Supporting Exploratory Search Tasks with Interactive User Modeling

Tuukka Ruotsalo[1], Kumaripaba Athukorala[2], Dorota Głowacka[2], Ksenia Konyushkova[2], Antti Oulasvirta[3], Samuli Kaipiainen[2], Samuel Kaski[1,2], and Giulio Jacucci[2]

[1]Helsinki Institute for Information Technology HIIT, Aalto University
PO Box 15600, 00076 Aalto, Finland
[2]Helsinki Institute for Information Technology HIIT, University of Helsinki
Department of Computer Science, PL 68, 00014 Helsinki, Finland
[3]Max Planck Institute for Informatics
Campus E1 4, 66123 Saarbrücken, Germany

ABSTRACT

This paper presents the design and study of *interactive user modeling* to support exploratory search tasks. Contrary to traditional interactions, such as query based search, query suggestions, or relevance feedback, interactive user modeling allows users to perceive the state of a user model at all times and provide feedback that directly rewards or penalizes it. The technique allows users to continuously tune the system's belief about their evolving information needs. We demonstrate that such functionality is useful in exploratory search where users need to get accustomed to a body of literature in a domain. We conducted two experiments where scientists carried out exploratory search tasks with our implementation of an interactive user modeling and retrieval system (SciNet) and two baselines: SciNet from which interactive user modeling was excluded and a real-world baseline (Google Scholar). The results show that interactive user modeling can help users to more effectively find relevant, novel and diverse results without compromises in task execution time.

Keywords

Search user interfaces, information-seeking behavior, exploratory search, user modeling.

INTRODUCTION

The performance of an information retrieval system is affected not only by its ability to produce documents relevant to a given query, but also by the users' ability to interact with the information space presented by its user interface (Marchionini, 2006). This paper contributes by presenting a novel approach that allows users to interactively control a

ASIST 2013, November 1-6, 2013, Centre Sheraton, Montreal, Quebec, Canada.

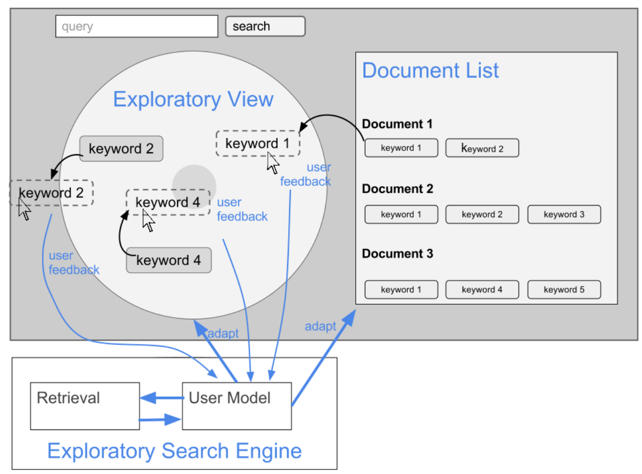

Figure 1: *SciNet* is a prototype system to study the concept of interactive user modeling. Keywords that are used as the user model's features are visualized on the exploratory view (left). Documents retrieved based on the model are shown in the document list (right). The user can move keywords (drag-and-drop within the exploratory view or from underneath the documents in the document list) to provide feedback to the user model. Proximity of a keyword to the center affects relevance. Every edit updates the interface.

user model that represents the user's information need transparently as they explore a complex information space. The approach is designed with the *exploratory search task* in mind in which the information needs are not discretely anticipated, but rather emerge as the users iteratively seek, learn, and reflect on complex information (Chowdhury et. al., 2011;Byström & Järvelin, 1995). Our *interactive user modeling* approach utilizes reinforcement learning and interactive visualization to enable *a model-based feedback loop*: the system actively learns from user interactions and proposes improved user model that the user can iteratively adjust. Two features are necessary:

1) *Transparent visualization and feedback of the user model* by allowing users to provide relevance feedback directly on the user model features. In our case these features are *key-*

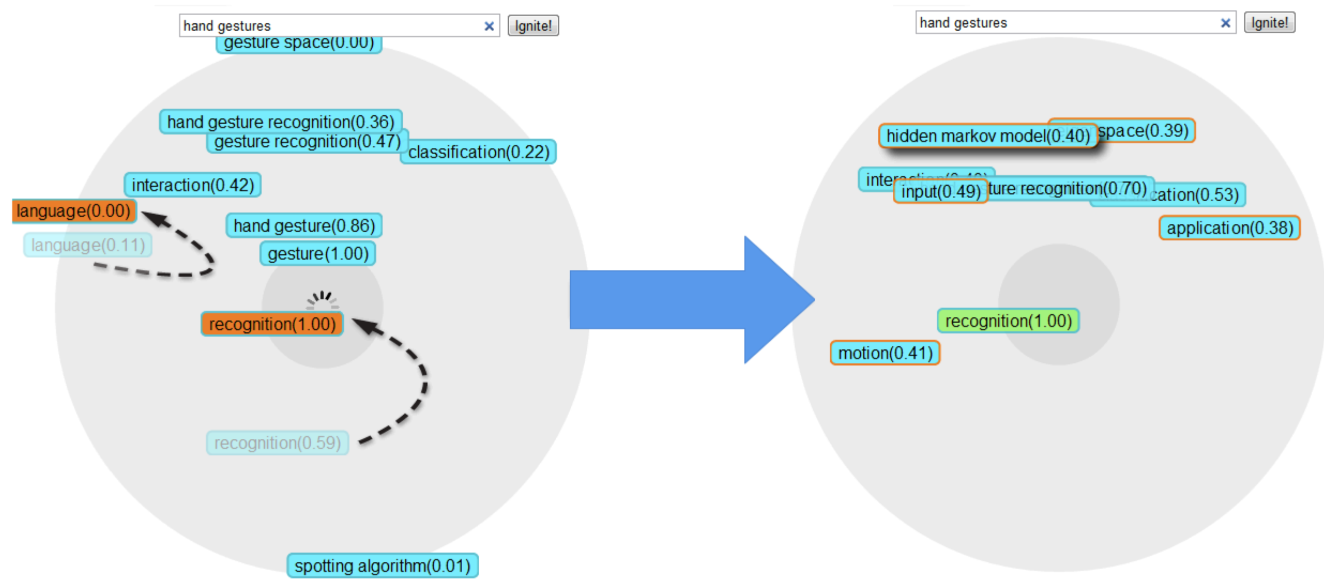

Figure 2: Interactions with the *Exploratory View*. In the first iteration (left) the user indicates an increased importance of the keyword "recognition" by dragging it towards the center of the exploratory view and indicates a reduced importance of the keyword "language" by dragging it outside the exploratory view. *SciNet* colors the keywords explicitly manipulated by the user to distinguish them. In the second iteration (right) new keywords have been predicted and are positioned on the exploratory view according to their estimated relevance.

words present and extracted form the documents and visualized for the user in the *exploratory view* (Figure 1).

2) Simultaneous user modeling of relevance and uncertainty by employing *exploration / exploitation* tradeoff of reinforcement learning. The modeling simultaneously employs exploitation (maximizing the relevance for the user) and exploration (minimizing the uncertainty of the system) based on proposing features to which the interactions can be targeted.

Figure 1 illustrates interactions in *SciNet,* a prototype system that implements the interactive user modeling, and Figure 2 shows the interactions with its key interface component, *exploratory view,* in more detail. The *SciNet* system is built to study these ideas in the domain of scientific information seeking. Previous work on query-based scientific search databases has found out that successful task performance is predicted by multiple query iterations, narrowing/broadening strategies, long evaluation times, and sound query formulation (Sutcliffe et. al., 2000). To critically assess the potential impact of our approach on information seeking behavior, we study scientists' performance in *exploratory search tasks* – one of the most complex domains of information seeking. We hypothesize that interactive user modeling can change the search process by allowing users to control the search process at a higher level through keywords, thus more effectively exploiting/exploring the space. The studies reported here address two research questions:

RQ1: Adoption by Users: *Will users make use of interactive user modeling when exploring information or will they rather resort to query-based search?*

RQ2: Task Success: *If users do adopt the features, will it improve their task success, i.e increase the quality or amount of relevant information they can acquire?*

We conducted two studies where *SciNet* was compared in a realistic task-based information seeking setting (Ingwersen & Järvelin, 2005) against two alternatives: First, a within-system baseline in which all other features were equal, but the interactive user modeling techniques were excluded in *SciNet*; Second, *Google Scholar*, a widely used real-world system representative of the state-of-the-art in query-based interaction. In addition to query-based search, *Scholar* employs many additional interaction techniques such as query suggestions, and text snippets that users can use as a source for cues to reformulate queries. At the time of the experiments, *SciNet* indexed a comparable proportion of scientific literature (over 60,000,000 documents). Performance was measured from two perspectives. First (RQ1), interaction with the system was measured by analyzing interaction logs and subjective assessments of the usability of the systems. Second (RQ2), information seeking efficiency was measured based on users' ability to find task-relevant documents and categorization using a give system (task performance), and system retrieval performance was measured by the relevance of the information returned by the compared systems in response to user interactions (system performance).

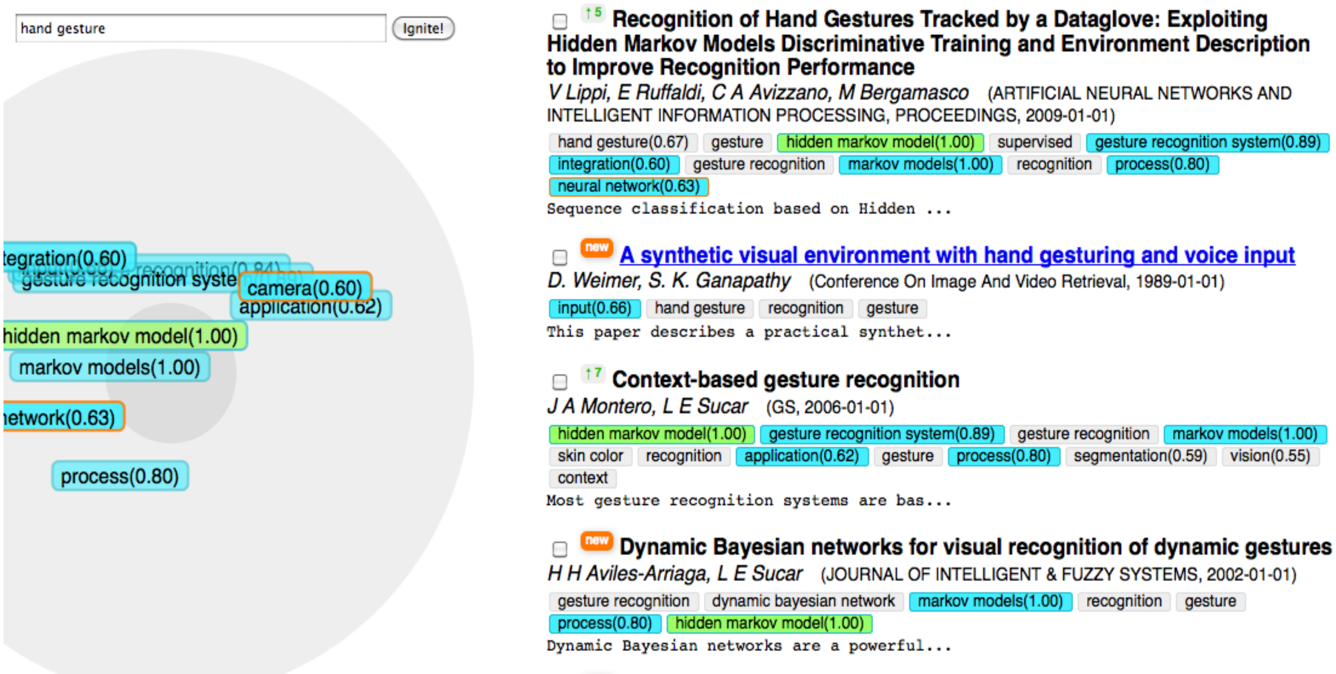

Figure 3: The document list after Iteration 2 has both new documents (labeled "new") and documents whose rank increased from the previous round. The user has now obtained documents matching the information need. The exploratory view (left) also offers options for continuing the exploration in other potentially relevant directions, such as the use of cameras, or neural networks in hand gesture recognition, or applications of hand gesture recognition.

The results show that interactive user modeling can significantly improve users' task performance by allowing more effective system performance without sacrificing task completion time. In particular, the interactive user modeling allows users to find more relevant, novel and diverse information without compromises in task completion time.

BACKGROUND

Exploratory search is a non-static information retrieval setting in which computational support is focused to assist users to interact, control, learn, and discover information during search process. It emphasizes on iterative dialogue between the system and the user through adaptive interfaces. A characterizing fact of exploratory search is the understanding of the search process as an investigatory process rather than a simple lookup function (Marchionini, 2006; Fox et. al., 2006). The target space and the nature of the problem of exploratory search is uncertain (White et. al., 2006), so every exploratory search system has an interactive user interface as the core component in order to implement the iterative exploration (Ahn & Brusilovsky, 2013; Glowacka et. al., 2013; Ruotsalo et. al., 2013). Often the interface can only visualize parts of the search space, simply because the whole potentially relevant space is too large. Personalization, filtering (e.g. faceted search), result categorization or clustering, and relevance feedback methods are often employed to limit and predict the relevant parts of the search space to help users to point their feedback to the currently relevant features. We briefly review these approaches and discuss their benefits and shortcomings, and contrast them to our approach.

Personalized information retrieval often focuses on adapting document rankings based on users' query logs or other interaction histories (Liu, 2009; Pitokow et. al., 2002; White et. al., 2010). However, personalization in most cases refers to techniques that use implicit feedback, i.e. feedback that is not explicitly acquired from the user, but observed based on links the user clicks or queries the user types. While personalization is based on user modeling, it is not interactive and does focus in providing explicit feedback mechanisms and engaging users to use them as a part of the search process.

Faceted search is an information filtering approach (Yee et. al., 2003) wherein users can navigate along conceptual dimensions that describe the content. It allows explicit feedback directly on topical categories. The problem is to keep the number of options, or facet categories, low enough for them to be interpretable for the user. Therefore, the facet categories must be based on either exploiting of what lies in the result set initially returned by a search engine or a global dataset independent of the query context. This may result in limited or overly general navigation options and facet categories that do not meet the user information needs that emerge within the seeking sessions. As a result of these limitations, the facets essentially function as filtering criteria, and users are forced to rely on typing queries whenever their expression of their information need is close, but not achievable with the current set of facet categories (Yee et. al., 2003).

Result categorization or clustering (Carpinento et. al., 2009; Cutting et. al., 1992; Hearst et. al., 1996 & 2006) is based on the idea of clustering the search results or the whole document space and visualizing the cluster set for the user to aid navigation in the information space. Search result clustering builds up on the idea that clustered groups of search results give users both overview and focus-view (Käki, 2005). After scanning the overall scope of the results, the user can focus on a specific cluster and further explore related documents. Clustering suffers from the same shortcoming as faceted search: in search result clustering the user is limited to exploring only within the initial query scope, and on the other hand if the clustering covers the entire document collection, the user risks losing the query context completely.

Mixed initiative interaction (Horvitz, 1999) is an alternative way to predict user intentions. It refers to a flexible interaction strategy, wherein an agent can contribute resolving the user's task in an interactive manner by initiating a dialogue for the user when it infers that the user may need assistance in navigation or problem solving. Mixed-initiative interaction is the principle most closely matching to our approach in the sense that both the user and the system are allowed and expected to be active but it has been traditionally developed as a pat of an agent system to assist users in an office tools environment, where its success has been limited by to the effort required from the user to correct prohibitive inferences and dialogue propositions mistakenly initiated by the system.

Our *interactive user modeling* approach is different from all the mentioned techniques in two ways. First, interactive user modeling allows exploration in addition to pure exploitation. The problem of pure exploitation employed by the existing techniques is that it produces search results and navigation options that are trapped inside user's initial query, and hence the offered interaction options allow user to access only very narrowly defined content. This forces users to repeat typed ad-hoc queries to explore beyond the initial query scope. Second, interactive user modeling exposes the relevant features of the user model (in contrast to only filtering criteria) for direct manipulation through visualization. Users can provide feedback and reduce system's uncertainty about user's needs in dialogue between the system and the user. In summary, our approach can avoid typical flaws of filtering systems that lead users to get stuck in suboptimal local contexts, or filter bubbles, due to suboptimal user interactions, and at the same time help the user to understand the navigation options most relevant to the current estimate of the user model.

INTERFACE AND INTERACTION DESIGN

We illustrate the interface and interaction design of the *SciNet* system through a walkthrough that exemplifies a real information-seeking task. *SciNet* enables interactive user modeling through a user interface composed of two main elements: the *exploratory view* and the *document list*, as shown in Figure 1. It additionally has a query typing area, as is traditional in query-based search interfaces. The *exploratory view* visualizes the user model on a radial layout and allows users to provide relevance feedback by moving keywords on this layout.

In our scenario, a user who is writing an essay about the topic *hand gestures* begins the seeking process by typing "*hand gestures*" as the query. The system then retrieves a set of documents and adapts the content to match the user s feedback (Figure 2).

On the first iteration the documents and keywords are chosen based on a direct match to the user's query and visualized for the user in the *document list* and the *exploratory view*. At this point the user's interest on hand gesture recognition increases (Iteration 1 in Figure 2) and she realizes that the keyword *language* is not related to her information need. She provides feedback for the system by moving it outside of the *exploratory view* and by moving the keyword "*recognition*" to the center of the *exploratory view*. The user then submits the feedback by clicking the center of the *exploratory view*. The system learns a more specific representation of the user's information needs from the feedback, expresses it in terms of keywords, and retrieves and predicts a new set of documents and keywords (Iteration 2 in Figure 2). At the end of Iteration 2 the user decides to take a look at documents about *hidden Markov model* (Figure 3). The *document list* now consists of the documents related to hand gesture recognition with hidden Markov models, but because of the novelty and diversity featured in the adaptation methods, the *exploratory view* allows alternative options for the user to select: applications, the user of neural networks, and the use of cameras in hand gesture recognition. By moving these towards the center the user could continue the seeking process either by drilling down in to alternative techniques, such as neural networks, or by building up a general overview of the information space by investigating the applications of hand gesture recognition.

INTERACTIVE USER MODELING

The intuition behind the interactive user modeling approach is that the system obtains user's feedback directly on the features of the current estimate of the user model through interactive visualization and uses exploration/exploitation paradigm of reinforcement learning to learn new estimates of the user model as interaction occurs. As opposite to conventional user modeling techniques that try to maximize relevance based on the available feedback, our approach allows continuous exploration by allowing the user and the computing system to control the user model transparently by both maximizing relevance but also reducing system's uncertainty about the user's information needs.

The user model is in our case represented as a weighted set of keywords. The model predicts relevance for potential future intents of the user based on this feedback, but at the same time selects keywords for the visualization not only by relevance (exploitation), but also based on how uncertain the system is about the most relevant keywords (exploration). This allows users to continuously improve the estimate of her intents by reinforcing their information needs.

The user can provide feedback by moving a keyword closer to or further from the center of the exploratory view: keywords in the center have relevance score 1 with the value getting smaller the further away from the center a keyword is moved (see Figure 2). Keywords placed on the edge of the exploratory view or beyond have relevance score 0. Keywords with relevance score 0 are excluded from appearing again in the exploratory view for the remainder of a given search session. Thus, the feedback is given by a relevance score $r \in [0, 1]$ for a number of keywords $1 \ldots i$.

The model to compute the estimates of other keywords that have not received direct feedback is as follows. We assume that the relevance score r_i of a keyword k_i is a random variable with expected value $r_i = k_i \cdot w$, such that the expected relevance score is a linear function of the keywords. The unknown weight vector w is essentially the representation of the user's information need and determines the relevance of keywords used in retrieval and visualized for the user on each iteration to allow feedback directly on the new estimates.

In order to solve the linear function and estimate value for each keyword, we use *LinRel* (Auer, 2002), an algorithm that has already been proven to work well in controlling exploration/exploitation tradeoff in interactive settings. The algorithm maintains a representation of the estimate w of the unknown weight vector. When selecting the next set of keywords to display, the system might simply select the keywords with the highest estimated relevance score by solving the linear regression problem. But since the estimate w may be inaccurate, this exploitative choice might be suboptimal. In other words, the feedback acquired from the user may result in an estimate that is suboptimal and does not represent the information need of the user nor the potential feedback options users would like to use to improve the estimate. Alternatively, the system can exploratively select a keyword for which the user feedback improves the accuracy of the estimate w, enabling better keyword selections in subsequent iterations. This is achieved by reducing uncertainty by requesting feedback for keywords that have the largest upper confidence bound when maximizing both relevance and uncertainty.

In each iteration, *LinRel* obtains an estimate w by solving a linear regression problem. Suppose we have a matrix K, where each row k_i is a feature vector of keywords presented so far. Let $r = [r_1, r_2 \ldots r_p]$ be the column vector of relevance scores received so far from the user as feedback, where p is a number of iterations. Thus, *LinRel* tries to estimate w by solving $r = K \cdot w$. Based on w, *LinRel* calculates an estimated relevance score $r = k \cdot w$ for each keyword k_i. As noted, instead of selecting the highest estimates based on the relevance scores, in order to deal with the exploration-exploitation trade-off, we select keywords not with the highest relevance score, but with the largest upper confidence bound for the relevance score.

Thus, if σ_i is an upper bound on standard deviation of relevance estimate r_i, the upper confidence bound of keyword k_i is calculated as $r_i + \gamma\sigma_i$, where $\gamma > 0$ is a constant used to adjust the confidence level of the upper confidence bound (i.e. the amount of exploration). In each iteration, LinRel calculates $s_i = \mathrm{K}(K^\top \mathrm{K} + \lambda \mathrm{I})^{-1} k_i$, where λ is a regularization parameter. The keywords that maximize $s_i{}^\top \mathrm{r} + \frac{\gamma}{2} \| s_i \|$ are selected for presentation and used in retrieval.

The selected keywords are then visualized for the user and their weights are used as an input for the ranking formula. We use a language modeling approach with Bayesian Dirichlet smoothing (Zhai & Lafferty, 2001) to retrieve a new set of documents and attached keywords by weighting each keyword k_i with their associated estimate of weight r_i. The results are then diversified via Dirichlet Sampling to ensure maximal coverage of different interests represented in the user model.

As a result of this procedure the system can estimate a weight for each keyword, visualize the keywords for the user to obtain feedback to both improve relevance estimates and reduce uncertainty related to each estimate, and retrieve documents for the user that match to the present estimate. This interactive user modeling both helps the user to explore the information space and allows the system to reduce the uncertainty related to potentially relevant keywords.

Task and domain	Task definition
Cognitive control *Psychology*	Which human functions (cognitive abilities) are related to this topic? Mention at least three. List at least three brain areas, two neurotransmitters and two mental disorders that have been shown to relate to cognitive control. Select 10 articles you find useful.
Reinforcement learning *Machine learning*	Which research areas make use of Reinforcement Learning. Mention at least five. Select 10 articles that you find useful.
Semantic search *Information retrieval*	Which kind of techniques and methods are used to acquire and utilize semantics in a search process. Mention at least five. Find research areas related to Semantic Search. Select 10 articles that you find useful..
Communication protocols *Computer networking*	Find research areas related to communication protocols. Mention at least five. List the protocols you have found. Mention at least 10. Select 10 articles that you find useful.
Fair use *Trademark law*	Which users are protected under "Fair use"? Find the economic uses for and against fair use. Select 10 articles that you find useful.

Table 1. Tasks and task definitions

EXPERIMENTS

We evaluate the interactive user modeling approach by comparing the *SciNet* system to two different baselines in two different studies. The first study measured retrieval performance, that is, the quality of results returned by the system in response to user interactions. The compared settings were the *SciNet* system with interactive user modeling and a within-system baseline setting in which users were only able to type queries (i.e. they did not benefit from user modeling or interaction features). The second study measured users' task performance, that is, the quality of information selected by users and compared *SciNet* with interactive user modeling support against a real-world baseline Google Scholar. In both studies the users were situated in an exploratory information-seeking scenario with a task-based setting (Ingwersen & Järvelin, 2005). That is, they were provided with a scenario describing information needs and asked to use the system to acquire information addressing these needs by using a given system. This experimental design allowed us to quantify performance, and to do so in a way that captures the essence of exploratory seeking behavior.

Tasks and Materials

We recruited five post-doctoral researchers as experts to both define tasks and evaluate the outcome in terms of the results provided by the participants, and the information returned by the system in response to user interactions. The experts were from five different research areas and each of them constructed a task in their area of expertise. The tasks were defined in accordance with a task template designed to situate the participants in a scientific writing scenario. The participants were asked to 1) find a representative set of scientific articles covering the topic and 2) find a more specific categorization or specific subfields under the topic. The tasks and their definitions are shown in Table 1. To minimize the effect of using different tasks, task structure, complexity and prior knowledge requirements of each task were normalized (Leide el al., 2007). The difficulty of the tasks was adjusted to be equal using NASA's Task Load Index (TLX) (Hart et. al., 1988) via trial studies.

We conducted all of the studies in a controlled setting, where participants used the given system with a computer connected to an 18"-21" LCD monitor and interacted with the device using mouse and keyboard.

Datasets and System Setups

At the time of the experiments, *SciNet* had indexed over 60 million documents from the following data sources: the Web of Science of Thomson Reuters, 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. The following fields of the original data were indexed: title, authors, publication forum, date of publication, abstract and keywords associated with each of the articles. The dataset and ranking formulas were the same for both system setups in the retrieval performance experiment and only interactive user modeling was excluded in the baseline condition. Google Scholar was used without modifications.

EXPERIMENT 1: Retrieval Performance

The purpose of the first experiment was to measure the effect of interactive user modeling for the retrieval performance of the system. That is, how accurate results the system is able to return in response to user interactions when the users were solving the task.

Experimental Design

The experiment used between-subjects design, i.e. each participant performed only a single task with one of the two system setups. The independent variables were the two system conditions: full system with interactive user modeling support and a system with only typed-query interaction. The retrieval effectiveness was the dependent variable. To ensure that we could collect enough data to reliably study retrieval effectiveness and gain enough data per task, for this experiment we used only two of the tasks defined by the experts: semantic search and reinforcement learning.

Participants and Procedure

We recruited 20 researchers from our university to participate in this study. All the participants were either faculty or students. To ensure that participants prior knowledge will not influence the exploratory nature of the tasks., we conducted a background survey of the participants.

The survey ensured that the participants had conducted literature search before but were not expert researchers in the topics of the search tasks The participants were explained each task and a short training session was given before they performed the task. The time to complete the task was set to 30 minutes.

	Relevance		Novelty		Obviousness	
	SciNet	B	SciNet	B	SciNet	B
F	0.25	0.15	0.18	0.09	0.22	0.20
P	0.69	0.72	0.40	0.33	0.26	0.34
R	0.15	0.09	0.12	0.05	0.20	0.17
#	882	552	570	228	253	223

Table 2: Precision, Recall, F-measure, and number of documents found in the Interactive user modeling condition (SciNet) and in the baseline condition (B). The differences between the systems in terms of relevance and novelty were found to be statistically significant (Wilcoxon Signed-Rank test, df=2, p<0.01).

Measurement

We measured effectiveness of the compared systems in terms of precision, recall, and F1 measure of the articles returned by the two systems in response to users' interactions with the system. We created a ground truth by pooling all articles found by any user with either of the system setups. This resulted in a pool of over 5283 articles that were all assessed by experts with respect to three properties: 1) relevance (relevant or not relevant article) 2) novelty (relevant article that is related to a specific aspect of the overall topic) and 3) obviousness (relevant, but obvious article that is well known in the field) (Clarke et. al., 2008). Overlapping assessments were conducted by two experts. Cohen Kappa test was then run to measure inter-annotator agreement between the experts. Kappa indicated a substantial agreement (Kappa = 0.71, p < 0.001).

Results

Table 2 shows the general system performance results. In all the assessed categories, i.e. relevance, novelty and obviousness, the *SciNet* system with interactive user modeling outperforms the baseline in terms of F-measure, i.e. the harmonic mean of the precision and recall achieved on average by a user throughout a search session with a gain from 0.15 to 0.25 in general relevance and from 0.09 to 0.18 in the case of novel documents. The higher F-measure is explained by an increase of recall, while precision in interactive user modeling condition is greater only for novelty. Still the condition with interactive user modeling achieves similar precision in other categories, 0.69 / 0.72 in precision for general relevance and 0.26 / 0.34 for obvious category. This indicates that interactive user modeling can improve users ability to acquire more relevant and novel documents while being able to acquire equally well obvious documents in the same seeking time.

Figure 4 presents cumulative F-measures of the two compared system conditions over time on the three relevance categories: relevant, obvious, and novel, i.e. illustrates the achieved gain as a function of time as the search progresses.

Interactive user modeling is beneficial for users as the improvement of the results achieved by the users in terms of F-measure is present throughout the search session. The underlying reason is that for the system that benefits from the interactive user modeling, temporal recall increases much faster than for the baseline already after a minute and at the end of the search session reaches more than a 30% greater value. This indicates that users are able to act on the cues offered by the interaction mechanism and can benefit from these actions.

In the first few minutes the performance of the setups is equal. A possible explanation is that at the beginning of the search, the users seeking with the baseline can come up with sufficient queries that result in obvious documents. As the search progresses and the users need to think of more specific queries, recall of relevant or obvious documents drops. However, when interactive user modeling is employed, the users are able to direct their search more efficiently while at the same time preserving the search context. We attribute this to the user modeling that allows users to obtain a wider set of relevant documents through enhanced interaction. This finding is supported by the analysis of the interaction logs. Users in the reinforcement interaction condition performed significantly more interactions with the system (on average 14.7) than users of the baseline (on average 8) within the same time restrictions. The *exploratory view* was used almost three times more than keyword typing and the interactions were in shorter intervals. Also participants spent more time on average after a typed query (60 seconds) than after manipulating the *exploratory view* (47 seconds). This suggests that the visualization of the user profile also assisted users to decide faster whether the returned information was sufficient and which directions to take to further refine their expressions of information needs.

EXPERIMENT 2: User performance

The purpose of the second experiment was to measure the *SciNet* system employing interactive user modeling against a real-world baseline in terms of user performance: i.e., the quality of answers that the users' provided as responses to a given task when using a given system.

Experimental Design

The experiment followed a within-subjects design. We compared two systems: *SciNet* and Google Scholar. Participants were asked to select at least 10 articles and at least three subtopics under each of the three questions. Task performance was measured based on expert assessments acquired for user responses (i.e. the documents and answers users provided in response to the task description were graded by experts). To minimize learning effects, we counterbalanced between the conditions and the tasks.

Participants and procedure

We recruited 20 participants from two different universities to participate in the study. All the participants were full-time researchers or faculty. They were all screened to ensure that their prior-knowledge was insignificant in influence on the search task by excluding participants who were either experts or complete novices regarding the topic of the task.

The participants were explained each task and use of the system. A short training session was given before the first task. The time allocated for each task was restricted to 10 minutes. Within this time period the participants had to both perform information seeking and find the subtopics. At the end of each task the users filled ResQue questionnaires (Pu et. al., 2011). Once they had completed all three tasks with each system they filled in the System Usability Scale (SUS) questionnaires (Brooke, 1996). We allowed the participants to take a break after they had completed tasks with one system. After performing all tasks in both systems they were interviewed on their overall experience with the system. Each study lasted for about 120 minutes. The participants received two movie tickets as a compensation of their time.

Measurement

Task performance was measured based on experts' double-blind assessments on each document and subtopic given as answers by the users, altogether 1800 graded relevance assessments were made for 600 individual items in the responses (article or category reported as an answer) in three relevance categories (relevant, novel, obvious). The experts were unaware of the particulars of participants and the system that they had used to perform the task. Similarly to the retrieval performance experiment, the experts evaluated each answer for three distinct properties: relevance, novelty, and obviousness. A five point Likert scale was used for rating. Participants' responses that did not contain the minimum number of required articles required were replaced with blank slots and marked as not relevant.

The subjective usability of the given system was measured using post-test ResQue questionnaires. ResQue consists of 60 questions falling into eight higher-level categories. It was chosen because exploratory search and recommender systems both highlight the importance of the users' overall satisfaction, including that with the interaction, and the ability to comprehend between the options offered by the system. The System Usability Scale (SUS) was used to evaluate overall system usability. In addition, we conducted post-test interviews to obtain users qualitative opinions about the interaction techniques and experiences of the systems.

The statistical significance of the results was measure using two-tailed t-test. The normal distribution of the data was first ensured using the Shapiro-Wilk test. The chosen significance levels were (*=$p<0.05$) and (**=$p<0.01$). We assessed inter-annotator agreement using a partially doubly annotated set of assessments by two experts. Intra-class-correlation was found to be 0.54 ($p<0.01$) which indicates a moderate to substantial agreement between the experts.

Results

Task performance results are illustrated in Figure 5. *SciNet* achieves significantly better relevance than Google Scholar (3,27/2,68, $p<0.05$). *SciNet* also significantly improves the ability of the scientists to find novel (2.6/2.2, $p<0.01$) and diverse (2.7/2.3, $p<0.01$) information. In terms of answers that were assessed obvious, *SciNet* users achieved equal performance to Google Scholar users, on average (3.3/3.2, no significant difference).

Subjective usability evaluation shows that the participants preferred *SciNet* over Google Scholar. The usability of the systems was found to be equal according to the SUS with *SciNet* scoring 53.2 and Google Scholar 50.8. A statistically significant difference between the SUS assessments was not found. A more detailed evaluation using ResQue, however, shows significant differences. The ResQue results were analyzed with both higher-level categories and individual questions. The result found is that *SciNet* outperforms Google Scholar in all higher-level categories in the standard ResQue method, except attitude towards the system, which was found to be equal. The results are illustrated for the eight higher-level categories in Figure 6.

The users felt the interface (3.5/2.9, $p<0.01$) and interaction (3.4/2.8,$p<0.01$) of *SciNet* to be better, felt that they obtained better results (3.4/3.0, $p<0.01$), felt *SciNet* to be more useful (3.3/2.9, $p<0.01$) and easier to use (3.6/3.0, $p<0.01$), and felt more in control with *SciNet* (3.4/2.9, $p<0.01$) and would use the system again (3.8/3.0, $p<0.01$).

In terms of the individual ResQue questions, the greatest differences were found in the clarity of the information provided by the system (3.9/3.0, $p<0.01$), ease of expressing preferences (3.5/2.7, $p<0.01$), and altering the outcome of the results (3.9/2.9, $p<0.01$), the presentation of the results (3.9/2.8, $p<0.01$) and assistance of users in the seeking process (3.7/2.6, $p<0.01$).

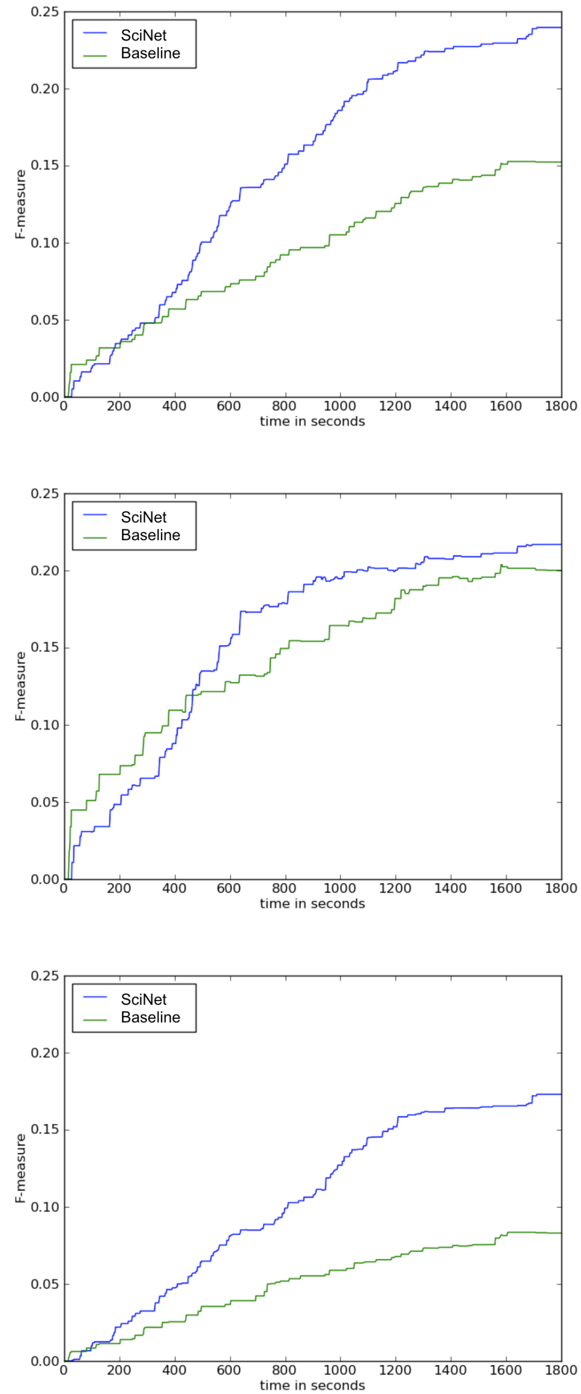

Figure 4: Cumulative F-measure over time for overall relevance (top), obvious documents (middle) and novel documents (bottom). The SciNet condition with direct information interaction significantly outperforms the typed-keyword baseline in the cases of overall relevance and novelty, and has equal performance in the case of obvious documents (Wilcoxon Signed Rank test, df=2, p<0.001).

The analysis of qualitative data originating from post-test interviews resulted in two findings. First, the participants mentioned an increased support of perceiving and giving feedback, and feel of control when using the *exploratory view*: "*Suggesting keywords make the system very easy to use and identify related keywords that I didn't know*.", "*visual search is awesome, seeing the centrality of the keywords from the circle.", and "the rich keyword selection option provides new cues, even for well known topics."*

Some participants also raised concerns about the interaction: *"The visualization of a circle was interesting, but sometimes I was loosing important keywords."*

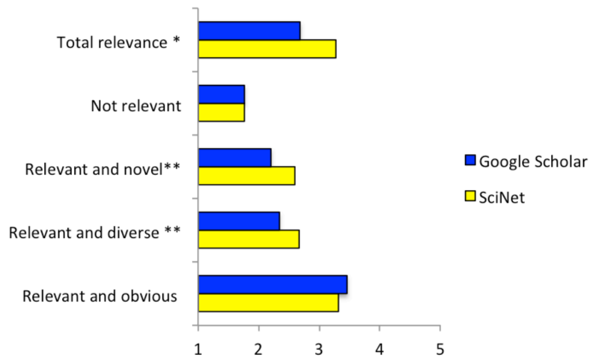

Figure 5: Average of combined expert scores from all tasks for article level assessments on a 5-point Likert scale. The SciNet condition resulted in better total relevance, more novel and more diverse set of articles than the Google Scholar condition. Both conditions resulted in an equal amount of not relevant and obvious articles. (Two-tailed t-test * p<0.05, ** p<0.01)

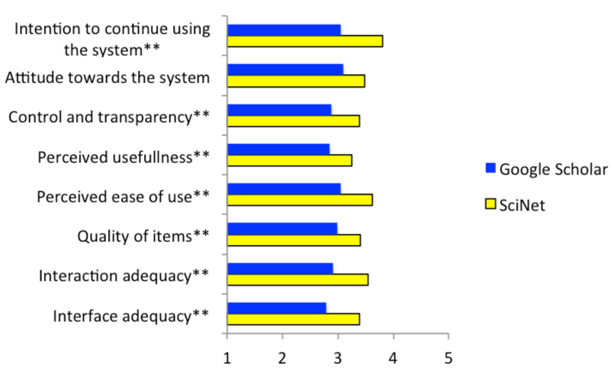

Figure 6: Average of the ResQue questionnaire categorized under the eight top-level categories on a 5-point Likert scale. The SciNet is favored in all categories over Google Scholar by participants, except attitudes toward the system, for which no statistical significance could be found. The statistical significance holds also on the level of each individual question under the seven significant categories. (Two-tailed t-test, ** p<0.01)

DISCUSSION

The results obtained in the two studies show that (RQ1), the information seeking behavior of users was affected by interactive user modeling. In particular, we found that when offered users adopted the interactive user modeling as their primary interaction technique. The frequency of user interactions was three times higher and interactions with the exploratory view were twice as common than typed queries.

This indicates that when offered, interactive user modeling was used as the main interaction mechanism, even if the users could have used only the query-based interface component. Interactive user modeling increased interaction with the system and partially replaced the need to type queries.

Second, (RQ2), while the change of the information seeking behavior is a sign of successful interaction design, the usefulness of the design should be measured by the *quality* of information users were able to find and select as a response to the given task. In our experiments, the change in information seeking behavior towards the use of interactive user modeling features turned in both 1) improved retrieval efficiency (better results returned by the system) and most importantly 2) improved task-performance (better answers provided by participants). The improvement in retrieval efficiency can be mainly explained by increased recall without losing precision. This improvement is manifested in 30%-50% increase in F-measure. Most importantly, the improved retrieval efficiency transfers in improved task-performance, even when evaluated against a strong real world baseline Google Scholar: interactive user modeling leads to better human task performance in providing answers to the given tasks.

Finally, the users subjectively preferred *SciNet*. In the subjective assessments users reported that they are able to utilize the visualization and interaction also to make sense of the search domain better and their ability to make decisions on the information available was improved.

CONCLUSIONS

Characteristic to exploratory search is that insights emerge *during* seeking to guide and structure the process. This paper has contributed a novel approach for interactive exploratory search. We demonstrated that interactive user modeling allows the user to control their exploratory search in an intuitive way and the user studies show that users can readily adopt this interaction to partially replace query typing as the input mechanism. Most importantly, this adoption leads to significantly improved retrieval and task performance.

Our results show that users adopt interactive user modeling and are more successful in their task performance using interactive user modeling. However, there is room for future work. The effect of different types of tasks and user modeling durations, such as long-term modeling that goes beyond individual search sessions should be investigated. Also the role of other personalization dimensions, such as difficulty of the retrieved contents and different levels of user pre-knowledge could be interesting to gain more insight beyond topical customization.

ACKNOWLEDGMENTS

The data used in the experiments is derived from the Web of Science prepared by THOMSON REUTERS, Inc., Philadelphia, Pennsylvania, USA: Copyright THOMSON REUTERS, 2011. All rights reserved; and the digital libraries of the ACM, IEEE and Springer.

REFERENCES

Ahn J.W, & Brusilovsky, P. Adaptive visualization for exploratory information retrieval, Inf. Proc. & Man., Available online 21 March 2013.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research,* (2002), 397–422.

Brooke, J. SUS: A Quick and Dirty Usability Scale. *Usability Evaluation in Industry*, (1996).

Byström, K, & Järvelin K. Task complexity affects information seeking and use, Inf. Proc. & Man., 31(2), (1995), pp. 191-213.

Carpineto, C., Osinski, S., Romano, G., & Weiss, D. A survey of web clustering engines. *ACM Comput. Surv., 41*, (2009), 1-38.

Chowdhury, S., Gibb, F., & Landoni, M. Uncertainty in information seeking and retrieval: A study in an academic environment. *Inf. Proc. & Man.,* (2011), 157–175.

Clarke C. L.A., Kolla M, Cormack G, Vechtomova O, Ashkan A, Büttcher S, & MacKinnon, I. Novelty and diversity in information retrieval evaluation. *Proc. SIGIR (2008)*, 659-666.

Cutting, D., Karger, D., Pedersen J., & Tukey, J. Scatter/gather: a cluster-based approach to browsing large document collections. *Proc. SIGIR,* (1992), 318-329.

Glowacka, D., Ruotsalo T., Konyushkova K., Athukorala K., Jacucci G., & Kaski S. Directing exploratory search: Reinforcement learning from user interactions with keywords. In Proc. IUI'13, 117–128, (2013).

Fox, E.A., Neves, F.,Yu, X., Shen, R., Kim, S., & Fan, W. Exploring the computing literature with visualization and stepping stones & pathways. *Commun. ACM 49*, (2006), 52-58.

Gersh, J., Lewis, B., Montemayor, J., Piatko, C., & Turner, R. Supporting insight-based information exploration in intelligence analysis. *Commun. ACM 49*, (2006), 63-68.

Hart, S.G., & Staveland, L.E. Development of NASA- TLX (Task Load Index): Results of empirical and theoretical research. *Human mental workload, Advances in psychology, 52*, (1988), 139-183.

Hearst, M., & Pedersen, P. Re-examining the cluster hypothesis: scatter/gather on retrieval results. *Proc. SIGIR,* (1996).

Hearst, M. Clustering versus faceted categories for information exploration. *Commun. ACM, 49*, (2006), 59-61.

Horvitz, E. Principles of mixed-initiative user interfaces. *Proc. SIGCHI,* (1999), 159-166.

Ingwersen, P., & Järvelin, K. The Turn: Integration of Information Seeking and Retrieval in Context. *Heidelberg: Springer,* (2005).

Käki, M. Findex: search result categories help users when document ranking fails. *Proc. SIGCHI,* (2005), 131-140.

Leide, J. E., Cole, C., Beheshti, J., Large, A., & Lin, Y. Task-based information retrieval: Structuring undergraduate history essays for better course evaluation using essay-type visualizations. *J. Am. Soc. Inf. Sci., 58*, (2007), 1227–1241.

Liu, T. Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval,* (2009), 225–331.

Marchionini, G. Exploratory search: from finding to understanding. *Comm. ACM 49*, (2006), 41-46.

Pitokow, J., Schütze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar E., & Breue, T. Personalized search. *Comm. ACM,* (2002), 50-55.

Pu, P., Li, C., & Hu, R. A user-centric evaluation framework for recommender systems. *In Proc. RecSys,* (2011), 157-164.

Ruotsalo, T., Peltonen J., Eugster, M.J.A., Glowacka, D., Konyushkova K., Athukorala K., Kosunen, I., Reijonen A., Myllymäki P., Jacucci G. & Kaski S. Directing Exploratory Search with Interactive Intent Modelling. *In Proc. CIKM,* (2013), To appear.

Sutcliffe, A., Ennis, M., & Watkinson, G. Empirical studies of end-user information searching. JASIST 51, 13 (2000), 1211-1231.

White, R. W., Bennett, P.N., & Dumais, S.T. Predicting short-term interests using activity-based search context. *Proc. CIKM,* (2010), 1009-1018.

White, R., Kules, B., Drucker, S., & Schraefel, M. Supporting exploratory search. *Commun ACM, 49,* (2006), 37.

Yee, K., Swearingen, K., Li, K., & Hearst, M. Faceted metadata for image search and browsing. *Proc. SIGCHI*, (2003).

Zhai, C., & Lafferty, J. A. study of smoothing methods for language models applied to Ad Hoc information retrieval. *Proc. SIGIR*, (2001), 334-34