Beyond Relevance: Adapting Exploration/Exploitation in Information Retrieval

Kumaripaba Athukorala[◦] **Alan Medlar**[⋆] **Antti Oulasvirta**[⋄] **Giulio Jacucci**[◦] **Dorota Głowacka**[◦]

[◦] Department of Computer Science, University of Helsinki

[⋆] Institute of Biotechnology, University of Helsinki

[⋄] Department of Communications and Networking, Aalto University

[◦] first.last@cs.helsinki.fi, [⋆] alan.j.medlar@helsinki.fi, [⋄] antti.oulasvirta@aalto.fi

ABSTRACT

We present a novel adaptation technique for search engines to better support information-seeking activities that include both lookup and exploratory tasks. Building on previous findings, we describe (1) a classifier that recognizes task type (lookup vs. exploratory) as a user is searching and (2) a reinforcement learning based search engine that adapts accordingly the balance of exploration/exploitation in ranking the documents. This allows supporting both task types surreptitiously without changing the familiar list-based interface. Search results include more diverse results when users are exploring and more precise results for lookup tasks. Users found more useful results in exploratory tasks when compared to a baseline system, which is specifically tuned for lookup tasks.

Author Keywords

Exploratory search; models of search behavior;
reinforcement learning; lookup search; adaptive systems.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

We present a new technique to support two common categories of search tasks: (1) exploratory search, where the goals are ill-defined and may change as search progresses, and (2) lookup search, where the user has a specific target in mind [33, 40]. A lookup search begins with the user expressing their information need as precisely as possible to reach the correct area of the information space, followed by quickly browsing through relevant results, until finally settling on the most appropriate item. By contrast, user behavior in exploratory search is highly dynamic. Users begin exploration with no clear search goals in mind and issue search queries that are imprecise at first. They browse through the search results and iteratively reformulate their queries using new keywords they discover. In contrast to lookup tasks, exploratory search tasks may last from a couple of hours to even years [33].

Our work addresses the observation that most information retrieval (IR) systems target only lookup search, even though exploratory search is almost as common [8]. Most IR systems present a ranked list of documents in descending order of their relevance to the issued search query, aiming to optimise the precision and recall of the search results. However, in exploratory tasks users are uncertain how to formulate search queries [14]. Retrieving the best matching results for the search query might trap the user in their initial query context and this may contribute to user perceptions that exploratory search is challenging.

While most prior work focuses on creating specialized interfaces for exploratory search [17], we present an approach that can support both lookup and exploratory search tasks with no changes to the familiar list-based search interface. Although specialized interfaces might be useful for exploratory tasks, they may not be ideal for lookup tasks. Research shows that, in general, users prefer the simple interfaces used to support lookup tasks [24]. We study an approach that can simultaneously support exploratory and lookup searches.

Our approach identifies the user's search task and adapts the search results accordingly. It only requires adaptations to the retrieval algorithm, which is reflected in the diversity of the returned results. Recent work demonstrates that setting the search engine's parameters so that the search results better match the level of user's knowledge of a specific area improves user satisfaction and performance in exploratory search [4]. Reinforcement learning techniques provide a way to parametrize the search system to trade off between exploration – presenting the user with diverse topics, and exploitation – moving towards more specific topics. While [4] determines an optimal parameter setting for exploratory search for a specific type of user, it does not differentiate between exploratory and lookup search. However, in order to support the diverse needs of users of search engines, it is vital for an adaptive search system to distinguish between exploratory and lookup search, and adjust its parameters and search results accordingly. Empirical studies show that lookup and exploratory search can be distinguished by behavioral characteristics, such as query length, reading time, and task completion time [2]. We show how these behavioral characteristics can be utilized by a search engine to automatically detect what type of search the user is conducting and thus allow the

IUI 2016, March 7–10, 2016, Sonoma, CA, USA.
Copyright © 2016 ACM 978-1-4503-4137-0/16/03 ...$15.00.
DOI: http://dx.doi.org/10.1145/2856767.2856786

search engine to provide tailored support for this specific task type. Unlike previous studies, our approach does not require any changes to the system's interface and does not require the user to specify beforehand what type of search they are planning to do.

To subject our proposed approach to empirical evaluation, we developed a prototype system that embodies this concept. Our prototype focuses on information-seeking tasks in the context of scientific literature search – but note that this approach is suited for other search domains as well. We selected scientific literature search because exploratory search is one of the most common purposes of search among researchers [3]. We used the arXiv data set as our document corpus because it is freely available and commonly used by researchers.

To evaluate the system, we designed an experiment that captures the key elements of exploratory and lookup tasks. In the experiment, researchers from the computer science field performed both exploratory and lookup tasks using our system and a baseline system. Our hypothesis is that by adapting the search system to set a higher exploration rate for exploratory tasks and a zero exploration rate for lookup tasks, we would improve user performance. As the baseline system, we use a traditional search system which sets the exploration rate to zero to maximize performance in lookup tasks. Results of our user study show that the adaptability of a search engine to both exploratory and lookup search behaviors improves user satisfaction and performance in search.

BACKGROUND

In this section we briefly review contributions of various research communities to the study of exploratory search. We revisit the existing approaches to shed light on some of the open problems in exploratory search. This review also guides us in formulating our approach to exploratory search.

Relevance Feedback

The first key approach developed by the IR community was relevance feedback: users mark documents as relevant or non-relevant and the query model is updated accordingly. Initial experimental studies showed that interactive IR systems benefited from term relevance feedback [30], though evidence from later studies showed these features to be rarely used and, therefore, do not improve retrieval results. The reasons are two-fold [30]: (1) relevance feedback often leads to context traps, and (2) the cognitive load of selecting relevant documents is high compared to typing a new query. Improved modeling of information-seeking behavior would allow an IR system to suggest more relevant results and avoid getting stuck in context traps.

Faceted Search

Faceted search aims to avoid the context trap by using global features instead of contextual ones [42]. It organizes information according to a faceted classification system, allowing users to explore collections of documents by applying multiple filters. Such systems classify information elements along explicit global dimensions, enabling the classifications to be accessed and ordered in multiple ways. As the number of global features is often very large, the process, however, can quickly become overly demanding as users have to go through a large number of options [42]. Improved modeling of user behavior and needs would allow a reduction in the number of facets and thus enhance user experience.

Reinforcement Learning

In exploratory search, the information-needs of a user are constantly evolving, and hence modeling is crucial for acquiring information on search intent as well as query reformulation [39]. Recently, reinforcement learning techniques have been applied to modeling exploratory search [20, 29, 37]. Reinforcement learning allows a system to balance exploitation (moving toward more specific topics) and exploration (presenting the user with alternative topics). The main disadvantage is the "cold start" problem: these systems need a number of sessions to adjust to a users information needs. A recent study showed that systems employing reinforcement learning techniques can optimally balance exploitation and exploration for a given set of users [4]. The study compared different levels of exploration in the retrieval algorithm and identified the right balance that improves user performance, satisfaction and interactions with results in exploratory search. This study provided useful cues for designing our adaptive search system.

Adaptive Systems

Text classification has been a popular topic in machine learning for decades. However, applications related to the problem of online adaptive learning only appeared relatively recently [19]. Most examples of text stream applications include e-mail classification [12], e-mail spam detection [32] and sentiment classification [9]. Various adaptive learning strategies have been used in this domain, including individual methods like case-based reasoning [11], and ensembles, either evolving or with an explicit detection of changes by means of change detectors [1, 10, 21]. All these methods adapt to the gradual change in either user interests or data distribution. Our proposed approach allows the system to adjust to the information needs of the user and the type of search they are conducting.

Exploratory Search Interfaces

The lack of success of relevance feedback techniques is often attributed to user interface design failing to conveniently provide feedback at suitable levels of granularity. In response, a variety of systems were designed to support user feedback, including intelligent user interfaces that assist the user in comprehending an information space [7, 13], visualizations that summarize results for faster relevance judgement [27, 31, 34], as well as interactive visualizations that allow the user to indicate the direction of exploration [20]. These systems are very useful but designed only for exploratory tasks. For lookup tasks, users find such visualizations to be a distraction and, in general, users prefer simple interfaces [24]. This brings us to an important question: is it possible to keep the interface simple and yet provide extra support for exploratory information seekers? To this end, we need a system that could predict the task type in the course of a search session and dynamically adapt the search engine according to the task type.

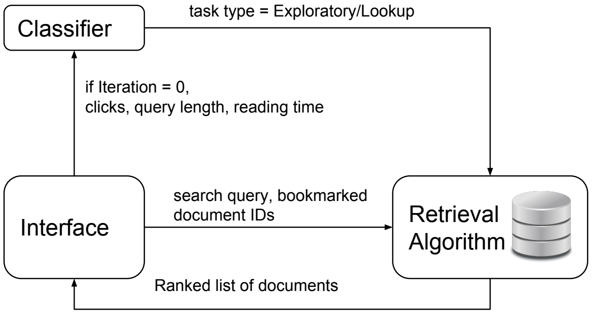


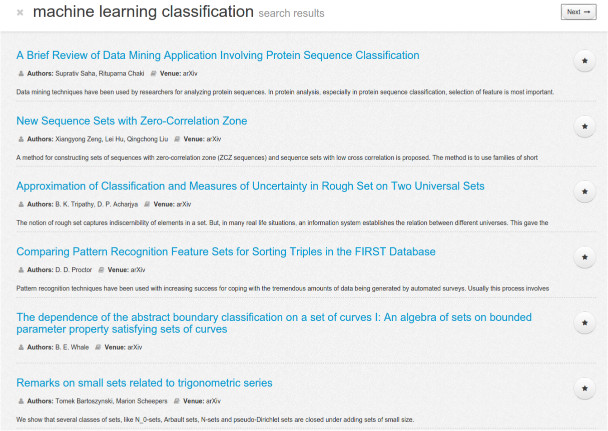
Figure 1. System Overview: the user types in a query and investigates the first page of the search results. Based on the user interaction with the search results displayed on the first page, the search is classified as either lookup or exploratory. The result of the classifier is passed on to the search engine, where the exploration rate of the retrieval algorithm is set according to the search task. Once set, the exploration rate is kept constant for the remainder of the search session.

Models of Exploratory Search

Models of user behavior in information search help us to provide personalized support. Information Foraging Theory is a promising model of exploration that predicts the information seeker's decision from the expectation of information gain [35]. There are also models that assist the search systems in predicting user perception of the results from implicit interactions, including search satisfaction [22], frustration [18], and disambiguating exploration and struggling scenarios [23], as well as predicting the domain knowledge of the user [16]. A recent study reports how to distinguish between exploratory and lookup tasks from implicit measures of search behavior [2]. This study has important implications for designing adaptive search systems. We integrate a similar model in our search system to predict the task type in the course of a search session.

Summary and Contribution

In this section, we briefly summarized some of the existing approaches to exploratory search ranging from IR methods, such as relevance feedback and faceted search, through machine learning methods, most notably reinforcement learning, to adaptive search systems and visualizations for exploratory search. Although all of these methods and systems can clearly improve support for exploratory search tasks, they often need a long training period to spot the gradual change in user behavior or require the users to explicitly state that they are planning to conduct an exploratory search. Additionally, most existing systems use specific visualization techniques that can only support exploratory search, which means that the user cannot rely on the same system to conduct both lookup and exploratory searches. Contrary to previous studies, our approach uses simple behavior characteristics, such as clicks and reading time, to automatically classify the search type and automatically adjust the search engine parameters without the need for additional user feedback or visualization techniques.



machine learning classification search results

A Brief Review of Data Mining Application Involving Protein Sequence Classification

Authors: Suprativ Saha, Rituparna Chaki **Venue:** arXiv

Data mining techniques have been used by researchers for analyzing protein sequences. In protein analysis, especially in protein sequence classification, selection of feature is most important.

New Sequence Sets with Zero-Correlation Zone

Authors: Xiangyong Zeng, Lei Hu, Qingchong Liu **Venue:** arXiv

A method for constructing sets of sequences with zero-correlation zone (ZCZ sequences) and sequence sets with low cross correlation is proposed. The method is to use families of short

Approximation of Classification and Measures of Uncertainty in Rough Set on Two Universal Sets

Authors: B. K. Tripathy, D. P. Acharjya **Venue:** arXiv

The notion of rough set captures indiscernibility of elements in a set. But, in many real life situations, an information system establishes the relation between different universes. This gave the

Comparing Pattern Recognition Feature Sets for Sorting Triples in the FIRST Database

Authors: D. D. Proctor **Venue:** arXiv

Pattern recognition techniques have been used with increasing success for coping with the tremendous amounts of data being generated by automated surveys. Usually this process involves

The dependence of the abstract boundary classification on a set of curves I: An algebra of sets on bounded parameter property satisfying sets of curves

Authors: B. E. Whale **Venue:** arXiv

Remarks on small sets related to trigonometric series

Authors: Tomek Bartoszynski, Marion Scheepers **Venue:** arXiv

We show that several classes of sets, like N_0-sets, Arbault sets, N-sets and pseudo-Dirichlet sets are closed under adding sets of small size.

Figure 2. Interface: the user types in a query at the top of the page and a set of search results is displayed. When the title of a document is clicked, the document's abstract is displayed below. The user can bookmark the documents that they like by clicking on the circle next to the title. By clicking the "next" button at the top of the page, the user is taken to the next search iteration.

APPROACH

Our system can be broadly divided into three parts: the user interface, the task type classifier and the search engine containing the document ranking algorithm. The current version of the system operates on $\sim$1 million documents obtained from the arXiv repository.

System Overview

The overview of the system is presented in Figure 1. After typing in the search query, the user is presented with 20 documents, where seven documents are visible without scrolling. Figure 2 shows the interface of the system. We display more documents than in traditional IR systems because in exploratory tasks users examine more results [41]. The initial set of documents is ranked based on the Okapi BM25 algorithm [38]. Next, to see more documents the user can indicate which documents interest him by clicking on the circle next to a given document and thus provide a relevance score of 1 to that document – we refer to this action as *bookmarking*. The user can bookmark as many or as few documents as he wishes. After clicking the "next" button in the top right corner of the page, a new set of documents is displayed based on the feedback provided by the user so far. Documents that do not receive an explicit relevance feedback are assumed to receive a relevance score of 0. While the user is interacting with the first Search Engine Results Page (SERP), the user interface logs three actions performed with the first SERP: the length of the query, total number of clicks (cumulative clicks), and the total time spent reading the clicked documents. These logged interactions are passed on to the classifier to categorise a given search session as either lookup or exploratory, which, in turn, allows the system to set the exploration rate at an appropriate level for the proceeding iterations. Note that the bookmark actions are not used by the classifier, rather they are used by the reinforcement learning algorithm to determine the relevance feedback given to the documents.

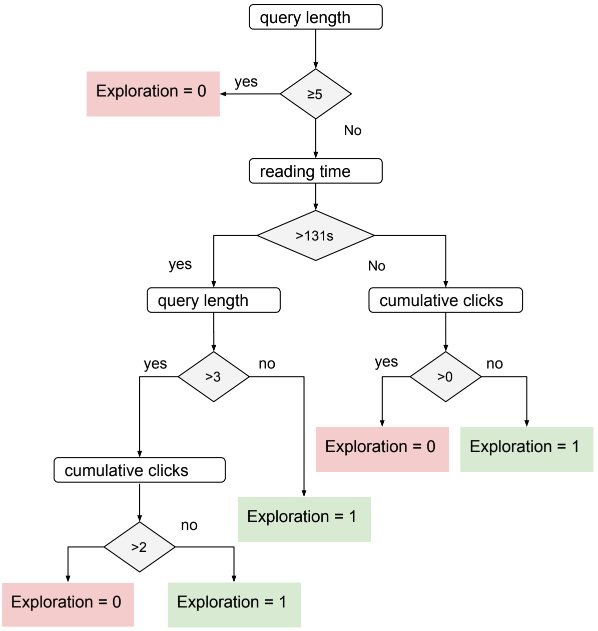


Figure 3. C4.5 decision tree trained using 10-fold cross-validation. When the user clicks the next button after exploring the first search engine results page, the interface sends query length, reading time, and cumulative clicks to the classifier. The classifier uses this decision tree to predict whether the search task type is exploratory or lookup.

Classifying Search Type

To build the classifier, we followed a previous study which identifies implicit search behaviors to distinguish exploratory search from lookup [2]. In that study, exploratory and lookup tasks were operationalized according to the preciseness of the search goal and the complexity of the search task. According to that operationalization, there are core exploratory/lookup tasks with dominant exploratory (high complexity and open-ended goals) or dominant lookup (low complexity and precise goals) characteristics. There are also borderline tasks with mix-characteristics. They designed an interface similar to Google scholar for the arXiv data. The interface included 40 results in first SERP and seven items were visible without the need to scroll. In all other aspects this interface is similar to our interface in Figure 2. The only difference is that our interface has the bookmark feature which is not a problem because the classifier does not use the bookmark interaction as a feature. They conducted a controlled study with six search tasks (two core exploratory and lookup tasks, and one borderline task from each category) and demonstrated how a classifier separates the two task categories from six observable search behaviors: query length, query duration, maximum scroll depth, cumulative clicks, proportion of browsing, and reading time. According to their data, query length, scroll depth, reading time, task completion time, and cumulative clicks are the most discriminative features. We used the data set reported in the same study and trained three different classifiers using 10-fold cross-validation to distinguish known item search and knowledge acquisition tasks [36]: C4.5 decision tree, bagging classifier, and random forest classifier. C4.5 decision tree performed the classification task with 80% accuracy, while bagging and random forest classifiers performed with 79% and 81% accuracy, respectively. We chose the C4.5 decision tree because its transparency allows us to simply understand the exact cause of classification failures, and its performance is comparable to the other classifiers tested.

We selected known item search and knowledge acquisition tasks because known item search is the most common lookup task and knowledge acquisition is one of the main purposes of exploratory search [33]. The classifier identified three features that could best distinguish the two tasks with 80% accuracy: (1) query length – total number of terms in the first query. A term is defined as "a string of characters separated by some delimiter, such as a space, a colon, or a period" [25]; (2) reading time – the time users spent investigating clicked documents; and (3) cumulative clicks – the total number of links in the results page clicked by the user. A schematic of the decision tree is given in Figure 3. We integrated the classifier into our system. As indicated in Figure 1, the classifier receives the user interaction data from the search interface when the user selects the next button to progress to the second iteration. Then, the classifier predicts whether the task is exploratory or lookup and passes this information on to the document ranking component, which sets the exploration rate according to the task type. The interface sends which articles were bookmarked to the retrieval algorithm.

Ranking Documents

In order to help the user to explore the document space, we use LinRel [6], which is used from the second iteration onwards. Suppose we have a matrix D, where each row d_i is a tf-idf feature vector representation of documents presented so far. Let $r = (r_1, r_2 ... r_t)^\top$ be the column vector of relevance scores received from the user up to time t. We estimate the expected relevance r_i of a document d_i as $\mathbb{E}[r_i] = d_i \cdot w$, where the vector w is estimated from user feedback. LinRel estimates $\hat{w}$ by solving $r = D \cdot w$ and estimates relevance score for each d_i as $\hat{r_i} = d_i \cdot \hat{w}$

In order to deal with the exploration/exploitation trade-off, we present documents not with the highest score $\hat{r_i}$, 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 $\hat{r}_i$, the upper confidence bound of document d_i is calculated as $r_i + \gamma\sigma_i$, where $\gamma \geq 0$ is a constant used to adjust the confidence level of the upper confidence bound. In each iteration, LinRel calculates

$$s_i = d_i \cdot (D^\top \cdot D + \lambda I)^{-1} D^\top, \quad (1)$$

where λ is the regularization parameter which is set to 1 if each of the feature vectors sums up to 1 (following [6]) and the documents that maximize $s_i \cdot r + \frac{\gamma}{2}\|s_i\|$ are selected for presentation. The first term $s_i \cdot r$ effectively ranks all documents based on their similarity to the documents the user has selected so far and thus it narrows the area of the search space (exploitation). The second term $\frac{\gamma}{2}\|s_i\|$ is an exploration term which ensures that the user is presented with a more diverse set of results. The value of γ can range from 0 to infinity.

The higher the value of γ, the more diverse, or exploratory, the results are. Based on the results of the classifier described above, i.e. whether a given search is classified exploratory or lookup, the value of γ is set to either 0 or 1. If $\gamma = 0$, then the system only exploits – this setting is used in searches classified as lookup. If a given search is classified as exploratory, then the value of γ is set to 1. According to [4], with $\gamma = 1$ the system provides optimal support for exploratory search, maximizing user satisfaction and the number of selected documents by the user.

USER STUDY

The purpose of this study is to empirically evaluate our hypothesis that an adaptive search system, which dynamically adjusts the exploration rate according to the task type, improves user performance compared to traditional search systems that are designed to support only lookup tasks. Our system uses a classifier that takes the user interaction data – cumulative clicks, query length, reading time – as input and predicts whether the user is performing an exploratory or lookup search task. We refer to this system as the *full system*. The *baseline system* that we use treats all the tasks as lookup tasks with the exploration rate fixed at 0. We particularly selected this setting for the baseline system because most of the existing search systems are similarly tuned to exploit the search query with no adaptation to task type [33].

Participants

We recruited 18 researchers from a Computer Science (CS) department because researchers have more experience in exploratory search tasks [3]. Eight of the participants were PhD students and 10 were masters thesis writers. All the participants are researchers in the machine learning domain. We selected the machine learning domain because our data set well covers this domain and also in our department it is easy to get help from experts in this domain to assess the task performance. Four participants were female, which reflects the 20% gender distribution in the CS department. According to the background questionnaire, all the participants often explore unfamiliar topics for research purposes (mean = 3 in 5 point Likert Scale where 1 = never, and 5 = daily). Google scholar is the primary literature search tool for 16 participants, while others use a combination of tools including Google Scholar, ACM digital library and arXiv.

Design

We designed a within subject study, where every participant performed two search tasks – one exploratory and one lookup – in each system (full system and baseline system) resulting in 4 tasks in total per participant. We needed both exploratory and lookup tasks to validate the performance of the classifier. To avoid order effects, we counter-balanced the sequence of performing the tasks and the search system using the Latin square design.

Tasks

Lookup tasks were designed according to a typical known item search scenario, where the participants look for an article that they have come across before. An expert researcher from the machine learning domain selected six articles for the known item search task, where three of them were from the topic image annotation, and the others are from the topic face recognition. We selected these topics because none of the participants were directly involved in research on these topics before, which creates a homogeneous set of participants to avoid biasing the results. Additionally, an expert tested the retrievability and availability of these articles in our data set. We randomly picked one article from each topic and asked the participants to skim through them two hours prior to the study. We described the lookup tasks using the following template:

> *"Do you remember the article that you read two hours ago about the topic X. Imagine that you are conducting a literature review on topic X and decided to refer to this particular article. Try to find this article using the given search system."*

In exploratory search tasks, users typically search for the purpose of learning [40]. To situate the participants in a natural exploratory search setting, prior to the study we asked them to provide a list of topics that they would like to learn about but have limited knowledge of. Then, we selected two unique topics from the given list and assigned one topic per system. In this way, we ensured that there was a genuine motivation to learn for the participants. Table 1 shows the search topics each participant selected for exploratory search tasks.

Measurements

We logged all the details related to the performed search tasks: task type (exploratory or lookup), system used (full system or baseline), search topic for exploratory tasks, search query, and target article for lookup tasks. The system logged the classifier predicted task type for all the tasks performed in both systems as well as all the user interaction data – clicks, duration of result page browsing, depth of scrolling, and time spent reading clicked articles, bookmarked articles, and user feedback on interacted articles. We also collected qualitative feedback on the search system through interviews. An external expert reviewer from the machine learning domain rated, on a binary scale, the relevance of all the articles returned by both systems for the exploratory tasks. The reviewer was unaware of which system was used by the participants.

Procedure

Prior to the study, we provided a background questionnaire to capture the knowledge and research experience of the participants. In this questionnaire, we also asked the participants to give a list of search topics that they would like to learn about for the exploratory search tasks. We only selected those participants who listed at least two machine learning topics that have sufficient articles in our data set.

To situate the participants in the known item search scenario, before the study we invited them to our lab and provided them with two articles to skim through. We asked the participants to imagine that they came across these two articles while searching for literature and instructed them to follow their natural skimming behavior in such a scenario. To prevent them from memorizing the content, we did not tell them

Participant	System	Search topic
1	Full	clustering
	Baseline	dimensionality reduction
2	Full	heart rate classification
	Baseline	unbalanced training set
3	Full	crowd sourcing
	Baseline	collaborative media
4	Full	semi-supervised learning
	Baseline	chaotic theory
5	Full	semantic analysis
	Baseline	lens correction algorithms
6	Full	abstract argumentation
	Baseline	read error correction
7	Full	probabilistic models
	Baseline	data mining
8	Full	dimensionality reduction
	Baseline	artificial neural network
9	Full	deep learning
	Baseline	topic models
10	Full	neural networks
	Baseline	random forests
11	Full	support vector machines
	Baseline	string kernels
12	Full	Bayesian nonparametric
	Baseline	computer vision
13	Full	brain signals
	Baseline	clustering
14	Full	topic segmentation
	Baseline	speech tagging
15	Full	graph matching
	Baseline	time series analysis
16	Full	bootstrapping
	Baseline	compressed sensing
17	Full	social network analysis
	Baseline	deep neural networks
18	Full	reinforcement learning
	Baseline	speech recognition

Table 1. List of search topics each participant selected for the exploratory tasks. We randomly assigned the selected topics for the full system (Full) and the baseline system.

about the upcoming lookup task where they had to re-find the same articles. We informed them that the purpose of this part of the study is to observe their skimming behavior. We did not restrict the time allowed for skimming and the participants spent on average 13 minutes skimming both articles. Then, we allowed them to return to their usual work and come back to the study after 2 hours. This way, we ensured that the participant has some memory of the article that they skimmed through but forgot the exact details, such as the title and author information [28].

All the studies were conducted in a controlled laboratory room, on a laptop computer with an external 27-inch display. Before the study began, we demonstrated how the search system works and the participants performed a training task to become familiarized with the system. Every participant performed one exploratory task and one lookup task with each system, resulting in four tasks in total. The participants were not aware of the two different systems as both the baseline and the full system have the same interface. For exploratory tasks, we allocated 25 minutes and instructed the participants to spend the entire time searching. For the lookup tasks, we allowed 15 minutes maximum and they could end the task as soon as they found the article. The system automatically terminated the search session at the end of the allocated time. Then, the system loaded the list of articles that the participant either clicked or bookmarked to provide an explicit feedback on the relevance of each article on the scale from 1 to 5 (where 1 = not relevant, 5 = highly relevant). After every task, we conducted a semi-structured interview about their experience with the search system. Each study lasted approximately 90 minutes and we compensated each participant with a movie ticket.

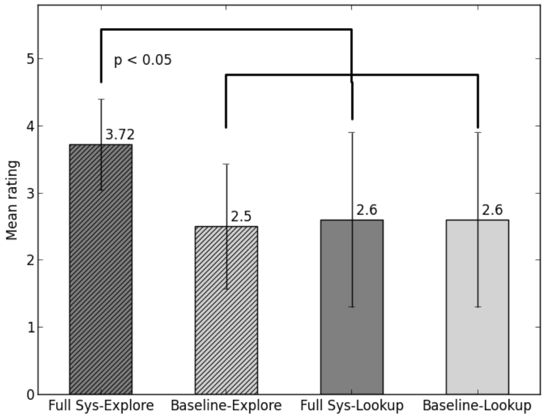


Figure 4. Mean rating given by the participants to the articles that they bookmarked in the full system (Full Sys) and the baseline system for exploratory (Explore) and lookup tasks.

RESULTS

The 18 participants performed 72 search tasks in total (18 exploratory tasks × 2 systems + 18 lookup tasks × 2 systems = 72). No data points were excluded. We analyzed both the quantitative and qualitative feedback from the participants as well as the interaction data and expert ratings.

User Perception

To understand whether there is a noticeable difference in the user perception of the two systems, we analyze the explicit scores users gave to the bookmarked items (Best score = 5, worst score = 1). We also analyze the qualitative feedback each participant gave during the interview to corroborate the scores given to the bookmarked items.

User rating: We found that users gave higher ratings for documents retrieved during exploratory tasks using the full system compared to the baseline system. Figure 4 shows the mean rating users gave to both systems for lookup and exploratory tasks (ratings are given on a 5 point Likert scale). Repeated measures ANOVA showed that the system has a statistically significant effect on the user given rating (F(1,71) =

		Classifier outcome	
		Exploratory	Lookup
Actual task	Exploratory	92%	8%
	Lookup	31%	69%

Table 2. Accuracy of the classifier according to the user study data.

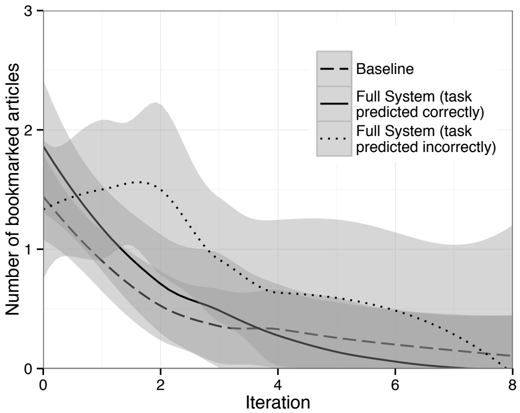


Figure 5. User behavior is influenced by incorrect classification of lookup tasks. When lookup tasks are incorrectly treated as exploratory tasks with a higher exploration rate, users tend to select more items.

4.7, $p < 0.05$). According to post hoc Tukey's test with bonferroni correction, in the exploratory tasks, participants gave a significantly higher rating for the results in the full system (mean = 3.72, s.d. = 0.6, $p < 0.05$) than the results in the baseline system (mean = 2.54, s.d. = 0.92). These results provide quantitative evidence that in exploratory tasks users perceived the results retrieved by the adaptive search system to be more useful.

However, users gave lower ratings for lookup tasks performed in both systems. The mean rating given for lookup tasks in both the full system and the baseline coincide and there is no statistically significant difference ($p > 0.05$, Tukey's test). This result suggests that even though users bookmark documents in lookup tasks, they do not find them to be relevant enough to give them a high score.

Qualitative feedback: Interviews further confirmed that users perceive the full system to be better for exploratory search tasks. Eight participants stated that the full system gave them different directions to explore: "I found new areas that I didn't know before" [Participant 15], "It gave me a range of options beyond what I directly typed. In other search engines I only see things related to my query" [P11]. Four participants mentioned that when they conduct exploratory searches in general, they spend a lot of time reading a single review article to learn about other related areas, but with the full system they found these areas right away: "normally I spend a couple of hours reading a book or a review to find the other topics, then in here I noticed a lot of articles not directly related to my query. When I checked them I realized, this is actually another interesting topic" [P2]. In the baseline system during exploratory tasks three participants stated that they got stuck in one topic: "Over iterations the topic got narrower and I did not know how to find other topics in here" [P8].

For lookup tasks, users gave, on average, lower ratings for articles retrieved in both systems. This is because in lookup tasks, the users' goal is to find a known item and their ratings are bimodal – rating the target item with score 5, and all the other items with the lowest scores, even if they are closely related to the target. For example, when rating a document that was very similar to the target, a participant gave it a score of 2, saying "This article is relevant, but not the one I want. So I give it 2" [P5].

According to the interviews with the participants whose lookup tasks were misclassified as exploratory, we found that most of them had forgotten the paper they were looking for: "I am sorry, I totally forgot [the details of] what I was looking for" [P16], "All I remember is one of the authors is from Facebook. I didn't find anything [relevant]" [P3].

Classifier Accuracy

We calculated the overall accuracy of the classifier by considering the entire 72 search tasks regardless of the system and calculated the percentage of the tasks that were correctly classified as exploratory/lookup. Among the 72 tasks, the classifier accurately classified 58, leading to an overall accuracy of 81%. Further analysis of how well the classifier distinguished exploratory and lookup tasks showed that it accurately classified 92% of exploratory tasks and 69% of lookup tasks (Table 2). The overall accuracy of the classifier in the full system was 78% (89% exploratory, 67% lookup), beating a random classification (50%) by a clear margin.

Further analysis of the incorrectly classified lookup tasks showed that these users issued very short search queries with two keywords, such as "deep learning", "image captioning" or "face recognition", and reading time was longer than 131 seconds. According to the classifier (Figure 3), when search queries are shorter than 3 words and reading time is longer than 131 seconds, they are classified as exploratory. This behavior is concordant with statements made by these users during their interviews that they had forgotten the details of the target article, which would have forced them to employ an exploratory strategy.

As further evidence that the users of misclassified lookup tasks behaved differently compared to those classified correctly, Figure 5 shows LOESS [15] regression lines for how the number of bookmarked articles varies as a function of the iteration number for different systems. LOESS is a

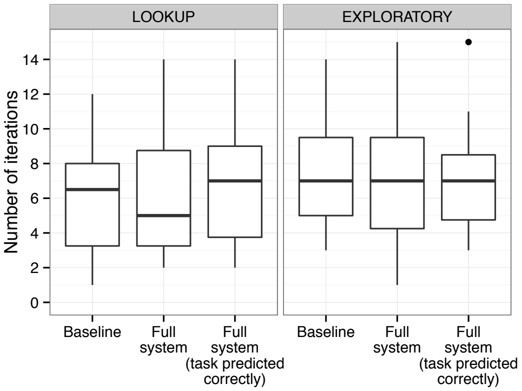


Figure 6. Boxplots of the number of iterations users spent performing lookup and exploratory tasks using the baseline, full system (all data) and full system when the classifier correctly predicted the task type.

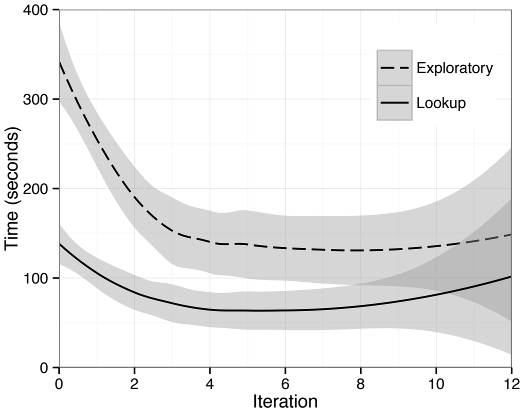


Figure 7. LOESS curves of the iteration duration for lookup and exploratory tasks. Plot only includes experiments from the full systems where the classifier correctly identified the task type. The shaded region indicates the 95% confidence interval.

method for local regression that produces a smooth curve using weighted least squares over a sliding window. Being nonparametric, LOESS does not require the data fit a single global function, but can instead highlight localised trends. Users with misclassified lookup tasks tended to bookmark more articles during the early iterations of the search session compared to both the baseline and the correctly classified full system. If these users are actually employing an exploratory search strategy, then it is indicative of a lack of user motivation in those lookup tasks, and we may, therefore, be underestimating the efficacy of the classifier.

User Behavior

We analyzed user search behavior in terms of: (1) the number of search results users clicked or bookmarked – referred to as *interactions*, (2) the number of times the user requests a new set of results by clicking the next button – referred to as *iterations*, (3) the number of bookmarked articles over iterations. Furthermore, to understand if there is a difference between the user perceived relevance and the actual relevance (according to an expert) in exploratory tasks, we compare the bookmarked articles with expert judgement.

Interactions: In exploratory search tasks, there is a higher number of interactions in the full system (mean = 15.7 interactions/participant) than in the baseline system (mean = 11.5 interactions/participant). In lookup tasks, there are fewer interactions than in exploratory tasks (lookup tasks in full system: mean = 5.2 interactions/participant, lookup tasks in baseline system: mean = 5.3 interactions/participant). According to the repeated measures ANOVA test, the task type has a significant effect on the number of interactions (F(1,71) = 15.5, $p = 0.0002$), however the effect of the system on the number of interactions was not significant (F(1,71) = 0.71, $p = 0.41$). We conclude that regardless of the system, users interact with more results in exploratory tasks than in lookup tasks.

Iterations: There were no significant differences between the number of iterations users spent with either system when performing either lookup or exploratory tasks ($p > 0.05$, paired t-tests). Figure 6 shows boxplots of the number of iterations users spent with each system performing lookup and exploratory tasks. During lookup tasks, users spent on average 6.11 (s.d. 3.20) and 6.06 (s.d. 3.57) iterations using the baseline and full systems, respectively. The average number of iterations was longer for exploratory tasks: 7.33 (s.d. 3.43) iterations with the baseline and 7.05 (s.d. 3.47) iterations with the full system. Discarding results where the classifier in the full system incorrectly predicted the task type did not lead to a statistically significant difference in the number of iterations spent in either system ($p > 0.05$, Welch's t-test).

There is a consistent difference, however, between the time taken for each iteration depending on the task type. In Figure 7 we consider only those experiments where the classifier correctly identified the task type. Users spent, on average, less time on each iteration during lookup tasks than exploratory tasks. For example, during the first iteration (between the search query and the classifier running) users spent on average 138 seconds performing lookup tasks, but 341 seconds for exploratory tasks.

User Bookmarks Vs. Expert Assessment: While expert assessment showed no difference between either system with respect to the number of relevant articles, users of the full system bookmarked more articles in the initial iterations compared to the baseline. Figure 8 shows how the number of bookmarked articles varies as a function of the iteration number for both baseline and full systems performing exploratory tasks. For a given search query, both baseline and full systems initially show the user the same articles, which is reflected in users bookmarking the same number of articles in the first iteration (iteration 0 in the graph). From the next iteration, the two systems diverge in terms of the number of articles that users bookmark. By the third iteration (iteration 2 in the graph), the users of the baseline system bookmarked less than 1 article per iteration on average. In the full system, however, the number of bookmarked articles declined gradually until

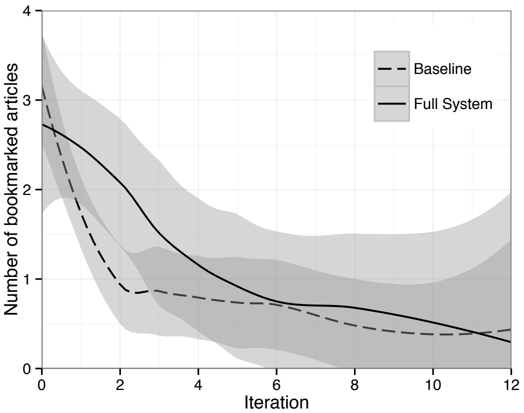


Figure 8. LOESS curves of the number of articles bookmarked at each iteration during exploratory tasks using the baseline and full system. The shaded region indicates the 95% confidence interval.

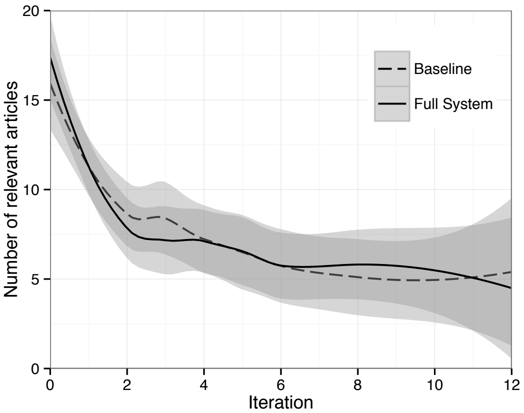


Figure 9. LOESS curves of the number of articles marked as relevant by an expert at each iteration during exploratory tasks using the baseline and full system. The shaded region indicates the 95% confidence interval.

converging with the baseline at around iteration 6. This trend suggests that while users of the full system are exposed to more relevant documents initially, the user feedback eventually overrides the exploratory terms in the ranking equation, leading to convergence with the baseline system.

We contrast these results with the expert assessment of article relevance at each iteration (Figure 9), which suggests that there is no difference between the performance of the full system compared to the baseline. This demonstrates the subjective nature of exploratory search and the need to support individual users' information needs.

DISCUSSION AND CONCLUSIONS

We contribute an approach for designing an information search system capable of supporting both exploratory and lookup search tasks. Exploratory search is gaining importance as more information is available through the web [33], yet it is considered to be one of the most challenging search tasks [3]. Perhaps this is due to the fact that most of the existing search systems are designed to support either lookup or exploratory searches but not both [13, 20]. Users are reluctant to switch between search systems for different search goals or deviate from familiar list based search interfaces to other visualizations introduced in exploratory search systems [24]. We demonstrate a new approach which aims to predict the search task type from implicit measurements of search behavior and dynamically set the parameters of the underlying retrieval algorithm.

Our approach involves using a classifier to predict whether the search goal is exploratory or lookup using three features – clicks, query length and reading time. In line with previous work [2], our results show that exploratory tasks can be distinguished from lookup tasks using search behaviors measured only from the interactions within the first SERP. The overall accuracy of our classifier is 81%, which is comparable with the 85% accuracy reported in [2]. This is an important finding given that the exploratory search tasks in our study are truly motivated by the participants. However, the accuracy of the classifier was lower in the lookup tasks (69%). According to the interviews with the participants of the misclassified lookup tasks, it is clear that they had forgotten the details of their target item and their behavior is very similar to the struggling behavior in lookup tasks reported in literature [23]. One possible reason is that the lookup tasks in our study setting were not as natural as exploratory tasks and the users did not have a true motivation to find the target or pay enough attention when skimming the articles. We see more opportunities in the future to further investigate other possible approaches to improve the classifier accuracy for lookup tasks by incorporating models that disambiguate struggling vs. exploration [23]. Overall, we showed empirically that it is possible to automatically separate search intentions early on in the course of the search session [8].

According to the mean scores users gave to the selected documents, it is evident that users noticed an improvement in the search results retrieved through our approach. A recent study which motivated the trade-off we made between exploration and exploitation in exploratory tasks also reported similar user feedback [4]. Even though the search interface and the visible features are identical in the two systems, users still favored the full system for exploratory tasks.

This finding shows that users are sensitive to the changes in the exploration parameter and a higher exploration rate is necessary for exploratory search tasks. Users' ratings also show how their expectations change according to the task type. In typical known item search tasks, as in our study, users are only interested in finding the target item [26] and they are less interested in other relevant items. For this reason, the overall mean rating for lookup tasks is lower. Another interesting finding is that users gave a similar rating for the baseline system regardless of the task type. This is unexpected because the baseline system does not adapt to the task type.

User behavior analysis sheds light on how the proposed approach influences search strategies – number of interactions, iterations and bookmarked articles. In exploratory tasks, users interacted with more results than in lookup tasks and went through more iterations [23, 33, 40]. Our approach had no significant influence on the number of interactions or iterations. These findings are interesting because they show that users were not overwhelmed by the diverse areas covered by the higher exploration rate in exploratory tasks. At the same time, we can conclude that even in the studies where the classifier incorrectly predicted the task type (22%), users followed the same strategy. This suggests that maybe the classifier incorrectly predicted the task type because the user is indeed following an exploratory strategy in lookup tasks and vice versa. However, further investigation with more studies in an uncontrolled setting would be required to make conclusive remarks.

We observed an interesting difference between the number of results that the users bookmarked as relevant and the expert rated as relevant. Even though the results from both the full system and the baseline system appear equally relevant to an expert, this is not the case for the user. Users tend to bookmark more documents as relevant in initial iterations in the full system compared to the baseline system. The expert reviewer only knew the initial search topic of the user but not the changes in user interests over time. This shows that the relevance of search results is highly subjective and can change throughout the search session [40]. As suggested in the literature, in exploratory search as users gain more knowledge on a given topic, their search interests might deviate from the original topic [5]. This is further justified by the interview findings, where users stated that they found more interesting topics different from their search topic in the full system. This finding further confirms that our approach is appropriate because in exploratory tasks it is not sufficient to retrieve results only relevant to the search query.

Our results show that in cases where lookup tasks are misclassified as exploratory, the user had to bookmark more items than in correctly classified cases (Figure 5). This suggests that users have to examine more items to find the target item. In known item search tasks the goal is to reach the target item as fast as possible [33], and having to bookmark several items indicates that it slows down the lookup process and would lead to frustration [18]. This behavior is similar to struggling behavior in lookup searches [18]. This finding points to the importance of accurately setting the exploration rate. While too few exploratory tasks were misclassified for a similar comparison, we note that tasks performed in the baseline are equivalent to misclassified exploratory tasks in the full system. In the baseline system users bookmarked fewer items than in the full system (Figure 8), which, combined with the interviews, suggests that users saw fewer articles of interest. In the current version of our system, the classifier is only run after the first iteration, however, these findings suggest a possible means to correct misclassified tasks from the second iteration onwards by monitoring user activity as the search progresses. We note that for correctly classified tasks, iteration time trivially distinguishes lookup from exploratory tasks during the initial iterations (Figure 7). The next version of the system could combine our current classifier with subsequent user bookmarking and timing information across iterations, taking a more holistic approach to task classification.

We see several future opportunities to further improve the proposed approach. An important open challenge is how to make the exploration/exploitation trade-off transparent to the user. This step needs more careful investigation of how to show the system predicted task type to the user and allow the user to correct the system's prediction. However, it is also important to avoid overloading the user with too much information as it would become a distraction [24]. It is also important to validate the classifier in a realistic setting by collecting longitudinal data. To this end, we intend to deploy our prototype system to the public. However, this brings additional challenges, such as how to identify the ground truth of the actual search task. We are investigating approaches how to unobtrusively obtain this data from the user.

ACKNOWLEDGEMENTS

This work was partly supported by The Finnish Funding Agency for Innovation (under projects Re:Know and D2I) and by the Academy of Finland (under the Finnish Centre of Excellence in Computational Inference).

REFERENCES

1. Adä, I., and Berthold, M. R. Eve: a framework for event detection. *Evolving systems 4*, 1 (2013), 61–70.

2. Athukorala, K., Głowacka, D., Oulasvirta, A., Vreeken, J., and Jacucci, G. Is exploratory search different? a comparison of information search behavior for exploratory and lookup tasks. *JASIST* (2015).

3. Athukorala, K., Hoggan, E., Lehtiö, A., Ruotsalo, T., and Jacucci, G. Information-seeking behaviors of computer scientists: Challenges for electronic literature search tools. In *ASIST* (2012).

4. Athukorala, K., Medlar, A., Ilves, K., and Głowacka, D. Balancing exploration and exploitation: Empirical parameterization of exploratory search systems. In *CIKM* (2015).

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

6. Auer, P. Using confidence bounds for exploitation – exploration trade-offs. *JMLR 3* (2002), 397 – 422.

7. Baldonado, M. Q. W., and Winograd, T. Sensemaker: an information-exploration interface supporting the contextual evolution of a user's interests. In *CHI*, ACM (1997), 11–18.

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

9. Bifet, A., and Frank, E. Sentiment knowledge discovery in twitter streaming data. In *Discovery Science* (2010).

10. Bifet, A., and Gavalda, R. Learning from time-changing data with adaptive windowing. In *Proc. of SDM* (2007).

11. Bouchachia, A. Incremental learning with multi-level adaptation. *Neurocomputing 74*, 11 (2011), 1785–1799.

12. Carmona-Cejudo, J. M., Baena-García, M., del Campo-Ávila, J., Morales-Bueno, R., and Bifet, A. Gnusmail: Open framework for on-line email classification. In *Proc. of ECAI* (2011).

13. Chau, D. H., Kittur, A., Hong, J. I., and Faloutsos, C. Apolo: making sense of large network data by combining rich user interaction and machine learning. In *CHI* (2011), 167–176.

14. Chowdhury, S., Gibb, F., and Landoni, M. Uncertainty in information seeking and retrieval: A study in an academic environment. *Information Processing & Management 47*, 2 (2011), 157–175.

15. Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association 74*, 368 (1979), 829–836.

16. Cole, M. J., Gwizdka, J., Liu, C., Belkin, N. J., and Zhang, X. Inferring user knowledge level from eye movement patterns. *Inf. Proc. Manag.* (2012).

17. Diriye, A. M. *Search interfaces for known-item and exploratory search tasks*. PhD thesis, UCL (University College London), 2012.

18. Feild, H. A., Allan, J., and Jones, R. Predicting searcher frustration. In *SIGIR* (2010).

19. Gama, J. a., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. A survey on concept drift adaptation. *ACM Comput. Surv. 46*, 4 (2014), 44:1–44:37.

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

21. Gomes, J. B., Menasalvas, E., and Sousa, P. A. Learning recurring concepts from data streams with a context-aware ensemble. In *Symposium on applied computing* (2011).

22. Hassan, A., and White, R. W. Personalized models of search satisfaction. In *CIKM* (2013), 2009–2018.

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

24. Hearst, M. *Search user interfaces*. Cambridge University Press, 2009.

25. Jansen, B. J., and Pooch, U. A review of web searching studies and a framework for future research. *JASIST 52*, 3 (2001), 235–246.

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

27. Käki, M. Findex: search result categories help users when document ranking fails. In *CHI*, ACM (2005), 131–140.

28. Kane, M. J., and Engle, R. W. Working-memory capacity, proactive interference, and divided attention: limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition 26*, 2 (2000), 336.

29. Karimzadehgan, M., and Zhai, C. Exploration exploitation tradeoff in interactive relevance feedback. In *CIKM* (2010), 1397–1400.

30. Kelly, D., and Fu, X. Elicitation of term relevance feedback: an investigation of term source and context. In *SIGIR* (2006), 453–460.

31. Kules, B., Wilson, M., Schraefel, M. C., Shneiderman, B., et al, and et al. From keyword search to exploration: How result visualization aids discovery on the web. Tech. Rep. HCIL-2008-06, University of Maryland, 2008.

32. Lindstrom, P., Delany, S. J., and Mac Namee, B. Handling concept drift in text data stream constrained by high labelling cost. In *Int. Florida Art. Intell. Research Society Conf.* (2010).

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

34. Matejka, J., Grossman, T., and Fitzmaurice, G. Citeology: visualizing paper genealogy. In *CHI Extended Abstracts*, ACM (2012), 181–190.

35. Pirolli, P., and Card, S. Information foraging. *Psych. rev. 106*, 4 (1999), 643.

36. Quinlan, J. *C4.5: Programs for Machine Learning*. Morgan-Kaufmann, Los Altos, California, 1993.

37. Radlinski, F., Kleinberg, R., and Joachims, T. Learning diverse rankings with multi-armed bandits. In *ICML* (2008), 784–791.

38. Sparck Jones, K., Walker, S., and Robertson, S. E. A probabilistic model of information retrieval: Development and comparative experiments. *Info. Proc. & Manag. 36*, 6 (2000), 779–840.

39. White, R. W., Bennett, P. N., and Dumais, S. T. Predicting short-term interests using activity-based search context. In *CIKM* (2010).

40. White, R. W., Kules, B., and Drucker, S. M. Supporting exploratory search. *Comm. ACM 49*, 4 (2006), 36–39.

41. White, R. W., and Roth, R. A. Exploratory search: Beyond the query-response paradigm. *Synth. Lec. on Inf. Conc., Retr., and Serv. 1*, 1 (2009), 1–98.

42. Yee, K.-P., Swearingen, K., Li, K., and Hearst, M. Faceted metadata for image search and browsing. In *CHI* (2003).