Interactive Faceted Query Suggestion for Exploratory Search: Whole-session Effectiveness and Interaction Engagement

Tuukka Ruotsalo

Department of Computer Science, Aalto University and University of Helsinki, Finland and Helsinki Institute for Information Technology HIIT

Giulio Jacucci

Department of Computer Science, University of Helsinki, Finland and Helsinki Institute for Information Technology HIIT

Samuel Kaski

Department of Computer Science, Aalto University, Finland and Helsinki Institute for Information Technology HIIT

Author Note

Corresponding author: Tuukka Ruotsalo, tuukka.ruotsalo@helsinki.fi

Abstract

The outcome of exploratory information retrieval is not only dependent on the effectiveness of individual responses to a set of queries, but also on relevant information retrieved during the entire exploratory search session. We study the effect of search assistance, operationalized as interactive faceted query suggestion, for both whole-session effectiveness and engagement through interactive faceted query suggestion. A user experiment is reported, where users performed exploratory search tasks, comparing interactive faceted query suggestion and a control condition with only conventional typed-query interaction. Data comprising of interaction and search logs show that the availability of interactive faceted query suggestion substantially improves whole-session effectiveness by increasing recall without sacrificing precision. The increased engagement with interactive faceted query suggestion is targeted to direct situated navigation around the initial query scope, but is not found to improve individual queries on average. The results imply that research in exploratory search should focus on measuring and designing tools that engage users with directed situated navigation support for improving whole-session performance.

Keywords: Whole-session relevance, Session search, Search assistance, Faceted Query Suggestions

Interactive Faceted Query Suggestion for Exploratory Search: Whole-session Effectiveness and Interaction Engagement

Introduction

An increasing fraction of search sessions span over multiple queries and involve exploratory search behavior (Marchionini, 2006). A user engaged in an exploratory search task is required to invest significant amounts of cognitive effort in identifying different aspects relevant to the task by evaluating intermediate results, reformulating queries, and inventing new queries to continue the search.

Users' targets in exploratory tasks are often intrinsically or topically diverse within a broad topical area (Raman, Bennett, & Collins-Thompson, 2014). For example, a user could be interested in a broad main topic of "machine vision", and then explore to specific aspects of the main topic, such as "edge detection", "pattern recognition", but also related but relevant topics, such as "optical sensors", or "deep learning".

A wide body of research has sought to address exploration support via search assistance, that is, interactive tools that assist users to browse or adjust their queries. These include, for example, faceted search (Yee, Swearingen, Li, & Hearst, 2003), techniques to help users formulate better queries (Shokouhi, 2013; Yee et al., 2003), and support for suggesting queries predicted using other users' search trails (White & Huang, 2010).

Previous studies have focused mostly on log analysis of short search sessions or task-based laboratory studies. The former has focused on analyzing relatively short search sessions that consist of a few clicks and query reformulations, and last a few minutes at best (Carterette, Clough, Hall, Kanoulas, & Sanderson, 2016; Raman et al., 2014). These studies have revealed the need and type of interaction that users rely on in a few subsequent queries, but are limited in understanding search behavior in longer exploratory search sessions.

Task-based laboratory studies, on the other hand, have studied various behavioral factors and their association with task factors, such as searchers pre-knowledge or topic familiarity (Duggan & Payne, 2008; Liu, Liu, & Belkin, 2013), task complexity (Byström & Järvelin, 1995; Liu et al., 2010), and task stage (Liu & Belkin, 2015). However, these studies have typically relied on a standard search user interfaces. Consequently, we know fairly little about the effects that changes in the search interface can have for search performance.

A line of research has also studied the effect of tools and interface designs on users' search behavior. Search assistance has been evaluated in restricted tasks, such as studying faceted classifications for filtering in known-item search (Yee et al., 2003) or clicks on other users' search trails (Capra, Arguello, Crescenzi, & Vardell, 2015). Despite the recognized importance of search assistance and the association of usage of assistance with session and task-level factors (Capra et al., 2015), previous work has provided limited evidence on the effect of search assistance on users' retrieval performance. Consequently, our knowledge on how the availability of search assistance affects users' retrieval performance, momentarily at single query-response level and over time in longer exploratory search sessions, is fairly limited.

Our previous work has shown incresed retrieval effectiveness and user satisfaction on a combination of intent modeling and visualization (Ruotsalo, Jacucci, Myllymäki, & Kaski, 2014; Ruotsalo et al., 2013). The present work focuses on examining differences in the system's retrieval performance in response to different types of user interfaces at query-response and whole-session levels. The studied systems vary in their availability of interactive faceted query suggestion. Here, interactive facets are defined as a set of meaningful labels organized in such a way as to reflect the concepts relevant to a domain. The facets can be used as probabilistic filters for a complex data structure by simultaneously filtering objects for maximum flexibility in information retrieval, a technique ofter referred as faceted search (Niu, Fan, & Zhang, 2019; Yee et al., 2003).

The main objective of the present study is to investigate the effect of search assistance, operationalized as interactive faceted query suggestions on a session-level information retrieval performance. To this end, we focus on studying whether the availability interactive faceted query suggestion, is effective for exploratory search where the user's target is to maximize long-term whole-session output, or at single query-response, or both. In addition, we quantify how the availability of search assistance affects the user's interaction with the search engine; whether it can reduce, complement or mitigate the need to invent and type queries.

More precisely, we ask the following research questions: 1) Is interactive faceted query suggestion associated with improved whole-session effectiveness? 2) Is interactive faceted query suggestion associated with improved query-response effectiveness? 3) Does interactive faceted query suggestion engage users with interaction to direct search?

As a result, our study helps to understand how search assistance, operationalized as interactive faceted query suggestion, affects users' query-response and whole-session retrieval performance in exploratory search tasks, and whether users engage to revise and adjust their information needs using such search assistance.

A user study was conducted, in which participants are carrying out extensive exploratory search tasks with and without interactive faceted query suggestion. Data comprising of search and interaction logs show that whole-session output in exploratory search is associated with the availability of interactive faceted query suggestion, but it provides limited or no advantage at query-response level. Participants seem to rely on interactive faceted query suggestion to complement, but not substitute typed query interaction, suggesting that it is mainly used for supporting exploration beyond the present query context.

Related Work

Our work builds on related work in session search, supporting task-based information needs, and studies on search assistance tools and techniques, which we review below. Then we explicate the contributions of the work in relation to previous work.

Session-based search

Information retrieval research has primarily focused on improving retrieval for a single query-response or sessions consisting of relatively short sequences of queries and clicks (Kanoulas, Carterettey, Hallz, Cloughx, & Sanderson, 2012). Less attention has been devoted for long-lasting exploratory search scenarios (Marchionini, 2006), even though exploratory search is a challenging research problem and can have potentially high impact for end users (Vakkari, 2001). Recently, session-based retrieval has become an increasingly popular research area. Studies have identified trends in user search sessions and introduced methods to improve search for such sessions (Guan, Zhang, & Yang, 2013; He, Bron, & de Vries, 2013; Raman et al., 2014).

An advantage of previous studies is that they employ large scale simulations by using existing log collections. More generally, research in session search has benefited from the introduction of the Session Track at TREC (Kanoulas et al., 2012), but has also limited the scope to the interaction types recorded in the TREC sessions. The key differences between the studies using TREC data and our study are that the TREC sessions are relatively short, typically including only few subsequent queries. Previous studies also rely on simulated retrieval performance based on these session logs available in the TREC collection.

It is unclear, however, if approaches and findings from studies using log data would lead to performance gains in real life information seeking contexts that go beyond simulations. To this end, previous approaches typically make assumptions on the interaction with the search engine being limited to typed queries, clicks, and page views, and the sessions being short and consist of samples that represent only partial user tasks (Vuong, Saastamoinen, Jacucci, & Ruotsalo, n.d.). This is a plausible assumption in order to improve the results for search systems via optimizing rankers relying on typed query interaction and click models. However, it assumes simple interactions, short sessions and at best a reactive user involvement in the search process. Conversely, we study long sessions in a user study allowing a comparison between two independent system conditions separating the effect of interactive faceted query suggestion in realistic exploratory search tasks.

Task-level User Behavior

Many exploratory search tasks, such as literature surveys or vacation planning, have been shown to elicit very different search behavior than what is observed in look-up searches (Liu, Liu, Cole, Belkin, & Zhang, 2012). While look-up searches have been thought to be the most common search type in Web search, it has been suggested that exploratory tasks are relatively common and can represent over a quarter of all tasks (Raman et al., 2014). Exploratory tasks may require the user to devote extensive amounts of time, queries, and other interactions to complete the task (Jones & Klinkner, 2008; Liu et al., 2012).

Whole-session analysis are affected by the corresponding tasks, and task factors have been associated with various search behaviors. Task-based laboratory studies have revealed associations between user factors, behavioral factors, and task factors. For example, searchers pre-knowledge or topic familiarity has been shown to be associated with variance in search behavior and performance (Liu et al., 2013), lead to less time being spent on searching, and faster decisions on issuing queries (Duggan & Payne, 2008). Task type has also been found to be associated with search behavior, including task completion time and the time taken to decide whether a document is useful or not (Liu et al., 2010). Moreover, associations haveo been shown between task performance and task complexity (Byström & Järvelin, 1995; Liu et al., 2010), and the stage of the task in which the search is conducted (Liu & Belkin, 2015).

Despite the knowledge gained from several empirical studies on the factors and their associations at task-level search behavior, these studies have typically relied on a standard search user interfaces and search engines. Consequently, we still know fairly little about the relationship of task success and the system features, such as the specifics of the search engine or the search user interface.

Empirical Studies on the Effects of Search Assistance

The early studies in search user interfaces have already demonstrated that even simple search assistance, such as manually curated faceted filtering interfaces, can be effective for various information retrieval tasks (Koren, Zhang, & Liu, 2008; Yee et al., 2003).

Also user-centered query autocompletion interfaces have been proven to be effective as evidenced by a series of personalized query auto completion approaches (Bar-Yossef & Kraus, 2011; Cai & de Rijke, 2016; Shokouhi, 2013). In particular, short-term search context of a user and query-level features have been shown to predict subsequent queries better than more general user features (J.-Y. Jiang, Ke, Chien, & Cheng, 2014). Query-autocompletion, however, is limited to suggesting queries that the user is already typing and, while effective, may be more useful for specifying existing information needs rather than allowing exploration to new directions in the information space.

In general, it has been shown that interaction with search assistance is typically more frequent in case of more complex tasks (Capra et al., 2015) and that users prefer specific hints over general ones. Users have a natural ability to recognize specific hints and query terms (Kangassalo, Spapé, Jacucci, & Ruotsalo, 2019), and specific search hints have been demonstrated to effectively improve searcher success rates and reduce perceived effort, while generic search hints can be detrimental in both search effectiveness and user satisfaction (Savenkov & Agichtein, 2014).

Some studies using search engine logs have shown that search assistance may also raise concerns about the correctness and utility of the results. It has been observed that struggling and exploring behaviors are interleaved in Web search (Hassan, White, Dumais, & Wang, 2014) and that increased usage of search assistance can be related to uncertainty of search success (Capra et al., 2015).

More detailed behavioral studies have shown that users vary in their use of search assistance, in particular query-autocompletion and suggestion, in terms of search activeness, browsing style, and query reformulation. The tendency to use assistance may also vary as search sessions progress; users shift their interests to focus less on the top results but more on results ranked at lower positions in browsing (J. Jiang, He, & Allan, 2014). Search assistance may also be used to browse around the information space. It has been shown that using search assistance and increasing interaction with search systems can degrade precision, but lead to better task outcomes (Vakkari & Huuskonen, 2012). Thus, human effort can compensate bad momentary retrieval results. These findings highlight the importance of system support to effectively interact with the search system to direct the search toward novel results.

Search Assistance Tools and Techniques

Various tools have been proposed for whole-session support via different kinds of interactive search assistance that offer affordances to filter search results and reframe and suggest queries. Proposed techniques include query auto-completion (J.-Y. Jiang et al., 2014; Li et al., 2014), query recommendation (Baeza-yates, Hurtado, & Mendoza, 2004; Boldi et al., 2008), and intent prediction for re-ranking (Hu, Zhang, Chen, Wang, & Yang, 2011). Research has also explored task-aware (Capra et al., 2015; Feild & Allan, 2013) and semantic models (Bing, Lam, Wong, & Jameel, 2015) for query recommendation.

Similar to our faceted approach, variety of methods have been proposed for modeling search intents for diversification and query suggestion. Improvements over conventional diversification methods have been achieved by clustering query refinements for intent detection (Sadikov, Madhavan, Wang, & Halevy, 2010) and using click-through data to intent-aware diversification (Hu et al., 2011). Researchers have also developed diversification methods that can make meaningful query suggestions context-aware by taking into account the immediately preceding queries as a context in query suggestion (Cao et al., 2008, 2009). Other techniques use query clustering to similar intent classes (Cheung & Li, 2012) and suggest a diverse set of queries using intent models that utilize a short-term context using the user's behaviour within the current search session, such as the previous query, the documents examined, and the candidate query suggestions that the user has discarded (Kharitonov, Macdonald, Serdyukov, & Ounis, 2013), or the page context that the user has browsed (Cheng, Gao, & Liu, 2010).

The main user interface oriented approaches to present richer information and allow navigation include filtering by facets (Yee et al., 2003), result visualization and navigation through clusters (Hearst, 1995). While these approaches provide means for visualizing search results and interacting with the underlying information space, they are not focused on supporting whole-session outcomes through modeling search sessions.

Various search systems employ visualization of the resulting information to enable faster relevance judgment and effective feedback (Matejka, Grossman, & Fitzmaurice, 2012; Terveen, Hill, & Amento, 1999). A variety of visualization approaches of search results have been explored, including multiple linked lists, scatter plots, graphs and their combinations (Kules, Wilson, Schraefel, & Shneiderman, 2008; Stasko, Görg, & Liu, 2008). These types of visual search systems are distinguished from familiar query composition based systems by their emphasis on rapid filtering to reduce result sets, progressive refinement of search parameters, continuous reformulation of goals, and visual scanning to identify results (Ahlberg & Shneiderman, 1994; Klouche et al., 2015).

Contributions

Our contributions can be summarized as follows:

1. According to our knowledge, we report the first user study quantifying the benefits of interactive faceted query suggestion for whole-session retrieval effectiveness and user engagement in exploratory search tasks.

2. Our results show that interactive faceted query suggestion is used for directed situated navigation to complement, but not substitute typed query interaction. This suggests that search assistance implemented as interactive faceted query suggestion is used for supporting exploration beyond the initial query context.

3. Our results suggest that whole-session relevance in exploratory search is associated with the availability of search assistance and that search assistance provides limited or no advantage at individual query-response level.

Methodology

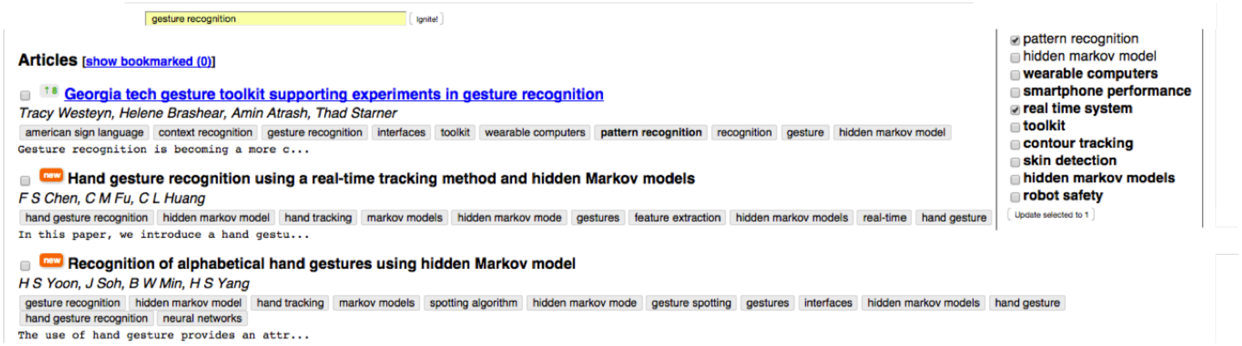
The target of the study was to determine whether the availability of search assistance, operationalized as interactive faceted query suggestion, has an effect on users' momentary query-response or whole-session retrieval performance. A user experiment was conducted to provide the methodological means to go beyond the earlier approaches that utilized simulations based on log data. The target of the present study is to quantify the effect of interactive faceted query suggestion for users' information retrieval performance *in-situ* in task-based user experiments. The research questions, experimental design, procedure, data, system conditions, tasks and apparatus, measures, and the details of the experiments are described in detail in the following subsections.

Experimental Design

The experiment followed a 2 × 2 between-subjects design with two system conditions: the experimental condition with search assistance, operationalized as interactive faceted query suggestion, and a control condition without the search assistance. This design was chosen to avoid the carryover effects, as each participant only used one of the systems and performed a single task with one system.

Experimental Search User Interface

A search user interface, shown in Figure 1, was designed to mimic conventional features of de-facto search user interfaces. It presents each document as a short snippet that can be expanded by clicking, and the associated metadata of the documents.



Articles [show bookmarked (0)]

□ ↑8 Georgia tech gesture toolkit supporting experiments in gesture recognition
Tracy Westeyn, Helene Brashear, Amin Atrash, Thad Starner
american sign language | context recognition | gesture recognition | interfaces | toolkit | wearable computers | **pattern recognition** | recognition | gesture | hidden markov model
`Gesture recognition is becoming a more c...`

□ new **Hand gesture recognition using a real-time tracking method and hidden Markov models**
F S Chen, C M Fu, C L Huang
hand gesture recognition | hidden markov model | hand tracking | markov models | hidden markov mode | gestures | feature extraction | hidden markov models | real-time | hand gesture
`In this paper, we introduce a hand gestu...`

□ new **Recognition of alphabetical hand gestures using hidden Markov model**
H S Yoon, J Soh, B W Min, H S Yang
gesture recognition | hidden markov model | hand tracking | markov models | spotting algorithm | hidden markov mode | gesture spotting | gestures | interfaces | hidden markov models | hand gesture | hand gesture recognition | neural networks
`The use of hand gesture provides an attr...`

☑ pattern recognition
□ hidden markov model
□ **wearable computers**
□ **smartphone performance**
☑ **real time system**
□ **toolkit**
□ **contour tracking**
□ **skin detection**
□ **hidden markov models**
□ **robot safety**

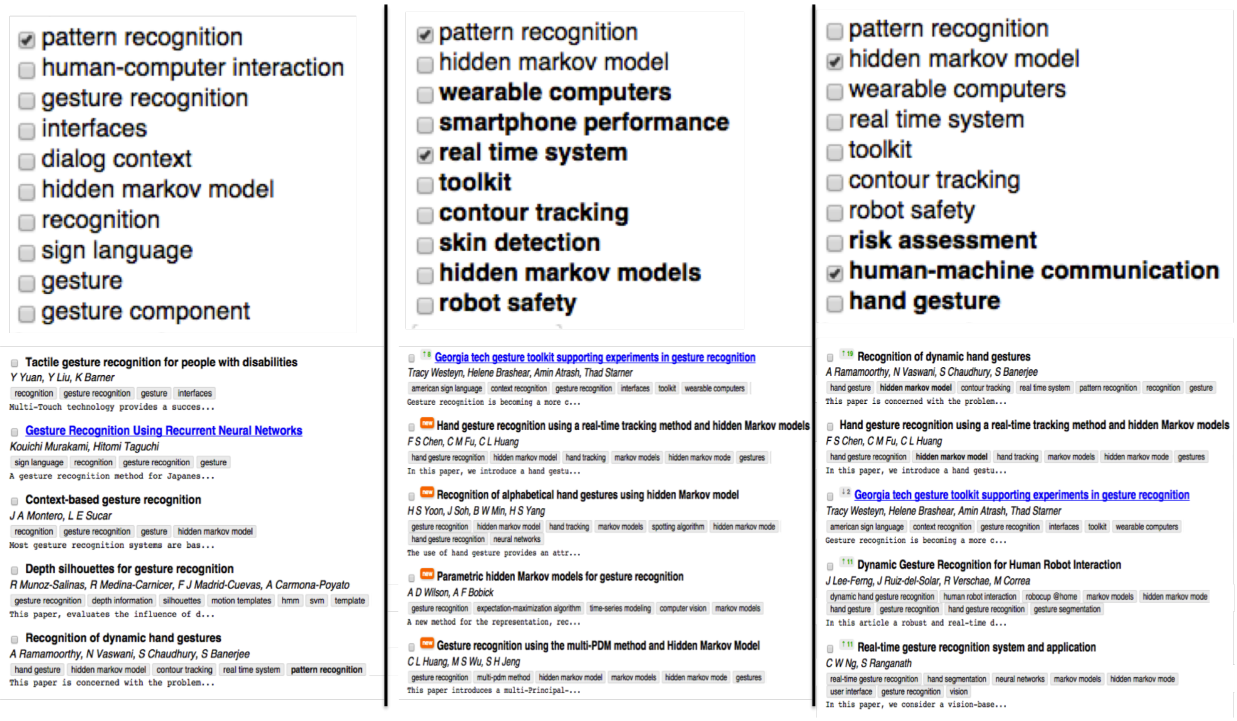
Figure 1**. The search user interface design with conventional typed query interaction and result list (left), and search assistance presented as a dynamic facet; keywords that the user can reward (right). Links to documents for which abstracts have been opened are colored in blue and documents that appear for the first time on the screen have been associated with a symbol "new" in front of the document title. The position change of a document is indicated in front of the document title with an arrow symbol along with the change in rank.**

The interactive faceted query suggestion component (right side of Figure 1) was designed to show ten facets at each iteration. The computational model to estimate the facets is described in detail in (Ruotsalo et al., 2014, 2013, 2018). The model utilizes an online machine learning algorithm to estimate keywords that are similar based on user feedback. It prefers keywords similar to the ones that the user has selected so far (exploit) and simultaneously keywords that are uncertain (explore) (Auer, 2003). This procedure ranks high keywords that are relevant for the user's interaction history, but guarantees that new and unseen keywords are available for interaction. The interface allows a conventional typed query interaction and adjusted query by clicking a keyword representing a facet. The typed query along with the facet selections are then used to compose a query to rank the documents.

The ranking model is a state-of-the-art unigram language model with Dirichlet smoothing (Zhai & Lafferty, 2004). For the experiment we fixed the parameters in the system to the following values: The number of retrieved documents that were visualized for the user at each iteration and ranked by the language model was set to 20. We set $\mu = 2000$ for the Dirichlet smoothing. The number of facets included in the interactive faceted query suggestion component was set to 10.

In order to provide the users with information about the change of the ranking, at each iteration, each document was presented with a marker indicating if it is a new document (orange icon) and an indication of the change in the rank (green icon) and a link to the full document.

Figure 2 illustrates an interaction sequence with the assistant. The left panel shows the document list and the assistant after the user has searched "gesture recognition". The assistant offers keywords, such as "pattern recognition", "gesture recognition", "interfaces", "dialog context", and "sign language" and the documents are generally about gesture recognition. The user provides feedback using the assistant by clicking the "pattern recognition" keyword. The middle panel shows the assistant after the feedback. The assistant reflects the pattern recognition aspect and offers keywords, such as "hidden Markov model" and "real time systems". The user then provides feedback on the keyword "real time system" and receives a new list of documents with the focus of pattern recogition in real-time gesture recognition. The third panel shows the third iteration after the user has selected the keyword "human-machine communication", which results in retrieved documents that are more generally about using pattern recognition models in gesture-based human-computer interaction.



☑ pattern recognition
□ human-computer interaction
□ gesture recognition
□ interfaces
□ dialog context
□ hidden markov model
□ recognition
□ sign language
□ gesture
□ gesture component

□ **Tactile gesture recognition for people with disabilities**
Y Yuan, Y Liu, K Barner
recognition | gesture recognition | gesture | interfaces
`Multi-Touch technology provides a succes...`

□ **Gesture Recognition Using Recurrent Neural Networks**
Kouichi Murakami, Hitomi Taguchi
sign language | recognition | gesture recognition | gesture
`A gesture recognition method for Japanes...`

□ **Context-based gesture recognition**
J A Montero, L E Sucar
recognition | gesture recognition | gesture | hidden markov model
`Most gesture recognition systems are bas...`

□ **Depth silhouettes for gesture recognition**
R Munoz-Salinas, R Medina-Carnicer, F J Madrid-Cuevas, A Carmona-Poyato
gesture recognition | depth information | silhouettes | motion templates | hmm | svm | template
`This paper, evaluates the influence of d...`

□ **Recognition of dynamic hand gestures**
A Ramamoorthy, N Vaswani, S Chaudhury, S Banerjee
hand gesture | hidden markov model | contour tracking | real time system | **pattern recognition**
`This paper is concerned with the problem...`

☑ pattern recognition
□ hidden markov model
□ **wearable computers**
□ **smartphone performance**
☑ **real time system**
□ **toolkit**
□ **contour tracking**
□ **skin detection**
□ **hidden markov models**
□ **robot safety**

□ ↑8 **Georgia tech gesture toolkit supporting experiments in gesture recognition**
Tracy Westeyn, Helene Brashear, Amin Atrash, Thad Starner
american sign language | context recognition | gesture recognition | interfaces | toolkit | wearable computers
`Gesture recognition is becoming a more c...`

□ new **Hand gesture recognition using a real-time tracking method and hidden Markov models**
F S Chen, C M Fu, C L Huang
hand gesture recognition | hidden markov model | hand tracking | markov models | hidden markov mode | gestures
`In this paper, we introduce a hand gestu...`

□ new **Recognition of alphabetical hand gestures using hidden Markov model**
H S Yoon, J Soh, B W Min, H S Yang
gesture recognition | hidden markov model | hand tracking | markov models | spotting algorithm | hidden markov mode | hand gesture recognition | neural networks
`The use of hand gesture provides an attr...`

□ new **Parametric hidden Markov models for gesture recognition**
A D Wilson, A F Bobick
gesture recognition | expectation-maximization algorithm | time-series modeling | computer vision | markov models
`A new method for the representation, rec...`

□ new **Gesture recognition using the multi-PDM method and Hidden Markov Model**
C L Huang, M S Wu, S H Jeng
gesture recognition | multi-pdm method | hidden markov model | markov models | hidden markov mode | gestures
`This paper introduces a multi-Principal-...`

□ pattern recognition
☑ hidden markov model
□ wearable computers
□ real time system
□ toolkit
□ contour tracking
□ robot safety
□ **risk assessment**
☑ **human-machine communication**
□ **hand gesture**

□ ↑19 **Recognition of dynamic hand gestures**
A Ramamoorthy, N Vaswani, S Chaudhury, S Banerjee
hand gesture | **hidden markov model** | contour tracking | real time system | pattern recognition | recognition | gesture
`This paper is concerned with the problem...`

□ **Hand gesture recognition using a real-time tracking method and hidden Markov models**
F S Chen, C M Fu, C L Huang
hand gesture recognition | **hidden markov model** | hand tracking | markov models | hidden markov mode | gestures
`In this paper, we introduce a hand gestu...`

□ ↓2 **Georgia tech gesture toolkit supporting experiments in gesture recognition**
Tracy Westeyn, Helene Brashear, Amin Atrash, Thad Starner
american sign language | context recognition | gesture recognition | interfaces | toolkit | wearable computers
`Gesture recognition is becoming a more c...`

□ ↑11 **Dynamic Gesture Recognition for Human Robot Interaction**
J Lee-Ferng, J Ruiz-del-Solar, R Verschae, M Correa
dynamic hand gesture recognition | human robot interaction | robocup @home | markov models | hidden markov mode | hand gesture | gesture recognition | hand gesture recognition | gesture segmentation
`In this article a robust and real-time d...`

□ ↑11 **Real-time gesture recognition system and application**
C W Ng, S Ranganath
real-time gesture recognition | hand segmentation | neural networks | markov models | hidden markov mode | user interface | gesture recognition | vision
`In this paper, we consider a vision-base...`

***Figure 2*. An example sequence of three iterations of interaction with the interactive faceted query suggestion component (top row) and the corresponding top-5 documents (bottom row) for each iteration. The user interactions to reward the prediction model are shown as activated tick boxes. The bold keywords are keywords presented for the first time within the session.**

Control Condition

The system used as the control condition used exactly the same ranking model, and the system was indexed with the same document collection, but the interactions were constrained to typed-keyword queries. The user interface was the same as in the experimental condition, but the interactive faceted query suggestion component, shown on the right side of Figure 1, was removed.

Data

We used a dataset of over 50 million scientific documents from the following data sources: the Web of Science prepared by Thomson Reuters, Inc., the digital library of the

Association of Computing Machinery (ACM), the Digital Library of the Institute of Electrical and Electronics Engineers (IEEE), and the digital library of Springer.

Search Tasks

We chose a task type that is complex enough to ensure that exploration is necessary for participants to acquire the information to accomplish the task, and complex enough to allow participants to choose the kind of interaction that best supports solving the tasks.

The tasks were defined to be scientific writing scenarios as follows. The participants were asked to prepare materials and an outline for writing an essay on a given topic. The assignments were (1) to search for relevant articles that they would be likely to use as references in their essay (2) to write an essay outline to structure the information.

Two post-doctoral researchers with computer and information science background were recruited to define two information seeking tasks. The task fields chosen by the experts were “semantic search” and “robotics”. The experts wrote task descriptions using the following template: “Imagine that you are writing a scientific essay on the topic. Search scientific information that you find useful for this essay”. In order to provide clear goals for exploration, the experts were asked to provide questions about specific aspects of the topic. The question defined by the experts for the robotics tasks was: “What are the sub-fields, application areas and algorithms commonly used in the field of robotics?”, while the question for the semantic search task was: “What are the techniques used to acquire semantics, methods used in practical implementation, organization of results, and the role of semantic Web technologies in semantic search?”. To ensure that the participants would carry out the search tasks as realistically as possible, the participants were asked to both search for documents to support their answers to these questions and to write short answers under each question to fill in the essay outline.

Participants

Twenty participants from two universities in the Helsinki capital area in Finland were recruited to participate in the study. All the participants were graduate students or postdoctoral researchers with a background in computer or information science. The participants were between 20–38 years old. There were 7 female and 13 male participants. All the participants had a technical background. They did not receive any compensation for participating the experiment. Through a prior background survey we ensured that every participant had conducted a literature search before and required participants to self-assess their prior knowledge on the topic of the search task. Participants with high or low prior knowledge were excluded from the study in order to avoid cognitive and expectation biases. Prior knowledge was self-assessed on a scale of 1 to 5 ((1) no knowledge at all, (2) some knowledge, (3) moderate knowledge, (4) knowledgeable, (5) expert knowledge). We only allowed participation if the prior knowledge was rated between 2 and 4.

Procedure

The basic protocol for each experiment scenario was the following. The system was first demonstrated to allow the participant to get familiar with the document collection, interface, and the functionality of the system they were using (5 minutes). All features of the system were demonstrated, but in the experiments the participants were free to use any of the features offered in the system, and they were not forced or encouraged to use particular features. Then the task was explained for the participant (5 minutes). After these the participant started the task. The participant was notified after 30 minutes to finalize the task and in case the participant was still working when 32 minutes was elapsed, the task was terminated. All participants used the full 30 minutes to complete the task.

Apparatus

The experiments were performed in an office-like environment using standard equipment (20”–24” monitor, mouse, and keyboard). The demonstration of the system was done by the instructor using a separate laptop computer. A separate computer was used with a purpose of not intervening with the machine that was used for the actual task, for example by causing queries to be cached. The participants were able to try out the system with an example query if they wanted, but they were not forced to do so.

Interaction behavior				
Measure	**Control**	**Query suggestion**	**p-val**	**Diff**
Typed query	8.7 ±1.31	7.1 ±1.41	p=0.41	-18.3%
Reformulation	5.7 ±1.37	4.6 ±1.09	p=0.25	-19.3%
Assistance	–	14.1 ±3.14	–	–
Typed + assistance	**14.4** ±3.15	**21.2**±4.20	**p=0.02**	**+179%**

Table 1

Interaction behavior results. The differences in bold are significant between the interactive faceted query suggestion condition and the control condition (unpaired Wilcoxon Signed-Rank Test).

Data Logging

All interactions performed by the users with the systems and the information retrieved in response to these interactions were logged. Data logged from the interactions in the interactive faceted query suggestion condition included typed queries, retrieved documents, the keywords representing the facets and interaction with the facets, and times at which each interaction occurred. Similarly, from the control system we logged the typed queries, retrieved documents, and the times at which participants entered queries to retrieve documents. The typed queries, query reformulations, and interactions with the keywords are further referred as interactions. Momentary query-response level measures were computed using a system response to an individual interaction and whole-session level measures were computed using the cumulative sets of information retrieved within the whole session.

Pooling and Relevance Assessments

After the completion of the experiments, the experts who designed the tasks conducted relevance assessments. An evaluation set consisting of 4649 documents (1491 in the semantic search task and 3158 in the robotics task). was created by pooling all retrieved documents. The documents in the pool were assessed according to binary relevance by the same experts who defined the tasks. A binary scale was chosen as it has been shown to sufficient for measuring retrieval performance according to topical relevance. It is also easy to use for assessors and robust across assessors (Kekäläinen, 2005). The assessments consisted of 3020 relevant documents (768 in the semantic search task and 2252 in the robotics task). To measure the inter-annotator agreement between the two experts, an overlapping subset of approximately 20% of the articles was assessed by both experts. The Cohen Kappa test indicated a substantial agreement between the experts (Kappa = 0.71, $p < 0.001$).

Measures

We focused on whole-session retrieval performance and momentary query-response retrieval performance. In addition, we quantified interaction behavior in order to understand the usage of the interactive faceted query suggestion component.

Search behavior was measured using the frequency of interactions. In the control condition the interactions recorded were typed queries and query reformulations. Following Huang and Efthimiadis (2009), a query was considered reformulated if it shared at least one word with the previously issued query. In the experimental condition we also recorded the interactions with the interactive faceted query suggestion component.

The rationale was that interactions were expected to increase in the interactive faceted query suggestion condition and the frequency of typed queries and query reformulations would decrease as the users would increasingly rely on the interactive faceted query suggestion component to adjust their search.

Retrieval performance was measured using variants of precision and recall. The first variant was computed to measure momentary query-response effectiveness in response to each interaction, i.e., the measures were computed at a single interaction level. The second variant was computed as a cumulative measure over the elapsed task time, i.e., the documents that users found were cumulatively added to a set avoiding duplicates to contribute to the measure, and characterized an average whole-session relevance. The denominator for recall was the total number of relevant documents assessed by the assessors and found by any of the participants when performing the task, i.e. maximum recall could be achieved if all relevant documents found by any of the participants for that task were found. This measure is similar to the residual versions of precision and recall proposed in (Qvarfordt, Golovchinsky, Dunnigan, & Agapie, 2013). The measures were computed at each iteration using the top-20 documents. The top-20 cutoff was chosen because the participants saw top-20 ranked documents in the user interface at each search iteration. We also compute the measures at the top-10 level for clarity. The momentary query-response recall included duplicates, i.e. it penalized for repeating queries that resulted in retrieving the same documents that were already been retrieved in the previous iterations. Then, precision and recall were computed at each iteration for the top twenty documents. These top twenty documents were also added to the cumulative set at each iteration to allow measuring the cumulative performance. The rationale for selecting the cutoff of 20 documents was based on the system configuration; in the experiments the participants were presented with 20 documents in response to queries or interactions with the interactive faceted query suggestion component.

We made a clear choice to not use session-level discounted cumulative gain (Järvelin, Price, Delcambre, & Nielsen, 2008) as that measure discounts for documents found later in the search session, which is against our intuition that exploratory search sessions should be evaluated based on session-level results rather than how fast information can be retrieved.

Hypotheses

Based on the research questions and the experimental design, the following hypotheses were defined:

H1: Whole-session effectiveness is higher for the condition with interactive faceted query suggestion than it is for the condition without interactive faceted query suggestion.

H2: Query-response effectiveness is higher for the condition with interactive faceted query suggestion than it is for the condition without interactive faceted query suggestion.

H3: Participants perform more interactions in the interactive faceted query suggestion condition and with the interactive faceted query suggestion component than they do in tbe condition without interactive faceted query suggestion.

Results

The experiment sought answers to whether the availability of search assistance, operationalized as interactive faceted query suggestion, is associated with improved whole-session or momentary query-response search performance. Next, the main results and their associations with interaction behavior are presented.

Whole-session Effectiveness

The effectiveness results are shown in Table 2. Cumulative whole-session effectiveness was found to be significantly improved in terms of recall (Wilcoxon Signed-Rank test, W =

<table>
<tr><th colspan="5">Retrieval performance</th></tr>
<tr><th>Measure</th><th>Control</th><th>Query suggestion</th><th>p-val</th><th>Diff</th></tr>
<tr><td colspan="5"><b>Whole-session (Cumulative)</b></td></tr>
<tr><td>P@10</td><td>0.75 ±0.04</td><td>0.83 ±0.03</td><td>p=0.14</td><td>+10.5%</td></tr>
<tr><td><b>R@10</b></td><td><b>0.02</b> ±0.006</td><td><b>0.04</b> ±0.02</td><td><b>p=0.01</b></td><td><b>+89.8%</b></td></tr>
<tr><td>P@20</td><td>0.76 ±0.04</td><td>0.83 ±0.03</td><td>p=0.21</td><td>+9.2%</td></tr>
<tr><td><b>R@20</b></td><td><b>0.05</b> ±0.007</td><td><b>0.09</b> ±0.01</td><td><b>p=0.01</b></td><td><b>+80.1%</b></td></tr>
<tr><td colspan="5"><b>query-response level (Momentary)</b></td></tr>
<tr><td>P@10</td><td>0.83 ±0.06</td><td>0.86 ±0.09</td><td>p=0.46</td><td>+3.4%</td></tr>
<tr><td>R@10</td><td>0.003 ±0.0002</td><td>0.003 ±0.0009</td><td>p=0.62</td><td>+1.8%</td></tr>
<tr><td>P@20</td><td>0.79 ±0.07</td><td>0.85 ±0.06</td><td>p=0.43</td><td>+7.6%</td></tr>
<tr><td>R@20</td><td>0.005 ±0.0005</td><td>0.006 ±0.0005</td><td>p=0.43</td><td>+20%</td></tr>
</table>

Table 2

Retrieval performance results measured at the top-20 and top-10 cutoff levels. The differences in bold are significant between the interactive faceted query suggestion condition and the control condition (unpaired Wilcoxon Signed-Rank Test)

75, p-value = 0.013). The mean cumulative recall was 0.09 for the condition with interactive faceted query suggestion and 0.05 for the control condition indicating over 80% improvement (Table 2, Whole-session). It is important to note that the absolute figures are low because on average an individual user was only able to retrieve a small portion of all relevant material in a 30-minute session. The participants retrieved on average 136 and 76 relevant documents, respectively. The improvement in recall was found to be linear throughout the search session and precision behaved similarity throughout the session (Figure 3). The effect sizes for recall are dependent on the amount of documents selected to the cumulative set at each iteration. The top-20 cutoff shows higher absolute effect sizes. However, the relative improvement is not dependent on the top-k cutoff level, but shows a robust and significant effect also for the top-10 cutoff level. No significant differences were found between the conditions with regard to precision (Wilcoxon Signed-Rank test, W = 61, p-value = 0.211). This suggests that the participants, in the interactive faceted query suggestion condition, were able to retrieve substantially more relevant materials without compromising precision.

This result confirms H1 for recall: Whole-session effectiveness is higher in terms of recall for the interactive faceted query suggestion condition than it is for the control condition.

Momentary Query-response Effectiveness

No significant differences were found in momentary query-response effectiveness (Table 2, query-response level). The mean precision (Wilcoxon Signed-Rank test, W = 55, p-value = 0.427) and recall (Wilcoxon Signed-Rank test, W = 55, p-value = 0.434) in response to an individual interaction were found to be similar throughout the sessions, i.e., we did not find any significant differences between the conditions. The mean recall for the relevant documents in response to an individual interaction for the control condition was 0.005 and 0.006 for the interactive faceted query suggestion condition. The participants retrieved on average 7.5 and 9 relevant documents (in the top-20 results) in response to an individual query, respectively. The temporal analysis (Figure 3) shows that the contribution of an individual interaction (query-response) to recall and the precision of documents were also found to be relatively stable throughout the search session for both conditions. Similarly to the cumulative whole-session analysis, the effect sizes for recall are dependent on the amount of documents selected to the cumulative set at each iteration, but the dfifferences are insignificant at the momentary query-response level.

This result does not allow to confirm H2: Query-response effectiveness is not higher

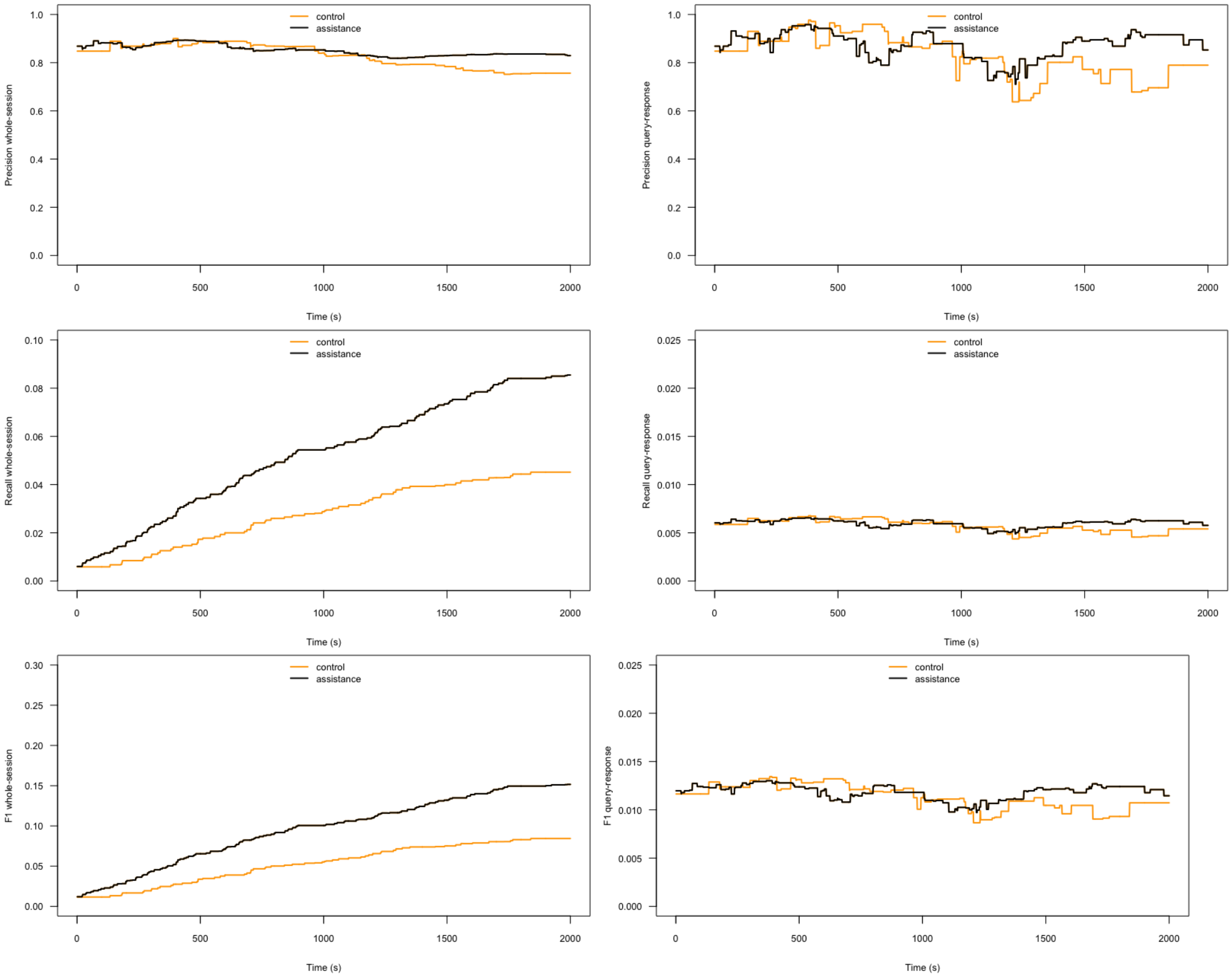


***Figure 3*. Temporal plots of the whole-session and momentary query-response effectiveness. The effectiveness measure on the y-axis and elapsed task time on the x-axis. The whole-session results are in the left column and the corresponding momentary query-response results are in the right column. Precision for relevant documents (upper row), recall for relevant documents (middle row), and F1 of relevant documents (lower row). The black line shows performance for interactive faceted query suggestion condition and the yellow line for the control condition.**

for the condition with interactive faceted query suggestion than it is for the control condition.

Interaction Behavior

Interaction behavior was analyzed to reveal the query typing behavior and amount of interactions the participants performed within the search session. The results are temporally illustrated in Figure 4. No significant differences were found in the amount of typed queries or typed query reformulations (Table 1). The participants who used the baseline system typed on average 8.7 queries and reformulated 7.1 times. The participants in the interactive faceted query suggestion condition typed 7.1 queries and reformulated 4.6 queries on average. No difference was found either in typed-query frequency (Wilcoxon Signed-Rank test, W = 34.5, p-value = 0.409) nor in typed-query reformulation (Wilcoxon Signed-Rank test, W = 59.5, p-value = 0.252) between the conditions. A possible explanation for the equal amount of typed queries across conditions could have been that participants were inspired by the suggestions, but interacted by typing queries. However, the typed-queries did not overlap with the interactive faceted query suggestion. Therefore, the amount of typed queries could not be attributed to the interactive faceted query suggestions, but query suggestions rather elicited more interactions in addition to typed queries.

The participants in the interactive faceted query suggestion condition performed significantly more interactions, on average 21.2 interactions, compared to the participants in the control condition who performed on average 8.7 interactions (Wilcoxon Signed-Rank test, W = 74.5, p-value = 0.0177). The substantial increase in interactions with the system with interactive faceted query suggestion indicates that the participants, in the control condition, were not willing or able to express their intentions to direct their search, or that the system did not help them to elicit exploration.

This result confirms H3: Participants perform more interactions in the interactive faceted query suggestion condition and with the interactive faceted query suggestion component than they do in tbe control condition.

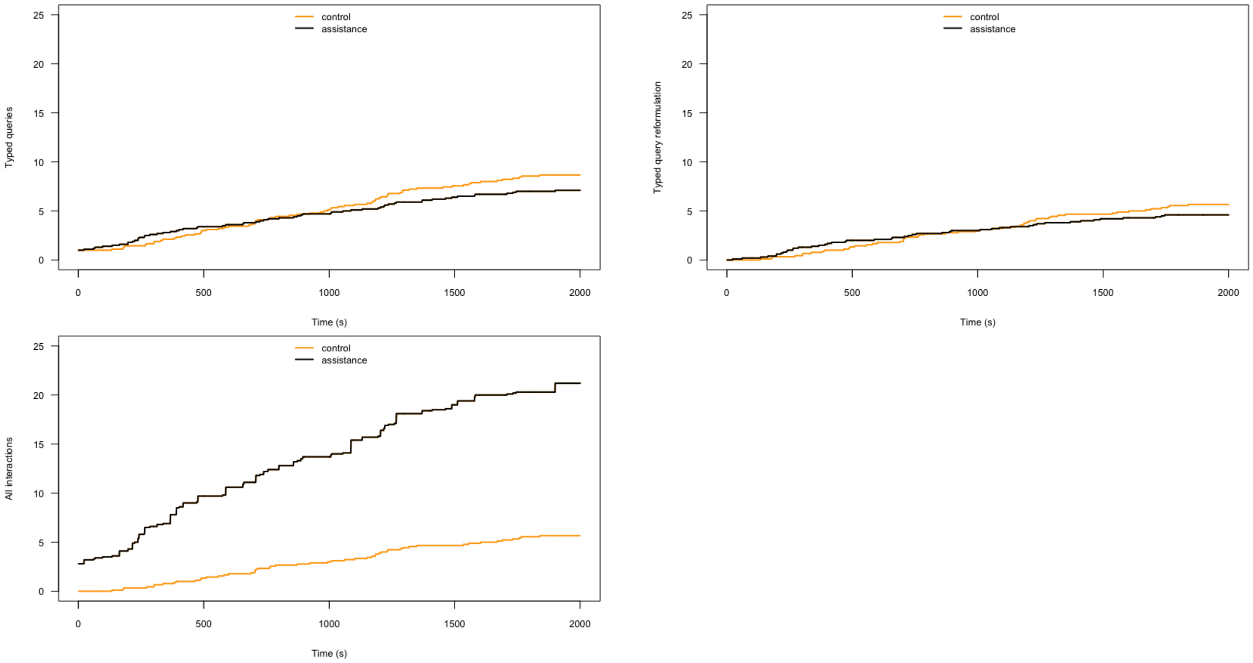


Figure 4**. Temporal plots of the interactions. The frequency is on the y-axis and elapsed task time on the x-axis. Typed-query frequency (top left), typed-query reformulation frequency (top right), and all interactions – typed query interactions in the control condition and typed query and interaction with the interactive faceted query suggestion (bottom).**

Discussion and Conclusions

We set out to study the effect of search assistance, operationalized as interactive faceted query suggestion, for users' momentary query-response level and whole-session level effectiveness in exploratory search tasks. We reported a user experiment demonstrating that users engage in interaction with interactive faceted query suggestion which lead to improvements in whole-session relevance over a control condition. We also show that the benefits are significant only at the whole-session level and do not manifest at query-response level search effectiveness.

Answers to the research questions

Is interactive faceted query suggestion associated with improved whole-session effectiveness? Yes, our results suggest that whole-session relevance in exploratory search is associated with the availability of interactive faceted query suggestion. Whole-session recall was significantly improved without compromising precision.

In detail, the cumulative recall for the system with interactive faceted query suggestion was over 80% higher when compared to the control condition. This suggests that the interactions with interactive faceted query suggestion component contributed substantially more at the whole-session level, despite no differences could be found at the level of individual interactions. In line with this finding, the results are also improved in terms of recall, but not in terms of precision.

Is interactive faceted query suggestion associated with improved query-response effectiveness? No, our results show that interactive faceted query suggestion provides limited or no advantage at individual query-response level; whenever participants interacted with the system, the conditions were equally effective in response to an individual query. This suggests that interactive faceted query suggestion is used to complement, but not substitute typed query interaction. As individual query-responses are equally effective, the result indicates that interactive faceted query suggestion is used for supporting exploration beyond the initial query context achieved via conventional typed-query interaction.

There are two plausible explanations for this result: anchoring behavior, and difficulty in expressing search intents with the typed query system. In the case of anchoring behavior the participants are submitting same or very similar queries in the control condition, resulting in overlap in the top-ranked results for a sequence of queries and lower cumulative recall. However, the typed queries are leading to equal precision, suggesting that interactive faceted query suggestion allows participants to specify information needs beyond the query scope allowing improved exploration and hence improved session-level recall.

Similarly, in the control condition the difficulty in expressing search intents may have caused the participants to struggle to come up with new queries, resulting in lower amount of interactions, and subsequently lower cumulative recall.

Does interactive faceted query suggestion engage users with interaction to direct search?

Yes, the participants issued significantly more interactions in the condition where interactive faceted query suggestions were available. Differences were not found in the amount of typed queries or the amount of reformulations. These findings suggest that interactive faceted query suggestions were used to complement the typed query interaction to conduct situated directed navigation around the initial query scope. Interactive faceted query suggestions seem to be particularly useful as exploration guidance; assimilating and expressing search intentions that were not obvious for the participants without the search assistance, but emerged in the interaction with the system and the user.

Implications

In summary, our results show increased whole-session recall without compromising precision when search assistance, operationalized as interactive faceted query suggestions, is available. No differences were observed in momentary query-response effectiveness indicating that individual interactions lead to equally good results, but the queries in the experimental condition are more diverse. Participants are also more engaged in searching when interactive faceted query suggestions are available, and issue over twice as many interactions with the system when compared to the control condition. These results are in line with recent work by Luo, Zhang, and Yang (2014), Raman et al. (2014), Capra et al. (2015), and Vakkari and Huuskonen (2012), suggesting that search effort may increase in exploratory search, but searchers obtain increased satisfaction and session-level effectiveness gains.

More generally, previous work has promoted the need to look at information search from a whole-session or whole-task point of view, but previous research has mostly focused on modeling or understanding search behavior from logs or developing evaluation approaches rather than studying the effects of system components enabling users to augment their interactions with search systems.

Our results show that performance gains can be attributed, not simply to create alternative, or reformulate and complement existing queries, but the ability of using search assistance for situated directed navigation. This has implications for both system design and evaluation methodologies for exploratory search systems. In the following, we summarize the implications derived from our results.

Search assistance to direct situated navigation

Our results show that whole-session output in exploratory search is associated with the availability of search assistance, but search assistance provides limited or no advantage at query-response level. Participants seem to rely on interactive faceted query suggestions to complement, but not substitute typed query interaction suggesting that it is mainly used for supporting situated navigation from the present typed-query context and that typed queries remain important for taking larger navigation steps. This implies that users gain new relevant information by using interactive faceted query suggestion and have access to a larger body of results, but the quality of system responses to individual interactions is unaffected: users do not on average issue better expressions of their information needs, but complement the typed queries by improved local navigation around the initial query scope.

Evaluation focus beyond the quality of response to an average query or rankings

The results also have implications for evaluation of exploratory search, suggesting a focus on measuring whole-session performance instead of performance of individual query-responses. The results suggest that whole-session output in exploratory search can substantially benefit from search assistance, but search assistance provides limited or no advantage at query-response level. This indicates that evaluating exploratory search performance via average query-response performance may provide only a limited view of the performance of the retrieval system. Evaluating whole-session effectiveness can be a proxy for user satisfaction at a session level and should be considered as an important measure for exploratory search performance.

Designing for engagement in situated directed navigation

Our results show that whole-session output in exploratory search is associated with the availability of search assistance, but the added value of the assistance comes only from active user engagements. Participants conducted more than double the number of interactions with the system where interactive faceted query suggestions were available than in the control condition. This implies that an important factor for the whole-session benefits to be realized is the design of user interfaces that enable and engage users in interaction with the recommendations provided by search assistance methods.

Limitations and Future work

There are some obvious limitations to our study. First, limit of our focus in exploratory search settings where users are willing to invest more time, required to explore, and are likely to benefit from search assistance. The task formulation may have led the participants to favor behavior that maximizes session outcomes. Consequently, our findings may have limited utility in shorter lookup search scenarios. While our experiments show significant effects in realistic task contexts with real user interactions, the effect sizes should be interpreted in such task context, and in comparison to the control condition.

Second, the levels of complexity and ambiguity of the tasks and topic domains were defined by the experimenters prior to the study, but have not been validated by external ratings or more objective measures. Thus, in future work a broader range of topic domains with clearly defined levels of complexity can be used to receive more detailed insights in the relationship between task complexity and the utility of search assistance.

Third, our experiments involved extensive search sessions, two conditions, and twenty participants. While this is a solid setup to study the effect of interactive faceted query suggestion, other interactive techniques, such as alternative interfaces and methods for query suggestions, query auto-completion, or similar techniques could be also used as control conditions in future experiments, and larger pools of participants could be used to validate the effects in real-world usage situations and tasks.

Finally, by predefining the topic domains the study did not address personal information needs naturally rising in real use, which could be required to observe fully realistic search behavior. To increase ecological validity, future experiments might involve setups in which the system would be exposed to real-life use enabling testing the methods in a more naturalistic setting.

References

Ahlberg, C., & Shneiderman, B. (1994). Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 313–317). New York, NY, USA: ACM.

Auer, P. (2003, March). Using confidence bounds for exploitation- exploration trade-offs. *J. Mach. Learn. Res.*, *3*, 397–422.

Baeza-yates, R., Hurtado, C., & Mendoza, M. (2004). Query recommendation using query logs in search engines. In *Proc. clustweb* (pp. 588–596). Springer.

Bar-Yossef, Z., & Kraus, N. (2011). Context-sensitive query auto-completion. In *Proceedings of the 20th international conference on world wide web* (pp. 107–116). New York, NY, USA: ACM.

Bing, L., Lam, W., Wong, T.-L., & Jameel, S. (2015, February). Web query reformulation via joint modeling of latent topic dependency and term context. *ACM Trans. Inf. Syst.*, *33*(2), 6:1–6:38.

Boldi, P., Bonchi, F., Castillo, C., Donato, D., Gionis, A., & Vigna, S. (2008). The query-flow graph: Model and applications. In *Proc. cikm'08* (pp. 609–618). New York, NY, USA: ACM.

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

Cai, F., & de Rijke, M. (2016). A survey of query auto completion in information retrieval. *Foundations and Trends in Information Retrieval*, *10*(4), 273-363.

Cao, H., D., Pei, J., He, Q., Liao, Z., Chen, E., & Li, H. (2008). Context-aware query suggestion by mining click-through and session data. In *Proc. kdd'08* (pp. 875–883). New York, NY, USA: ACM.

Cao, H., Hu, D. H., Shen, D., Jiang, D., Sun, J.-T., Chen, E., & Yang, Q. (2009). Context-aware query classification. In *Proc. sigir'09* (pp. 3–10). New York, NY, USA: ACM.

Capra, R., Arguello, J., Crescenzi, A., & Vardell, E. (2015). Differences in the use of search assistance for tasks of varying complexity. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval* (pp. 23–32). New York, NY, USA: ACM.

Carterette, B., Clough, P., Hall, M., Kanoulas, E., & Sanderson, M. (2016). Evaluating retrieval over sessions: The trec session track 2011-2014. In *Proceedings of the 39th international acm sigir conference on research and development in information retrieval* (pp. 685–688). New York, NY, USA: ACM.

Cheng, Z., Gao, B., & Liu, T.-Y. (2010). Actively predicting diverse search intent from user browsing behaviors. In *Proc. www'10* (pp. 221–230). New York, NY, USA: ACM.

Cheung, J. C. K., & Li, X. (2012). Sequence clustering and labeling for unsupervised query intent discovery. In *Proc. wsdm '12* (pp. 383–392). New York, NY, USA: ACM.

Duggan, G. B., & Payne, S. J. (2008). Knowledge in the head and on the web: Using topic expertise to aid search. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 39–48). New York, NY, USA: ACM.

Feild, H., & Allan, J. (2013). Task-aware query recommendation. In *Proc. sigir'13* (pp. 83–92).

Guan, D., Zhang, S., & Yang, H. (2013). Utilizing query change for session search. In *Proceedings of the 36th international acm sigir conference on research and development in information retrieval* (pp. 453–462). New York, NY, USA: ACM.

Hassan, A., White, R. W., Dumais, S. T., & Wang, Y.-M. (2014). Struggling or exploring?: Disambiguating long search sessions. In *Proceedings of the 7th acm international conference on web search and data mining* (pp. 53–62). New York, NY, USA: ACM.

He, J., Bron, M., & de Vries, A. P. (2013). Characterizing stages of a multi-session complex search task through direct and indirect query modifications. In *Proceedings of the 36th international acm sigir conference on research and development in information retrieval* (pp. 897–900). New York, NY, USA: ACM.

Hearst, M. A. (1995). Tilebars: visualization of term distribution information in full text information access. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 59–66). New York, NY, USA: ACM Press/Addison-Wesley Publishing Co.

Hu, B., Zhang, Y., Chen, W., Wang, G., & Yang, Q. (2011). Characterizing search intent diversity into click models. In *Proc. www '11* (pp. 17–26). New York, NY, USA: ACM.

Huang, J., & Efthimiadis, E. N. (2009). Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th acm conference on information and knowledge management* (pp. 77–86). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/1645953.1645966` doi: 10.1145/1645953.1645966

Järvelin, K., Price, S. L., Delcambre, L. M. L., & Nielsen, M. L. (2008). Discounted cumulated gain based evaluation of multiple-query ir sessions. In *Proc. ecir'08* (pp. 4–15). Berlin, Heidelberg: Springer-Verlag. Retrieved from `http://dl.acm.org/citation.cfm?id=1793274.1793280`

Jiang, J., He, D., & Allan, J. (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 &#38; development in information retrieval* (pp. 607–616). New York, NY, USA: ACM.

Jiang, J.-Y., Ke, Y.-Y., Chien, P.-Y., & Cheng, P.-J. (2014). Learning user reformulation behavior for query auto-completion. In *Proc. sigir '14* (pp. 445–454). New York, NY, USA: ACM.

Jones, R., & Klinkner, K. L. (2008). Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In *Proc. cikm '08* (pp. 699–708). New York, NY, USA: ACM.

Kangassalo, L., Spapé, M., Jacucci, G., & Ruotsalo, T. (2019). Why do users issue good queries?: Neural correlates of term specificity. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval* (pp. 375–384). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/3331184.3331243` doi: 10.1145/3331184.3331243

Kanoulas, E., Carterettey, B., Hallz, M., Cloughx, P., & Sanderson, M. (2012). Overview of the trec 2011 session track. In *Trec'11.*

Kekäläinen, J. (2005, September). Binary and graded relevance in ir evaluations-comparison of the effects on ranking of ir systems. *Inf. Process. Manage.*, *41*(5), 1019–1033. Retrieved from `https://doi.org/10.1016/j.ipm.2005.01.004` doi: 10.1016/j.ipm.2005.01.004

Kharitonov, E., Macdonald, C., Serdyukov, P., & Ounis, I. (2013). Intent models for contextualising and diversifying query suggestions. In *Proc. cikm '13* (pp. 2303–2308). New York, NY, USA: ACM.

Klouche, K., Ruotsalo, T., Cabral, D., Andolina, S., Bellucci, A., & Jacucci, G. (2015). Designing for exploratory search on touch devices. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (pp. 4189–4198). New York, NY, USA: ACM.

Koren, J., Zhang, Y., & Liu, X. (2008). Personalized interactive faceted search. In *Proceedings of the 17th international conference on world wide web* (pp. 477–486). New York, NY, USA: ACM.

Kules, W., Wilson, M. L., Schraefel, M. C., & Shneiderman, B. (2008). *From keyword search to exploration: How result visualization aids discovery on the web* (Tech. Rep.). University of Southampton.

Li, Y., Dong, A., Wang, H., Deng, H., Chang, Y., & Zhai, C. (2014). A two-dimensional click model for query auto-completion. In *Proc. sigir'14* (pp. 455–464). New York, NY, USA: ACM.

Liu, J., & Belkin, N. J. (2015). Personalizing information retrieval for multi-session tasks:

Examining the roles of task stage, task type, and topic knowledge on the interpretation of dwell time as an indicator of document usefulness. *Journal of the Association for Information Science and Technology*, *66*(1), 58-81.

Liu, J., Cole, M. J., Liu, C., Bierig, R., Gwizdka, J., Belkin, N. J., ... Zhang, X. (2010). Search behaviors in different task types. In *Proceedings of the 10th annual joint conference on digital libraries* (pp. 69–78). New York, NY, USA: ACM.

Liu, J., Liu, C., & Belkin, N. (2013). Examining the effects of task topic familiarity on searchers' behaviors in different task types. *Proceedings of the American Society for Information Science and Technology*, *50*(1), 1-10.

Liu, J., Liu, C., Cole, M., Belkin, N. J., & Zhang, X. (2012). Exploring and predicting search task difficulty. In *Proc. cikm '12* (pp. 1313–1322). New York, NY, USA: ACM.

Luo, J., Zhang, S., & Yang, H. (2014). Win-win search: Dual-agent stochastic game in session search. In *Proc. sigir '14* (pp. 587–596). New York, NY, USA: ACM.

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

Matejka, J., Grossman, T., & Fitzmaurice, G. (2012). Citeology: Visualizing paper genealogy. In *Chi '12 extended abstracts on human factors in computing systems* (pp. 181–190). New York, NY, USA: ACM.

Niu, X., Fan, X., & Zhang, T. (2019, January). Understanding faceted search from data science and human factor perspectives. *ACM Trans. Inf. Syst.*, *37*(2), 14:1–14:27. Retrieved from `http://doi.acm.org/10.1145/3284101` doi: 10.1145/3284101

Qvarfordt, P., Golovchinsky, G., Dunnigan, T., & Agapie, E. (2013). Looking ahead: Query preview in exploratory search. In *Proc. sigir '13* (pp. 243–252). New York, NY, USA: ACM.

Raman, K., Bennett, P. N., & Collins-Thompson, K. (2014, October). Understanding intrinsic diversity in web search: Improving whole-session relevance. *ACM Trans. Inf.*

Syst., *32*(4), 20:1–20:45.

Ruotsalo, T., Jacucci, G., Myllymäki, P., & Kaski, S. (2014, December). Interactive intent modeling: Information discovery beyond search. *Commun. ACM*, *58*(1), 86–92.

Ruotsalo, T., Peltonen, J., Eugster, M., Głowacka, D., Konyushkova, K., Athukorala, K., … Kaski, S. (2013). Directing exploratory search with interactive intent modeling. In *Proc. cikm '13* (pp. 1759–1764). New York, NY, USA: ACM.

Ruotsalo, T., Peltonen, J., Eugster, M. J. A., Głowacka, D., Floréen, P., Myllymäki, P., … Kaski, S. (2018, October). Interactive intent modeling for exploratory search. *ACM Trans. Inf. Syst.*, *36*(4), 44:1–44:46.

Sadikov, E., Madhavan, J., Wang, L., & Halevy, A. (2010). Clustering query refinements by user intent. In *Proc. www '10* (pp. 841–850). New York, NY, USA: ACM.

Savenkov, D., & Agichtein, E. (2014). To hint or not: Exploring the effectiveness of search hints for complex informational tasks. In *Proceedings of the 37th international acm sigir conference on research & development in information retrieval* (pp. 1115–1118). New York, NY, USA: ACM.

Shokouhi, M. (2013). Learning to personalize query auto-completion. In *Proc. sigir '13* (pp. 103–112). New York, NY, USA: ACM.

Stasko, J., Görg, C., & Liu, Z. (2008). Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization*, *7*(2), 118–132.

Terveen, L., Hill, W., & Amento, B. (1999). Constructing, organizing, and visualizing collections of topically related web resources. *ACM Transactions on Computer-Human Interaction*, *6*(1), 67–94.

Vakkari, P. (2001). A theory of the task-based information retrieval process: a summary and generalization of a longitudinal study. *J. Docum.*, *57*, 44–60.

Vakkari, P., & Huuskonen, S. (2012). Search effort degrades search output but improves task outcome. *JASIST*, *63*(4), 657–670.

Vuong, T., Saastamoinen, M., Jacucci, G., & Ruotsalo, T. (n.d.). Understanding user behavior in naturalistic information search tasks. *Journal of the Association for Information Science and Technology*. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24201` doi: 10.1002/asi.24201

White, R. W., & Huang, J. (2010). Assessing the scenic route: Measuring the value of search trails in web logs. In *Proc. sigir '10* (pp. 587–594). New York, NY, USA: ACM.

Yee, K.-P., Swearingen, K., Li, K., & Hearst, M. (2003). Faceted metadata for image search and browsing. In *Proc. chi '03* (pp. 401–408). New York, NY, USA: ACM.

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, *22*(2), 179–214.