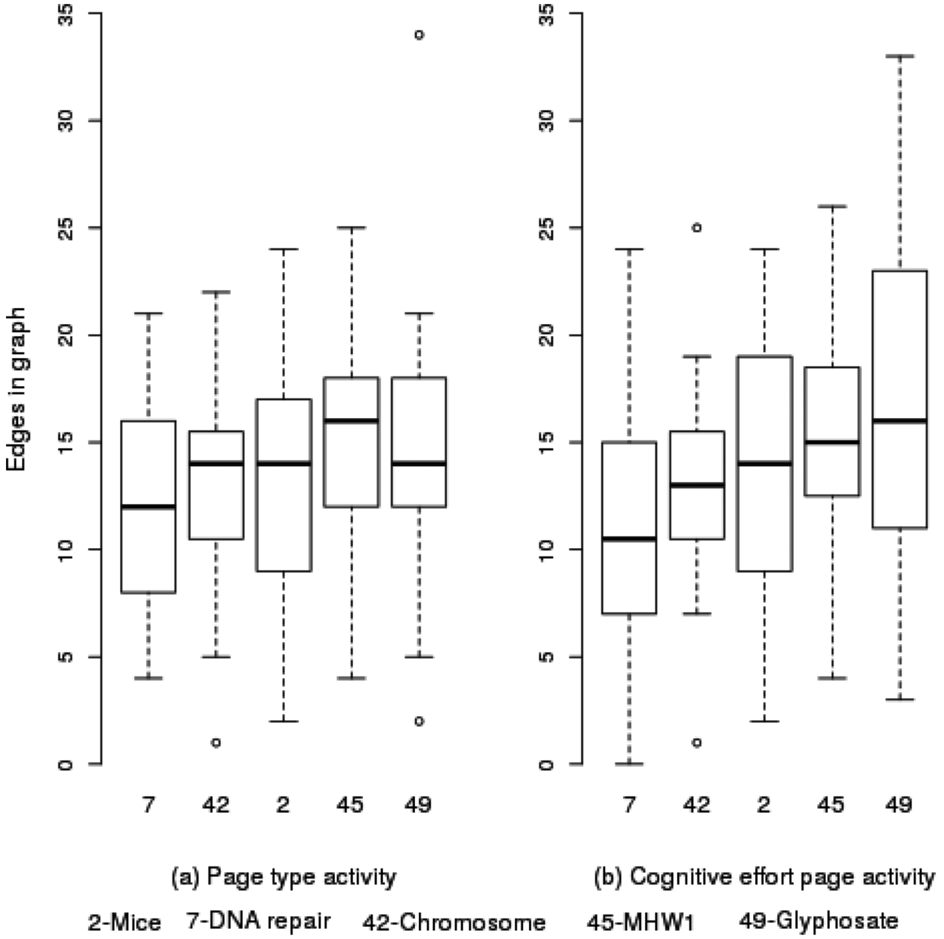


Fig. 9. DKE study: number of edges in activity graphs ordered by posttask difficulty assessment.

The number of graph edges and the overall graph density are closely related to one another. Graph density is the ratio of the number of edges to the number of possible edges if the graph was fully connected. This was calculated for each task session, some of which had fewer states than other sessions. So the graph density of two sessions with the same number of edges could be different. The number of edges provides a count of the connections between states and so indicates how many moves were available in the session. The graph density describes the coverage of the Markov chain in the session state space. Figure 9(a) shows that the number of graph edges for page-type activity sessions varies little between tasks. In contrast, the task ordering at the page cognitive effort activity level (Figure 9(b)) corresponds well with the posttask difficulty assessments.

7. DISCUSSION

7.1. Overall Results

Our research questions address the validity of activity pattern analysis and its capacity to distinguish between tasks and learn some aspects of task properties. Table XII summarizes the relative success in detecting differences between the tasks using the

Table XII. Measures on Activity Patterns: Success in Distinguishing Tasks

	TCE		DKE	
Measure	Page Type	Cognitive Effort	Page Type	Cognitive Effort
Distributions	✓	✓	X	X
Transition probabilities	✓	✓	X	X
RLE compression	X	✓	?	X
RLE codebook size	✓	✓	✓	✓
Graph density	X	✓	?	✓
Graph num. edges	✓	✓	✓	✓
Graph max. clique size	?	✓	X	X

various measurements applied to the user activity patterns. A check mark means the measurement was reasonably successful. A question mark means there was some success but also a bad miss, or that a pattern of successful discrimination is visible but not as cleanly distinguished as the reasonably successful measurements. An “X” means the measurement generally failed to distinguish tasks in the expected way, even if it correctly handled a few cases. These judgments do not address the confidence with which tasks could be detected or compared with other tasks if one were given an individual task session.

Our first research question concerned the ability to detect aspects of a user’s task type by observing his or her activities. The results for the TCE study showed that measurements on activity patterns consistently differentiated between tasks that had different facet value characteristics. The cluster distributions and state distributions differentiated the TCE tasks in the expected way. Further, there was a correspondence with posttask difficulty assessments and the facet value distance between the task pairs. RLE codebook size also differentiated the tasks well at both levels, but RLE compression was successful only at the cognitive effort level. The graph properties, which reveal structure within the transition probability matrices, did a good job at the cognitive effort level. The number of graph edges was successful at both levels.

The DKE study used tasks of a single task type that were most similar to CPE (Copy editing) in the TCE study. Measured by facet value distance, the DKE tasks were closer to CPE than any of the other TCE tasks. The DKE tasks varied by search engine retrieval task difficulty. We found the activity pattern technique was able to distinguish between the DKE tasks on the basis of difficulty. However, only codebook size and the number of graph edges were able to differentiate tasks reasonably well at both the page-type and cognitive effort page activity levels. For a number of measurements, the most difficult (49 – Glyphosate) and easiest (7 – DNA repair) tasks could be distinguished, but not other task pairs.

The TCE results show that instances of distinct task types can be ranked in expected order using the technique for many measures, but the DKE results show that tasks of a single task type can also be distinguished by at least some activity pattern measures. We conclude that the results provide good evidence for a positive answer to our first research question. However, further exploration is needed to resolve the relationship between task difficulty and task-type facets. This is discussed further later.

Our second research question asked if user activity patterns were able to distinguish between tasks. The results provide good evidence that evaluation of user activity patterns can distinguish tasks based on the results for both the TCE and DKE studies for at least some measures.

Our third research question concerned the generalizability of the technique. The main generalizability test in this article was to compare task discrimination results from two independent user studies. Furthermore, the facets that were used for constructing the TCE tasks were taken from Li [2009]. That faceted scheme was based on studies of tasks quite different from journalism tasks and, indeed, of a quite wide variety of tasks within an academic community. Thus, we believe some claim can be made for generalizability to tasks in other domains in which tasks can be characterized by these facets and their values.

For the DKE study, the technique produced the expected subjective task difficulty ranking when RLE codebook complexity and the number of graph edges were considered. However, a more detailed analysis on discrimination between pairs of tasks does not reveal a simple relationship between task difficulty and these measurements of activity patterns. For the TCE study, the technique produced the expected task identification based on the task-type facet values. For both studies, these task identification and characterization results agreed with the rankings we found in previous work using various task session measures, including time on task, number of pages processed, dwell time, and others [Liu et al. 2010; Cole et al. 2011a, 2011b; Liu et al. 2011].

On balance, there are enough similarities between the TCE and DKE results at both the high and low representation levels to say there is evidence for the generalizability of the technique. In particular, one would not expect to see similar rankings against independent variables (task facet design, task difficulty design, and user difficulty assessments) if the technique itself was not valid or reliable. The results show matches across the studies as well as between levels. Even when there is not a strong match, the results show partial matches. The results of similarities between the levels can be interpreted as evidence for a causal relationship between the unit low-level tasks of the textual information acquisition process and the high-level page use.

The technique might be distinguishing between task types only by a subset of facets, for example, objective difficulty. To distinguish between task-type instances using activity patterns would then rest on a subsidiary hypothesis that the task difficulty construct can be resolved into task facet value components. The task difficulty construct is complex, and further research is needed to better model aspects of task difficulty. Activity pattern analysis might be able to contribute to such a research program.

Finally, user activity pattern analysis is centered on a dynamic representation of individual search activity patterns. Another dimension of generalization, then, is to learn if the technique can be applied within specific task sessions to explore the details of the user's moment-to-moment interaction within the task session. For example, can the technique be used to detect significant events in the task session? One goal is to model aspects of the user's flow during search, which might be correlated with his or her experience of the search session and with observable events in the session, for example, patterns suggesting frustration or learning. The technique holds promise for gaining specific understanding of an individual's moment-to-moment actions.

In summary, regarding our research questions, we found that the activity pattern analysis supports the hypothesis that search tasks in the higher band of rational activity, such as page use, click actions, and so forth, during the information search process are reflected in rather low-level patterns of the moment-to-moment experience of search. The results also show that the activity pattern technique appears to be able to discriminate between tasks. Further, the consistent discrimination of the tasks that matches intuitions about differences between tasks is evidence that activity patterns can detect task properties to some degree. That is, differences in patterns of user activity appear to be indicative of both the type of task in which the user is engaged and at least one general characteristic of tasks, namely, difficulty. Since the activity patterns do not involve the content, except to the degree that the reading eye movements relate to the processing of content, we believe that activity pattern models for task-type prediction and difficulty can be constructed and tested. Such models could then be applied during sessions to sequences of new data to make predictions of the task type or difficulty. A system might apply reranking algorithms and other rules and responses that have been previously learned to give better performance for tasks with those characteristics.

7.2. Activity Pattern Complexity Measures

Complexity-related measures were most correlated with the ability to distinguish between tasks. For both the page-type and cognitive effort activity patterns, the size of the run length encoding codebook distinguished tasks in the expected way. Likewise, graph properties of the Markov chains aggregated by task that related to the number of states visited and their connectivity also did a good job of distinguishing the tasks in the studies. This was seen in the RLE codebook and in Markov chain graph properties that measure the density and structure of state transitions, both in page types and cognitive effort states.

The success of the technique in discriminating between the tasks is evidence that this activity pattern approach is capturing differences in how participants made progress in their tasks. For both the TCE tasks and the DKE tasks, there are measures of activity patterns that provided successful task discrimination.

For the TCE tasks, it is not hard to describe how the observed success of complexity measures can be related to the task differentiation success. The results suggest a correlation between complexity-related measures on activity and task difficulty or complexity. One explanation is that simple tasks, specific tasks, and easy tasks are carried out in activity patterns that are less complex than those used when confronted with difficult search tasks or ones that are complex or amorphous.

Li's system has an explicit objective complexity measure based on the number of essential steps to complete the task. However, other facets, such as whether the task is amorphous or specific, can be seen as also relating to complexity. In an amorphous task, a user must make choices from a great range of (ill-defined) possibilities, whereas in a specific task, one expects that the required choices and decisions are both well-defined and more limited sets. It seems reasonable to expect a bias toward more complex patterns of exploration to meet the needs of amorphous tasks. Other facets, for example, whether the task product is intellectual, are also amenable to thinking about the range of choices confronting a user during the process. This can be seen to fit with Fuhr's [2008] PRP principle applied within the changing situations of users as they move through the search task session.

Such a relationship would not be surprising. Similar observations and results have been reported for whole-session analysis involving, for example, query structures over time [Qiu 1993; Wildemuth 2004].

For DKE, it is more challenging to provide an explanation for the activity pattern technique because the tasks varied only by topic and the objective retrieval difficulty of finding relevant documents. How might this be explained? To grapple with this challenge, we need to explore the task difficulty construct.

7.3. Task Difficulty

Objective task difficulty and perceived task difficulty measurements are complex constructs. In the task classification system, there are several facets that contribute to making a task difficult. In comparing the results for the two studies, the relationship between task difficulty and these basic properties of task type needs to be untangled.

For the TCE study, the tasks were designed to be different in facets that were expected to affect search behavior. The DKE study did not use designed tasks. The DKE study used a single task type that was similar to the Copy editing task in the TCE study. The DKE tasks were fashioned from the TREC Genomics Track topics using the objective difficulty of finding the documents that were judged relevant to the topic by the TREC expert assessors. So, the DKE tasks are of one task type with five topics in the genomics domain and they vary by difficulty. The variation in the difficulty of the DKE tasks was correlated by the user posttask difficulty assessments (Figure 2).

There are several challenges to making a connection between the TCE and DKE tasks to interpret the results. First, several facets in the task classification system contribute to making a task difficult. In addition, the DKE study was designed to elicit domain knowledge influence on search behaviors, and participants were selected to have varying domain knowledge. Activity patterns could be quite different for those with high levels of domain knowledge as compared with those with lower levels. Even the understanding of the task goal, that is, understanding the task goal concept well enough to be able to recognize useful documents, likely varied among participants. In contrast, the TCE study participants had experience with the types of tasks and their topic domain knowledge was not so important to task performance.

In summary, measurements involving subjective task difficulty raise many questions when thinking about how it can be used to model the search process, tasks, and task sessions. Which task difficulty assessment should be used? How do anticipated and in-task assessments affect search session biases? When do users form their posttask assessments during the session? These are interesting but thorny questions. It is tempting to see subjective task difficulty as a way to bridge between the results for TCE and DKE, but it is hard to trust that bridge to bear much weight.

7.4. Summary: Task Difficulty, Task Facets, and Activity Complexity

There are intuitive connections between task difficulty and the task facets. The TCE results show that the designed tasks were differentiated using the activity patterns in ways that are consistent with participants' posttask difficulty assessments. In the DKE study, the tasks were of one task type. Those tasks were also distinguished using the same activity pattern measures in ways that are consistent with participants' posttask difficulty assessments. The complexity of the task difficulty construct makes it hard to say precisely what is being measured. While some results suggest activity patterns are able to distinguish both the TCE and, to a lesser degree, DKE tasks by subjective task difficulty, we believe such a strong claim is unwarranted. Further investigation is needed to understand task difficulty and how it relates to task sessions, including the evolution of task sessions.

The user activity pattern technique can be applied during the task session. We intend to use search activity patterns as a tool to explore relationships between user assessments of the current task difficulty and dynamic assessments of user actions during search.

The usefulness of in-session local activity pattern detection for specific personalization moves is another direction for future research. User activity patterns can be calculated for segments of a task session, so practical application does not require knowledge of task session boundaries or, in particular, the start of the session. The task-type results suggest that user activity pattern complexity may be associated with task properties that presumably involve greater intellectual effort, for example, because they are not specific or involve more challenging concepts. We do not address the problem of examining in-session activity patterns in this work. The complexity and graph properties of local activity chains can be continuously calculated. Relating the immediate cognitive effort of a search task to properties of the search task as a whole can also have practical implications. For example, a system might use activity pattern changes as points to make an intervention decision because the system would have a hint as to what type of help could be beneficial. For example, a system could act when it sees complexity peaks by automatically taking notes, bookmarking, or allowing direct, rather than linear, navigation to any of the recent pages or query results. Implicit support could be implemented by adjusting ranking algorithms by document readability or reweighting specific metadata. Activity pattern changes might identify significant events in a task session, such as instances of learning. In such cases, one could examine the action(s) associated with the event, for instance, clicking through to a page or reading a particular passage. A system could hypothesize that the event induced a change in the user's knowledge and so privilege some content in modeling the user search as part of a personalization scheme.

7.5. Explaining Changes in Activity Patterns

One motivation in selecting the features to fashion the cognitive effort vectors was to get features at different levels of textual information acquisition from the document. Classification of the features in the cognitive effort vectors by level can be used to explore the structure of the learned cognitive effort clusters. Each cluster is a state in the Markov chains used to represent activity sequences. These states can be labeled with the levels of the features that are most important in defining the cluster membership. That is, one can say State N is dominated by cognitive effort features that draw on the entire page, while State M is dominated by features that concern a single reading passage in the page.

The user cognitive states learned for the TCE and DKE user studies show some variance in their most significant features. This suggests hypotheses about the relationships between state transitions and the user's task session activities and search intentions. Section 4.5 presented some analysis of the most important cognitive effort features in the page clusters (Table VI) and suggested how they might be used to explain cluster state transitions.

8. CONCLUSIONS

One goal of our research is to improve personalization of information systems by predicting the task type and other aspects of the user's current task. Task is known to affect user search behaviors, so we have looked at predicting task type from observations of user behaviors. Several challenges exist in exploring and modeling behaviors. Behaviors, per se, can be hard to detect. They emerge from user actions, and sequences of actions express behaviors. Behaviors are labels for classifications of user action sequences. The segmentation of user actions that correspond to behavior instances is not obvious. Expression of a behavior is extended and variable in time. Different behaviors are likely to have different expected lengths of user actions.

Hendahewa and Shah [2013] developed the idea of representing human search sessions as fixed temporal units of interaction and focused on representing the search activity as local patterns of actions. This representation may be a useful way to get around some of the challenges of representing behaviors directly.

Our research questions relate to a hypothesis about the cognitive activity of users during information search tasks. The hypothesis that search tasks in the information search process are reflected in rather low-level patterns of the moment-to-moment experience of search is implicit in user-centered work. This is a source of the power of context to improve system performance. Our results show that the activity pattern technique applied at the low level appears to be able to distinguish between tasks and detect differences in the high-level tasks by characterizing them consistently. It suggests that differences in the patterns of user activity are able to indicate aspects of the higher levels of task in which the user is engaged. That is support for this cognitive-level linkage hypothesis for information search tasks.

We have introduced a technique to represent users' activity patterns as they work during extended information search sessions. One contribution of this article is the development and validation of this technique and demonstrating that it can detect aspects of tasks that are germane to personalization, such as task complexity and difficulty. Moreover, this technique provides a way to represent and detect properties of the user's current and historical situation in the task session.

The results suggest a correlation between complexity-related measures on activity and task difficulty or complexity. One explanation for our results is that simple tasks, specific tasks, and easy tasks are carried out in activity patterns that are less complex than those used when confronted with difficult search tasks or ones that are complex or amorphous. Such a relationship is not a surprise, and similar observations and results have been reported for whole-session analysis involving, for example, query structures over time [Qiu 1993; Wildemuth 2004]. While there is much more work to be done, we are pleased with the positive results of our analysis of information search task session datasets from two independent user studies. The activity pattern technique presented shows promise as a useful tool for identifying the nature of the information seeker's motivating task and for representing the user's ongoing search experience during an information-seeking session. Both of these contexts of search are crucial to personalization of support for effective information seeking.

In this work, the activity pattern models were built by aggregating task sessions across users. Future work will explore the potential to fashion cognitive explanations for differences in task behaviors using the general cognitive effort activity pattern states. We will also contrast the success of the general models with activity pattern models made for individuals across their tasks. It is plausible that individual differences can be found in the complex cognitive effort states for activity pattern state spaces. Comparing general models and individual models may help one better understand the degree to which individual differences affect search behaviors.

APPENDIX A

TCE Study Tasks

Background Information Collection (BIC): Product—Factual; Goal—Specific; Complexity—High; Level—Document

Your assignment: You are a journalist at the *New York Times* working with several others on a story about "whether and how changes in U.S. visa laws after 9/11 have reduced enrollment of international students at universities in the U.S." You are supposed to gather background information on the topic, specifically, to find what has already been written on this topic.

Your Task: Please find and save all the stories and related materials that have already been published in the last 2 years in the *New York Times* on this topic and also in five other important newspapers, either U.S. or foreign.

Interview Preparation (INT): Product—Factual, Intellectual; Goal—Amorphous and Specific; Complexity—Low; Level—Document

Your assignment: Your assignment editor asks you to write a news story about "whether state budget cuts in New Jersey are affecting financial aid for college and university students."

Your Task: Please find the names of two people with appropriate expertise that you are going to interview for this story and save just the pages or sources that describe their expertise and how to contact them.

Advance Obituary (OBI): Product—Factual; Goal—Amorphous; Complexity—High; Level—Document

Your assignment: Many newspapers commonly write obituaries of important people years in advance, before they die, and in this assignment, you are asked to write an advance obituary for a famous person.

Your Task: Please collect and save all the information you will need to write an advance obituary of the artist Trevor Malcolm Weeks.

Copy Editing (CPE): Product—Factual; Goal—Specific; Complexity—Low; Level—Segment

Your assignment: You are a copy editor at a newspaper and you have only 20 minutes to check the accuracy of the three underlined statements in the excerpt of a piece of news story to follow:

New South Korean President Lee Myung-bak takes office

Lee Myung-bak is the 10th man to serve as South Korea's president and the first to come from a business background. He won a landslide victory in last December's election. He pledged to make economy his top priority during the campaign. Lee promised to achieve 7% annual economic growth, double the country's per capita income to US$4,000 over a decade, and lift the country to one of the top seven economies in the world. Lee, 66, also called for a stronger alliance with top ally Washington and implored North Korea to forgo its nuclear ambitions and open up to the outside world, promising a better future for the impoverished nation. Lee said he would launch massive investment and aid projects in the North to increase its per capita income to US$3,000 within a decade "once North Korea abandons its nuclear program and chooses the path to openness."

Your Task: Please find and save an authoritative page that either confirms or disconfirms each statement.

APPENDIX B

DKE Study Tasks

CATEGORY I: GENETIC PROCESSES

7: DNA repair and oxidative stress

Need: Find correlation between DNA repair pathways and oxidative stress. Context: Researcher is interested in how oxidative stress affects DNA repair.

CATEGORY II. GENETIC PHENOMENA

45: Mental Health Wellness-1

Need: What genetic loci, such as Mental Health Wellness-1 (MWH1), are implicated in mental health?

Context: Want to identify genes involved in mental disorders.

42: Genes altered by chromosome translocations

Need: What genes show altered behavior due to chromosomal rearrangements? Context: Information is required on the disruption of functions from genomic DNA rearrangements.

CATEGORY III: GENETIC STRUCTURE

49: Glyphosate tolerance gene sequence

Need: Find reports and glyphosate tolerance gene sequences in the literature. Context: A DNA sequence isolated in the laboratory is often sequenced only partially, until enough sequence is generated to identify the gene. In these situations, the rest of the sequence is inferred from matching clones in the public domain. When there is difficulty in the laboratory manipulating the DNA segment using sequence-dependent methods, the laboratory isolate must be re-examined.

2: Generating transgenic mice

Need: Find protocols for generating transgenic mice.

Context: Determine protocols to generate transgenic mice having a single copy of the gene of interest at a specific location.

ACKNOWLEDGMENTS

This research was supported by a Google Faculty Research Award to N.J. Belkin and C. Shah and by IMLS grants LM-06-07-0105-07 and RE-04-12-0105-12.

REFERENCES

Mikhail Ageev, Qi Guo, Dmitry Lagun, and Eugene Agichtein. 2011. Find it if you can: A game for modeling different types of web search success using interaction data. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 345–354. DOI:http://dx.doi.org/10.1145/2009916.2009965

Omar Alonso, Michael Gertz, and Ricardo A. Baeza-Yates. 2007. On the value of temporal information in information retrieval. *SIGIR Forum* 41, 2 (2007), 35–41.

T. W. Anderson and Leo A. Goodman. 1957. Statistical inference about Markov chains. *Annals of Mathematical Statistics* 28, 1 (March 1957), 89–110. Retrieved from http://www.jstor.org/stable/2237025.

Samur Araujo, Gebrekirstos Gebremeskel, Jiyin He, Corrado Bosscarino, and Arjen de Vries. 2012. CWI at TREC 2012, KBA Track and Session Track. In *Proceedings of the 21st Text Retrieval Conference Proceedings (TREC'12)*, E. Vorhese and I. Soboroff (Eds.). Retrieved from http://trec.nist.gov/pubs/trec21/t21.proceedings.html.

Anne Aula, Rehan M. Khan, and Zhiwei Guan. 2010. How does search behavior change as search becomes more difficult? In *Proceedings of CHI 2010*, Elizabeth D. Mynatt, Don Schoner, Geraldine Fitzpatrick, Scott E. Hudson, W. Keith Edwards, and Tom Rodden (Eds.). ACM, 35–44. DOI:http://doi.acm.org/10.1145/1753326.1753333

Marcia J. Bates. 1979a. Idea tactics. *Journal of the American Society for Information Science* 30 (Sep. 1979), 280–289.

Marcia J. Bates. 1979b. Information search tactics. *Journal of the American Society for Information Science* 30, 4 (1979), 205–214.

Marcia J. Bates. 1989. The design of browsing and berry-picking techniques for online search interface. *Online Review* 13 (1989), 407–424.

Harold Bekkering and Sebastiaan F. W. Neggers. 2002. Visual search is modulated by action intentions. *Psychology Science* 13, 4 (July 2002), 370–374.

Nicholas J. Belkin. 2008. Some(what) grand challenges for information retrieval. *SIGIR Forum* 42, 1 (2008), 47–54. DOI:http://dx.doi.org/10.1145/1394251.1394261

Frank Bickenbach and Eckhardt Bode. 2001. *Markov or Not Markov – This Should Be a Question.* Kiel working paper, Vol. 1086. Institut für Weltwirtschaft an der Universität Kiel, Kiel, Germany.

Ralf Bierig, Jacek Gwizdka, and Michael J. Cole. 2009. A user-centered experiment and logging framework for interactive information retrieval. In *Proceedings of the SIGIR 2009 Workshop on Understanding the User: Logging and Interpreting User Interactions in Information Search and Retrieval*, Nicholas J. Belkin, Ralf Bierig, Georg Buscher, Ludger van Elst, Jacek Gwizdka, Joeman Jose, and Jamie Teevan (Eds.). CEUR, 8–11.

Pia Borlund. 2003. The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research* 8, 3 (2003). Retrieved from http://informationr.net/ir/8-3/paper152.html.

Leo Breiman. 2001. Random forests. *Machine Learning* 45 (2001), 5–32.

Georg Buscher, Andreas Dengel, and Ludger van Elst. 2008. Query expansion using gaze-based feedback on the subdocument level. In *Proceedings of SIGIR'08.* ACM, 387–394.

Katriina Byström. 2002. Information and information sources in tasks of varying complexity. *Journal of the American Society for Information Science and Technology* 53, 7 (2002), 581–591.

Katriina Byström and Kalervo Järvelin. 1995a. Task complexity affects information seeking and use. *Information Processing & Management* 31, 2 (1995), 191–213. DOI:http://dx.doi.org/10.1016/0306-4573(95)80035-R

Katriina Byström and Kalvero Järvelin. 1995b. Task complexity affects information seeking and use. *Information Processing and Management* 31, 2 (1995), 191–213.

Michael J. Cole, Jacek Gwidzka, Chang Liu, Nicholas J. Belkin, and Xiangmin Zhang. 2012. Inferring user knowledge level from eye movement patterns. *Information Processing & Management* 49 (Nov. 2012), 1075–1091. DOI:http://dx.doi.org/10.1016/j.ipm.2012.08.004

Michael J. Cole, Jacek Gwizdka, Chang Liu, and Nicholas J. Belkin. 2011a. Dynamic assessment of information acquisition effort during interactive search. In *Proceedings of the American Society for Information Science and Technology Conference* (2011), Vol. 48. ASIS&T, 1–10.

Michael J. Cole, Jacek Gwizdka, Chang Liu, Nicholas J. Belkin, Ralf Bierig, and Xiangmin Zhang. 2011b. Task and user effects on reading patterns in information search. *Interacting with Computers* 23, 4 (July 2011), 346–362. DOI:http://dx.doi.org/10.1016/j.intcom.2011.04.007

Persi Diaconis. 1988. Recent progress on de Finetti's notions of exchangeability. *Bayesian Statistics* 3 (1988), 111–125.

Susan T. Dumais and Nicholas J. Belkin. 2005. The TREC interactive tracks: Putting the user into search. In *TREC. Experiment and Evaluation in Information Retrieval*, E. M. Voorhees and D. K. Harman (Eds.). MIT Press, Cambridge, MA, 123–145.

David Ellis, Deborah Cox, and Katherine Hall. 1993. A comparison of the information seeking patterns of researchers in the physical and social sciences. *Journal of Documentation* 49 (1993), 356–356.

David Ellis and M. Haugan. 1997. Modeling the information seeking patterns of engineers and research scientists in an industrial environment. *Journal of Documentation* 53, 4 (1997), 384–403.

John M. Findlay and Iain D. Gilchrist. 1998. Eye guidance and visual search. In *Eye Guidance in Reading and Scene Perception*, Geoffrey Underwood (Ed.). Elsevier Science Ltd., New York, 295–312.

John M. Findlay and Iain D. Gilchrist. 2003. *Active Vision: The Psychology of Looking and Seeing.* Oxford University Press, New York.

Norbert Fuhr. 2008. A probability ranking principle for interactive information retrieval. *Information Retrieval* 11, 3 (2008). 251–265. DOI:http://dx.doi.org/10.1007/s10791-008-9045-0

Jacek Gwizdka. 2008a. Cognitive load on web search tasks. In *Proceedings of Workshop on Cognition and the Web*. 83–86.

Jacek Gwizdka. 2008b. Revisiting search task difficulty: Behavioral and individual difference measures. *Proceedings of the American Society for Information Science and Technology* 45, 1 (2008), 1–12.

Jacek Gwizdka. 2009. Cognitive load and web search tasks. In *Proceedings of the 3rd Workshop on Human-Computer Interaction and Information Retrieval*. Catholic University of America, 54–57.

Jacek Gwizdka and Ian Spence. 2006. What can searching behavior tell us about the difficulty of information tasks? A study of web navigation. *Proceedings of the American Society for Information Science and Technology* 43, 1 (2006), 1–22.

Shuguang Han, Zhen Yue, and Daqing He. 2013. Automatic detection of search tactic in individual information seeking: A hidden Markov model approach. In *iConference 2013 Proceedings*. 712–716. DOI:http://dx.doi.org/10.9776/13330

Chathra Hendahewa and Chirag Shah. 2013. Segmental analysis and evaluation of user focused search process. In *Proceedings of the International Conference on Machine Learning and Applications (ICMLA'13)*, Vol. 1. 291–294. Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6784629.

William R. Hersh, Aaron M. Cohen, Jianji Yang, Ravi Teja Bhupatiraju, Phoebe M. Roberts, and Marti A. Hearst. 2005. Genomics track overview. In *The 14th Text Retrieval Conference (TREC 2005)*, Gaithersburg, MD. NIST, 14–25.

Jukka Hyönä, R. F. Lorch, and J. K. Kaakinen. 2002. Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. *Journal of Educational Psychology* 94, 1 (2002), 44–55.

Albrecht W. Inhoff and Weimin Liu. 1998. The perceptual span and oculomotor activity during the reading of Chinese sentences. *Journal of Experimental Psychology: Human Perception and Performance* 24, 1 (1998), 20–34.

Jiepu Jiang, Daqing He, and James Allan. 2014. Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'14).* ACM, New York, NY, USA, 607–616. DOI:http://dx.doi.org/10.1145/2600428.2609633

Soohyung Joo and Iris Xie. 2012. Exploring search tactic patterns in searching digital libraries. In *ICADL*, Hsin-Hsi Chen and Gobinda Chowdhury (Eds.). Lecture Notes in Computer Science, Vol. 7634. Springer, 349–350. DOI:http://dx.doi.org/10.1007/978-3-642-34752-8_48

Evangelos Kanoulas, Ben Carterette, Mark Hall, Paul Clough, and Mark Sanderson. 2013. Overview of the TREC 2012 session track. In *Proceedings of the 21st Text REtrieval Conference (TREC'12)*. NIST. Retrieved from http://trec.nist.gov/pubs/trec21/t21.proceedings.html.

Melanie Kellar, Carolyn Watters, and Michael Shepherd. 2007. A field study characterizing web-based information-seeking tasks. *Journal of the American Society for Information Science and Technology* 58, 7 (2007), 999–1018.

Jeonghyun Kim. 2006. Task as a predictable indicator for information seeking behavior on the web. Ph.D. Dissertation. Rutgers University, New Brunswick, NJ.

Kyung-Sun Kim and Bryce Allen. 2002. Cognitive and task influences on Web searching behavior. *Journal of the American Society for Information Science and Technology* 53, 2 (2002), 109–119.

Yong-Mi Kim and Soo Young Rieh. 2005. Dual-task performance as a measure for mental effort. In *Proceedings of the 68th Annual Meeting of the American Society for Information Science and Technology (ASIST'05).* ASIST, Charlotte, NC.

Yuelin Li. 2008. *Relationships Among Work Tasks, Search Tasks, and Interactive Information Searching Behavior.* Ph.D. Dissertation. Rutgers University, New Brunswick, NJ.

Yuelin Li. 2009. Exploring the relationships between work task and search task in information search. *Journal of the American Society for Information Science and Technology* 60, 2 (2009), 275–291.

Yuelin Li and Nicholas J. Belkin. 2010. An exploration of the relationships between work task and interactive information search behavior. *Journal of the American Society for Information Science and Technology* 61, 9 (2010), 1771–1789.

Chang Liu, Michael J. Cole, Eun Baik, and Nicholas J. Belkin. 2012. Rutgers at the TREC 2012 session track. In *Proceedings of the 21st Text Retrieval Conference Proceedings (TREC'12)*, Ellen Vorhese and Ian Soboroff (Eds.). NIST.

Chang Liu, Jingjing Liu, Nicholas J. Belkin, Michael J. Cole, and Jacek Gwizdka. 2011. Using dwell time as an implicit measure of usefulness in different task types. *Proceedings of the American Society for Information Science and Technology Conference* 48, 1 (2011), 1–4.

Jingjing Liu, Michael J. Cole, Chang Liu, Nicholas J. Belkin, Jun Zhang, Jacek Gwizdka, Ralf Bierig, and Xiangmin Zhang. 2010a. Search behaviors in different task types. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries (JCDL'10)*. ACM, 69–78. DOI:http://doi.acm.org/10.1145/1816123.1816134

Jingjing Liu, Jacek Gwizdka, Chang Liu, and Nicholas J. Belkin. 2010b. Predicting task difficulty for different task types. In *Proceedings of the 73rd ASIS&T Annual Meeting (ASIS&T'10)*. American Society for Information Science, 16:1–16:10.

Ying-Hsang Liu and Nina Wacholder. 2008. Do human-developed index terms help users? An experimental study of MeSH terms in biomedical searching. *Proceedings of the American Society for Information Science and Technology* 45, 1 (2008), 1–16.

Jiyun Luo, Sicong Zhang, and Hui Yang. 2014. Win-win search: Dual-agent stochastic game in session search. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 587–596.

Gary Marchionini. 1989. Information-seeking strategies of novices using a full-text electronic encyclopedia. *Journal of the American Society for Information Science* 40, 1 (1989), 54–66.

Gary Marchionini and Hermann Maurer. 1995. The roles of digital libraries in teaching and learning. *Communications of the ACM* 38, 4 (1995), 67–75.

Robert E. Morrison. 1984. Manipulation of stimulus onset delay in reading: Evidence for parallel programming of saccades. *Journal of Experimental Psychology: Human Perception and Performance* 10, 5 (1984), 667–682.

Vanessa Murdock, Diane Kelly, W. Bruce Croft, Nicholas J. Belkin, and Xiao-Jun Yuan. 2007. Identifying and improving retrieval for procedural questions. *Information Processing & Management* 43, 1 (2007), 181–203. DOI:http://dx.doi.org/10.1016/j.ipm.2006.05.009

Alexander Pollatsek, Keith Rayner, and David A. Balota. 1986. Inferences about eye movement control from the perceptual span in reading. *Perception & Psychophysics* 40, 2 (1986), 123–130.

Liwen Qiu. 1993. Markov models of search state patterns in a hypertext information retrieval system. *Journal of the American Society for Information Science* 44, 7 (1993), 413–427.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124 (1998), 372–422.

Keith Rayner, Kathryn H. Chace, Timothy J. Slattery, and Jane Ashby. 2006. Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading* 10, 3 (2006), 241–255.

Keith Rayner and Susan A. Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition* 14, 3 (1986), 191–201.

Keith Rayner and Alexander Pollatsek. 1989. *The Psychology of Reading.* Lawrence Erlbaum Associates, Mahwah, NJ.

Erik D. Reichle, Alexander Pollatsek, and Keith Rayner. 2006. E–Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research* 7, 1 (2006), 4–22.

Erik D. Reichle, Keith Rayner, and Alexander Pollatsek. 2004. The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences* 26, 04 (2004), 445–476.

Eyal M. Reingold and Keith Rayner. 2006. Examining the word identification stages hypothesized by the EZ Reader model. *Psychological Science* 17, 9 (2006), 742–746.

Phoebe M. Roberts, Aaron M. Cohen, and William R. Hersh. 2009. Tasks, topics and relevance judging for the TREC Genomics Track: Five years of experience evaluating biomedical text information retrieval systems. *Information Retrieval* 12, 1 (2009), 81–97. DOI:http://dx.doi.org/10.1007/s10791-008-9072-x

Stephen E. Robertson. 1977. The probability ranking principle in IR. *Journal of Documentation* 33, 4 (December 1977), 294–304.

Dario D. Salvucci. 1999. Inferring intent in eye-based interfaces: Tracing eye movements with process models. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 254–261.

Carolin Strobl, Anne-Laure Boulesteix, and Thomas Augustin. 2007. Unbiased split selection for classification trees based on the Gini index. *Computational Statistics & Data Analysis* 52, 1 (2007), 483–501.

Elaine Toms, Tayze MacKenzie, Chris Jordan, Heather L. O'Brien, Luanne S. Freund, Sandra Toze, Emille Dawe, and Alexandra MacNutt. 2007. How task affects information search. In *Workshop Pre-Proceedings in Initiative for the Evaluation of XML Retrieval (INEX)*. 337–341.

Jochen Triesch, Dana H. Ballard, Mary M. Hayhoe, and Brian T. Sullivan. 2003. What you see is what you need. *Journal of Vision* 3 (2003), 86–94.

Chih-Hao Tsai and George W. McConkie. 1995. The perceptual span in reading Chinese text: A moving window study. In *Proceedings of the 7th International Conference on the Cognitive Processing of Chinese and Other Asian Languages*.

Ryen W. White, Paul N. Bennett, and Susan T. Dumais. 2010. Predicting short-term interests using activity-based search context. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*. ACM, 1009–1018.

Ryen W. White and Diane Kelly. 2006. A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. ACM, 297–306. DOI:http://dx.doi.org/10.1145/1183614.1183659

Barbara M. Wildemuth. 2004. The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology* 55, 3 (2004), 246–258.

Laurence B. Wolfe and Chein-I Chang. 1993. A complete sufficient statistic for finite-state Markov processes with application to source coding. *IEEE Transactions on Information Theory* 39, 3 (1993), 1047–1049. DOI:http://dx.doi.org/10.1109/18.256512

Yisong Yue, Yue Gao, Olivier Chapelle, Ya Zhang, and Thorsten Joachims. 2010. Learning more powerful test statistics for click-based retrieval evaluation. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 507–514.

Zhen Yue, Shuguang Han, Jiepu Jiang, and Daqing He. 2012. Search tactics as means of examining search processes in collaborative exploratory web search. In *Proceedings of the 5th Ph.D. Workshop on Information and Knowledge (PIKM'12)*. ACM, 59–66. DOI:http://dx.doi.org/10.1145/2389686.2389699

Received April 2014; revised August 2014; accepted October 2014