Narrow or Broad?
Estimating Subjective Specificity in Exploratory Search

Kumaripaba Athukorala° Antti Oulasvirta•⋆ Dorota Głowacka° Jilles Vreeken• Giulio Jacucci°

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

• Max Planck Institute for Informatics and Saarland University, ⋆ Aalto University

° first.last@cs.helsinki.fi, • {oantti,jilles}@mpi-inf.mpg.de

ABSTRACT

Supporting exploratory search is a very challenging problem, not least because of the dynamic nature of the exercise: both the knowledge and interests of the user are subject to constant change. Moreover, whether the results for a query are informative is strongly subjective. What is informative to one user, is too specific for the other; specificity differs between users depending on their intent and accumulated knowledge about the domain.

We propose a formal model—motivated by Information Foraging Theory—for predicting the subjective specificity of search results based on simple observables such as result-clicks. Through two studies including both controlled and free-form exploratory search we show our model allows us to differentiate between levels of subjective result specificity with regard to the current information need of the user.

Categories and Subject Descriptors

H.1 [**Models and Principles**]: User/Machine Systems—*Human information processing*

Keywords

Exploratory search; models of search behavior; click data; subjective specificity; Information Foraging Theory.

1. INTRODUCTION

Search tasks are commonly divided into two broad types: navigational, or known-item search, and exploratory search. In the former the user has specific search results in mind, while in the latter the problem is open ended, the user does not yet know exactly what she wants to find, and her goals may change as the search progresses [19, 36]. Traditional information retrieval (IR) techniques concentrate mostly on known-item search. Exploratory search is less well-studied, even though it is rapidly gaining importance as more and more knowledge is available through the web and knowledge bases [25]. While exploratory search is naturally challenging for users, at the same time it is also rather difficult for IR systems to offer support: search goals are poorly defined, users lack knowledge to formulate precise queries, user knowledge, search goals and information needs can all change throughout the search process [25, 36]. Recent years show an increased interest in techniques to support exploratory search, such as novel user interfaces [24], retrieval techniques [2, 6, 14], and studies of exploratory search [3, 24, 25]. One of the key open problems is that we need a better understanding of the dynamic nature of the user's information needs in exploratory search. In this paper we formalize a model that allows us to estimate the specificity of search results with respect to the user's *subjective* information need.

Exploratory search starts when a user has an interest in finding information on a topic in which she has little or no knowledge [36]. Generally, the user starts with vague queries using broad search terms, which allows them to obtain cues about new keywords and repetitively reformulate queries with specific terms [36]. Formulating good queries, however, is difficult—as is reformulating queries when the results are not satisfactory. When users try out queries exploratorily, some queries will return results that are overly specific with regard to the knowledge of the user by going into far too much detail. Alternatively, results can also be too broad, covering so many sub-topics that it is difficult for the user to get an overview.

For example, consider an undergraduate who has just started a course on data mining issuing "data mining" as her first query to explore this domain. Data mining is a broad subject and so the search results might cover a diverse information scope, which might make the results too *broad* for the user. Later this user might obtain cues about a new keyword "subgroup discovery" and formulate a new query. If the user has gained sufficient knowledge on this topic, then the search results might be just right for her. If not, however, then the search results may contain very specific technical details that are not comprehensible for a novice and so the user might find the results too *narrow*.

We are interested in modeling the specificity of search results with respect to the user's information need, which we refer to as *subjective result specificity*, here on we use "subjective specificity" as a short hand. We envision IR systems that *automatically* detect the subjective specificity of a result so that they can effectively support the user's exploratory search. In particular, depending on the subjective specificity, users will benefit from different types of support. For example, if the results are too broad, then visualizations of the information space and guided tours would help the

CIKM'14, November 3–7, 2014, Shanghai, China.
Copyright 2014 ACM 978-1-4503-2598-1/14/11$15.00.
http://dx.doi.org/10.1145/2661829.2661904 .

user to understand the new domain [14, 16]. If the results are too narrow, users might prefer introductory material explaining the new concepts, such as Wikipedia articles, or literature reviews [3]. Subjective specificity detection is also useful for IR systems supporting exploratory search through techniques such as results clustering, keyword suggestion or query expansion to determine whether the generated results are too broad or narrow for the user. To the best of our knowledge, there is no prior work for estimating subjective specificity in exploratory search.

We formalize a model that allows an IR system to infer subjective specificity from easily observable aspects of user behavior. That is, our model relies only on implicit click data, and we do not require any extra sensors such as eye-trackers [7]. Further, our model is sensitive in a predictable manner to moderating factors such as prior search experience and in-session learning.

The model captures how *information gain* [27] in exploration behavior is affected by the subjective specificity. We define information gain as the number of search results that a user *clicks* expressed as a function of the number of search results *seen* by the user. We assume that the information gain follows a natural logarithmic distribution. We adapt the formalism of Information Foraging Theory (IFT) [27] to predict how the slope of the information gain curve, the rate at which users click results, changes when subjective specificity becomes low (broad) or high (narrow) with respect to the user-specific reference curve.

The key idea is that the same search result can have very different information content for a user depending on how well it matches her current information needs. Consider the user in our previous example with two search queries: "data mining" and "subgroup discovery". The first query would retrieve broad results that include information about many areas, inviting the user to explore further. Consequently, the user would spend more time on every item [33], hence a higher slope of the curve. The second query would retrieve too narrow results with overly specific titles that make little sense to a novice, so she would probably estimate only few items as informative and worthy of further exploration, making the slope of the information gain curve shallow.

To evaluate our model, we design two experiments that capture the key elements of exploratory search. We focus on information-gathering [19] in the context of scientific essay writing—but note that our modeling approach is suited for other tasks as well. In the first study, we let computer science students explore scientific information to gather material for an essay on a domain they are not familiar with. We varied the subjective specificity at three levels (broad, intermediate, and narrow) and the order in which these appeared in a search session. It is our belief that this is the first study to manipulate the subjective specificity in a search session consisting of several searches on the same topic. In the second study, we consider the natural setting of free-form exploratory search, having users explore a topic of their interest. Empirical evaluation shows that our model estimates subjective specificity well in both settings.

In summary, our main contributions are: 1. a formal model to predict subjective specificity based on user behavior; 2. extensive empirical validation of the model; including 3. examination of moderating factors, such as prior search experience and in-session learning.

2. RELATED WORK

In recent years, exploratory search has attracted attention from, among others, IR, HCI, and cognitive science research communities. Below, we review contributions of these communities to understand user behavior, and develop retrieval techniques and user models to support exploratory search.

2.1 Studies of Exploratory Search Behavior

We briefly review studies on exploratory search to understand how user strategies affect observable exploratory search behavior. Overall, they clearly point to the dynamic nature of the exploratory search process which motivates our model. According to prior studies, exploration begins with concept formulation and then it narrows down to more specific concepts [36, 25], and as domain knowledge changes so do search tactics [37]. Previous studies provide detailed descriptions of exploratory search strategies such as narrowing and broadening search queries [31, 30, 33], users spending more time evaluating unfamiliar topics than familiar ones [20], change in search behavior with increasing domain knowledge [34]. Prior studies show that people with exploratory information needs are inclined to click more results following a query [36]. Literature suggests that users browse many results at the beginning of complex search tasks but they become selective when search goals become clearer [33].

2.2 Techniques to Support Exploratory Search

IR and machine learning communities propose several techniques to facilitate exploratory search. Some of the initial solutions include result clustering [8], relevance feedback [22], and faceted search [40]. However, these techniques are rarely used in practice, perhaps due to the high additional cognitive load of providing feedback for a large number of items [22]. In response, new techniques were designed to visualize search results and engage the user into the feedback loop. Some of them include interactive visualizations combined with learning algorithms to support users to comprehend the search results [6], and visualization and summaries of results [24]. These solutions give users more control, however, they do not adapt to the moment-by-moment information-needs of the user [32]. Recently, reinforcement learning (RL) techniques have been used to facilitate exploratory search [14, 21, 29]. These systems look promising, however, the modeling process can take a few iterations while the user has to deal with suboptimal results. With the help of our model, RL-based and other adaptive exploratory search systems could improve their performance.

2.3 Models of Information-Seeking

Many models of information-seeking have been designed to disambiguate user behaviors in known-item search, including search satisfaction [11, 17], frustration [9], and struggling [18]. Other related work includes models using eye-gaze [7] to predict the domain knowledge of the searcher, however, such models require extra sensors such as eye-trackers and are not sensitive to in-session knowledge gain and changes in user interests [37]. There exist probabilistic models for predicting the next interactions of the searcher [9], disambiguating short-term search interests [35], and estimating the relevance of results to information need [1]. Prior work suggests that there is a positive correlation between the information need specificity and query length [26]. However, in exploratory search users may issue queries of varying

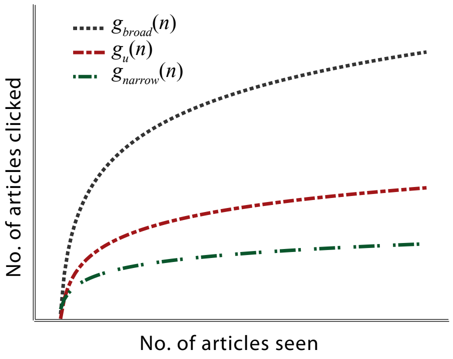


Figure 1: Hypothetical example of information gain as a function of the number of articles seen, and clicked to be explored in closer detail (Seen–Clicked). $g_u(n)$ is the user-specific effective information gain function. Our model predicts that the gradient of this Seen–Clicked curve increases (e.g., $g_{broad}(n)$) when results become more broad, and the gradient decreases (e.g., $g_{narrow}(n)$) when the results become more specific than the current information need of the user.

lengths with little understanding. This work shows the importance of user models in improving IR systems. Despite the benefits, the development of user models for exploratory search has seen little research attention.

Exploratory information-seeking is related to Information Foraging Theory (IFT) [27]. IFT includes several quantitative models of user search. The key idea is that decisions on what to do are made according to the expectation of information gain. As a user is searching and learning more about the content, she is continuously updating "information scent", i.e. her estimate of information gain by selecting a particular item. Information scent, in turn, affects whether to investigate an element or not. The theory makes predictions on how information gain, expressed as a function of time, changes with interface design [8]. When search results are unordered, information gain is a linear function of time. When they are ordered, it shifts to a diminishing returns curve. IFT has been used to explain how presentation techniques, such as result clustering, change the information gain rates and when is the optimal time to stop searching. More recently [13] used information scent to predict rankings of links. Existing work on IFT, however, does not consider the effect of specificity of search results and user-specific parameters, such as background knowledge. While IFT shows the basic shape of the gain function, it does not give a mathematical formulation that we can use in an IR system.

Berry-picking [5] is another human-centered model that assumes search is a constantly evolving phenomenon with the user constantly updating her cognitive model of the information being searched for. Like IFT, this model lacks a quantifiable formal approach that can be easily applied in an IR system. We contribute to IFT by providing a formal model that allows us to predict subjective specificity in exploratory search thus allowing IFT to be incorporated into an IR system.

3. MODEL OVERVIEW

Our goal is to predict the effect of subjective specificity on exploratory information-seeking where multiple search engine result pages (SERPs) are examined. We aim to capture the iterative and evolving nature of search, i.e. as the user explores a new domain, the search results become narrower and user knowledge expands.

Our model links two observable aspects of user exploratory behavior into what we call *effective information gain*, i.e. 1. the number of search results *seen* on a list; 2. number of search results *clicked*. By *clicked* we mean the action of opening a link to a search result for further investigation. We introduce a formal model for the effective information gain, (g), curve of a user, (u), as a function of the number of result items seen, (n), as shown, $g_u(n)$, in Figure 1. We refer to this graph as the *Seen–Clicked* curve. Any gain function is affected by the objective relevance of the search results. In our case, when results are ranked according to relevance, the function takes the shape of a diminishing returns curve.

We describe the relationship as a logarithmic regression model parameterized by λ and α:

$$g_u(n) = \lambda \ln(n) - \alpha \tag{1}$$

where n is the number of items seen so far on a result list which is a positive integer with no upper bound and λ determines the slope of this curve. α is a case-specific term which affects the maximum gain—it is determined by several factors; subjective specificity and case-specific factors such as the search task, and the maximum number of search results the user is expecting to gain. We make an assumption that when n is one item, α is -1 if the subjective specificity is broad and 0 otherwise. However, in reality n will be greater than one. We used a logarithmic function to capture the information gain, (g), as this is the most natural foraging distribution [15] commonly used in human behavior models [10]. In our model, we focus on the gradient of the gain function, λ, which is dependent on two parameters:

λ_u**:** user-specific factor

λ_r**:** results–specificity factor

The user-specific factor, λ_u, may depend on the user's experience with exploratory information-seeking or the search tool. For every user of a search tool a distinct Seen–Clicked curve is defined by λ_u. The $g_u(n)$ curve plotted in Figure 1 shows an instance of such a Seen–Clicked graph.

The results–specificity factor, λ_r, determines the effect of the subjective specificity on the gradient of the curve. For search results with high subjective specificity, *narrow*, the gradient of the curve reduces to a new effective gain function, as shown in graph $g_{narrow}(n)$ in Figure 1. An instance of a Seen–Clicked curve for search results that have low subjective specificity, *broad*, is shown in $g_{broad}(n)$. Although a single click carries little information about subjective specificity, our empirical data show that aggregated clicking behavior on a result page suffices for distinguishing among three levels (broad, intermediate, narrow).

An IR system would monitor the clicking and viewing actions of a user in a session. It would derive λ_u from the user's previous session, and throughout a given session it would derive λ_r from the user actions. Thus, the gain function in Equation 1 can be a combination of λ_u and λ_r:

$$g_u(n) = \lambda_r \lambda_u \ln(n) - \alpha. \tag{2}$$

A parameterized model predicts the subjective specificity of SERPs for the user and then compares the gradient of the

Seen–Clicked graph based on the user's clicks in a current query with that of the user's baseline Seen–Clicked graph. Such a baseline graph can be constructed by observing the everyday interactions of a user with a search tool. Then, if this user formulates a particular query to explore a research topic, the gradient of the new Seen–Clicked graph can be compared against the gradient of her baseline graph, and so the system can predict whether the search results are too narrow or too broad for her information-need—and adjust the behaviour of the system accordingly.

4. STUDY 1: CONTROLLED QUERY

The purpose of this study is to validate our model in a controlled setting as well as explore how prior experience and in-session learning affect the model. To this end, we designed the study by manipulating subjective specificity at three levels and permuting them over a session consisting of three result pages.

4.1 Pre-Study: User Observations

Prior to Study 1, we observed the information seeking behavior of computer scientists in order to understand their natural exploratory search behaviors. Our sample included two participants from each category: PhD students, post-doctoral and senior researchers. We asked them to inform us when they are exploring literature for a real need. We then visited their workstations and uninterruptedly observed and video recorded their search process. Based on these observations and prior work [4], we identified a common search strategy in exploratory search which initiates with query formulation, followed by scanning the SERP while clicking links of results that seem interesting. According to our observations, users process the clicked links after scanning the SERP. This information helped us to plan the design parameters of the data collection process.

4.2 Participants

We recruited 24 university-based computer science researchers who were not overly familiar with the topics of the search tasks. We selected computer science researchers as these generally have much experience with electronic literature search tools [3]. In order to explore the influence of prior experience on our model, we selected participants with varying levels of experience, that is, MSc and PhD students. Ten of the participants were in the process of writing their master's thesis and 14 were PhD students. Nine of the participants were female and 15 were male. The average age of the participants was 26.7 years, with the minimum age being 23 and the maximum 37. With a pre-study questionnaire we quantified their experience with scientific information-seeking (*mean* ± *std.dev*) [PhD students (5.6 ± 1.2), MSc students (5 ± 1.7)], frequency of exploratory search [PhD students (4.0 ± 1.03), MSc students (3.7 ± 1.6)] (ratings are given in a 7 point Likert scale where 1 ="not at all familiar/never" and 7 = "very familiar/often"). Table 1 reports the search topics and participant's familiarity with them.

Table 1: Tasks and queries used in the study. (B = Broad, I = Intermediate, N = Narrow). Familiarity with search topics is rated in a 7-point Likert scale and mean ± standard deviation of the participants' familiarity with each topic is given below the topic.

Topic, Familiarity		**Query**
Clustering (3.66 ± 1.57)	B	Clustering
	I	Density-based clustering
	N	Subspace clustering
Data mining (2.96 ± 1.46)	B	Data mining
	I	Pattern mining
	N	Subgroup discovery
Data privacy (1.75 ± 1.03)	B	Database privacy
	I	Differential privacy
	N	Differential privacy with continual leakage
Encryption (2.38 ± 1.21)	B	Encryption
	I	Identity-based encryption
	N	Certificateless encryption
Ergonomics (1.67 ± 1.09)	B	Ergonomics
	I	Task ergonomics
	N	EMG in ergonomics studies
Security (2.45 ± 1.06)	B	Computer Security
	I	Computer viruses
	N	Stuxnet

4.3 Design, Tasks, Materials, and Procedure

The study involved performing exploratory search on different topics. Every task involved going through three article lists generated from three queries that retrieved results with varying specificity: *broad* (B), *intermediate* (I) and *narrow* (N). The broad results covered a wide information scope; the intermediate ones a sub-field of the broad topic; and the narrow ones a very specific topic. To explore how in-session learning affects the model, we altered the order of presenting the results, which resulted in six permutations: Broad followed by Intermediate followed by Narrow (or, *BIN* for short), and likewise *BNI*, *INB*, *IBN*, *NIB* and *NBI*.

In order to cover all the six permutations, we created six unique tasks for six different topics. We asked senior researchers from these six computer science disciplines to define a task on their topic of expertise consisting of three search queries to retrieve results of varying specificity for a novice information-seeker in that domain. The experts also analyzed Google Scholar results for each query to ensure that they complied with the subjective specificity. The search topics and the queries are given in Table 1. For the purpose of counterbalancing, we randomised the order of the tasks and the query permutation for each participant.

4.3.1 Tasks

The tasks were defined in accordance with a task template designed to situate the participants in a scientific essay writing scenario, which is most suitable for creating exploratory search tasks [38]. To preserve consistency among the tasks, all the task descriptions followed the same template. Note that we refer to the results of one query as a list of articles:

> *"Imagine that you are writing a scientific essay about topic X. We provide you with three lists of articles that we have retrieved using three different queries. Go through each list in the order we give you and tick articles that you are interested in further reading to consider in your essay. Follow your natural scientific literature review-*

ing style when scanning the article lists. You have three minutes to go through each list. We will inform you when the three minutes are over and then you can move on to the next list."

4.3.2 *Materials and User Interface*

Google Scholar is the most commonly used literature search tool by computer scientists [3]. We used Google Scholar to retrieve 100 articles per query (from 10 SERPs), ranked according to the relevance for that query. As each task consisted of three queries, we retrieved 300 articles in total per task (100 articles/query × 3 queries/task). According to user observations (see Section 4.1), searchers decide to click on a result based on the Google Scholar information snippet, therefore we extracted all the primary information provided with each result item in Google Scholar.

Prior work suggests that search queries affect the user perception of search results [26], hence the query display could prime the search behavior. However, users do not always see the actual search query in systems that support exploratory search through techniques such as query expansion. Therefore, to avoid the influence of the search query on the search behavior, we only displayed results of the query and not the query itself. Participants could see the results retrieved for one query at a time. Once they completed scanning and ticking interesting articles from one list, then the list of next 100 articles for the next query was displayed. We provided a tick-box on the left side of every article and the participants could tick the articles that they were interested in. We informed the participants that ticking an article is analogous to clicking the URL and opening an article.

4.3.3 *Measures and Procedure*

We conducted the experiment on a desktop computer in a controlled room. We first gave the participants the printed task description. Next, we provided the first list of articles. We logged all the articles that the participants ticked and the time. While the participants were performing the tasks, we logged their gaze distribution over the articles to corroborate the number of articles seen before clicking an article. We instructed the participants to think aloud while performing the tasks and we used a voice recorder to record their thinking aloud. A pilot study showed it takes approximately three minutes to examine one list of articles without getting overly exhausted. Hence, the participants were given three minutes to go through one list. We used a timer and informed the participants when the three minutes had passed.

4.4 Results of Study 1

Every participant performed six search tasks and each search task involved searching through three lists of articles, therefore we obtained data from 432 search sessions (3 results lists × 6 search tasks × 24 participants). All together there were 4,414 click actions. We used all data without removing any outliers to keep the prediction task realistic.

4.4.1 *Subjective Specificity*

According to our model, the gradients of the Seen–Clicked curves should decrease with the increase in the subjective specificity (or narrowness of the results), and they should follow a natural logarithmic distribution. In order to confirm this, we analysed the overall distribution of the user information gain over information seen for the three types of results.

Table 2: Logarithmic regression models and model fit (R^2) for number of articles Seen–Clicked. Breakdown per Broad, Intermediate and Narrow search results.

Results Type	**Model**	**Fit (R^2)**
Broad	$3.83\ln(n) - 3.59$	0.97
Intermediate	$2.40\ln(n) - 2.06$	0.97
Narrow	$2.05\ln(n) - 1.96$	0.97

Table 3: The Wilcoxon signed-ranked test on (left) the gradients between the models and (right) the case-specific term, α, for Broad (B), Intermediate (I), and Narrow (N) results.

	Gradient λ			**Alpha α**		
Results	**Z**	**p-val**	**r**	**Z**	**p-val**	**r**
B & I	−4.20	<.001	−.60	−3.77	<.001	−.54
B & N	−4.29	<.001	−.62	−3.60	<.001	−.52
I & N	−2.71	<.01	−.39	−.057	.954	−.01

Figure 2a shows the overall number of articles Seen–Clicked averaged over all the participants over the three types of search results. As our model predicts, the gradient of the Seen–Clicked curve decreases as the results become narrower for the user's information need.

Next, we constructed gain curves for the three types of results for each participant averaging over the six tasks they performed. Using logarithmic regression, we calculated the model including the gradient (λ) and the case-specific term (α) of predicted curves for every participant (Section 3). Table 2 provides the summary of the prediction models of the three types of results and the model fit, R^2, calculated for the overall click data.

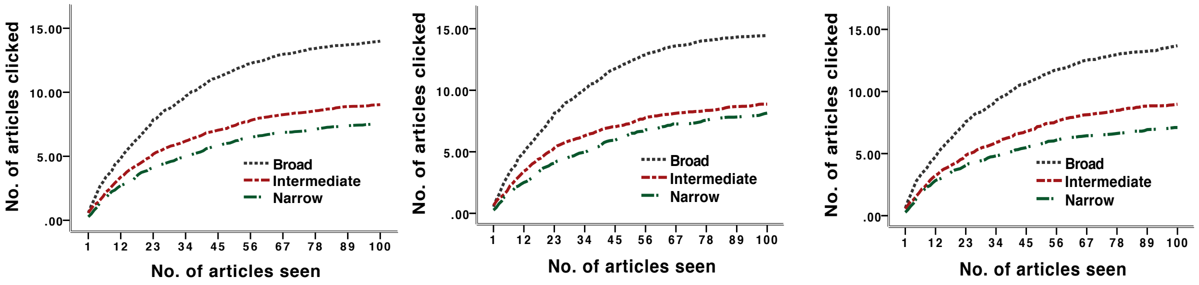
We used Wilcoxon signed-ranked test to statistically compare the gradients of the predicted models of each type of results. Gradients of the broad results (*median* 3.56) were significantly greater than those of intermediate (3.08) and narrow results (2.04). The gradients of the predicted models of the intermediate results were significantly greater than that of narrow results.

To see whether, and how, subjective specificity affects the case-specific term α, we conducted a Wilcoxon signed-ranked test. For broad results, the values for α (*median* 2.41) were significantly greater than for either intermediate (1.16) or narrow results (1.69). However, the difference was not significant between intermediate and narrow results, suggesting that the case-specific term is not as sensitive to subjective specificity as the gradient. Table 3 provides test results for both λ and α.

To summarize, the results confirm that when the subjective specificity increases, the gradient of the Seen–Clicked curve decreases. This suggests that the effective information gain reduces with an increase in narrowness of the results, which is validated by our model.

4.4.2 *Order Effects*

Since the participants went consecutively through the three types of article lists, it is necessary to ensure that the order of the articles lists has no effect on the information gain



(a) Averaged over all tasks (b) Results Order: Narrow after Broad (c) Results Order: Narrow before Broad

Figure 2: Seen–Clicked curves (2a) constructed by averaging over all the tasks performed by all the users with the broad, intermediate, and narrow search results in Study 1. Note the difference between Seen–Clicked curves of broad and narrow results is already visible within first 10 result items. Second and third images from the left, show Seen–Clicked curves for Broad, Intermediate and Narrow results w.r.t. whether Broad results where considered before (2b) or after (2c) the Narrow results. Note the difference in gradients for the Narrow results.

curves in Figure 2. To this end, we compared the number ($mean \pm std.dev$) of articles all the users clicked from the first (6.1 ± 2.6), second (5.6 ± 1.9), and third (6.0 ± 2.1) result lists according to the order they were presented. The results show that the order of results presentation has no effect on the number of articles clicked. In order to validate this statistically, we performed a Friedman test on the average number of articles participants clicked from the first, second, and third article lists. It shows there are no significant differences between the number of articles clicked in each list ($p = .717$).

4.4.3 *Prior Experience*

To understand the effect of prior experience on our model, we compared the gradients of the models predicted for the results with broad, intermediate, and narrow subjective specificity between participants with different levels of experience. The pre-study questionnaire confirms that PhD students have more experience than MSc students in scientific information-seeking and exploratory search (Section 4.2). According to our model, we expect the prior experience of the participants to affect the gradients of the Seen–Clicked graphs. As expected, Figure 3 shows that the gradients of the Seen–Clicked graphs for the participants with lower level of experience (MSc students) is higher than that of the more experienced participants (PhD students).

Table 4: Results of the Mann-Whitney's U test for MSc and PhD students comparing gradients of predicted models of the broad, intermediate, and narrow results. The Seen–Clicked curves for MSc students show significantly steeper gradients for all three types of results.

Results	**U**	**Z**	**p-value**	**r** (effect size)
Broad	36	−1.99	< .05	−.41
Intermediate	33	−2.17	< .05	−.44
Narrow	26	−2.58	< .05	−.53

We used Mann-Whitney's U test to compare the gradients of the models predicted between the two groups. The results are summarized in Table 4. The gradients of the predicted models of broad results of MSc students (*median* 4.48) were significantly steeper than that for the PhD students (2.94). The gradients of the predicted models for results with intermediate subjective specificity of MSc students (3.08) were also significantly greater than that of PhD students (1.67). Similarly, the gradients of the predicted models of the narrow results of the MSc students (2.46) were significantly greater than that of the PhD students (1.49).

Table 5: Correlation analysis between the model gradients for Broad, Intermediate and Narrow results with respect to user familiarity with scientific information-seeking (left), and how often they explore unfamiliar research topics (right). ($N = 24$)

	Familiarity		**Frequency**	
Query	r_τ	**p-value**	r_τ	**p-value**
Broad	−.35	< .05	−.46	< .01
Intermediate	−.49	< .01	−.47	< .01
Narrow	−.52	< .01	−.39	< .05

The results clearly show that the participants with lower level of experience in scientific information-seeking (MSc students) click more results indicating a lower subjective specificity for all the three types of results than the more experienced participants (PhD students). In order to understand this behavior we analyzed the correlation between prior experience and gradients of the models, and think-aloud recordings.

The correlation analysis suggests that the gradients for Seen–Clicked curves are lower for the users with more experience in scientific information-seeking and exploratory search (PhD students) (Table 5). According to the voice recordings, PhD students had more specific criteria for the type of articles they needed. For example, 11 PhD students explained that they were more interested in review articles than articles about a specific topic. They also distinguished scientific articles from books to refrain from clicking too many books and paid more attention to the publication year to avoid older articles. However, MSc students clicked all the articles that have relevant titles.

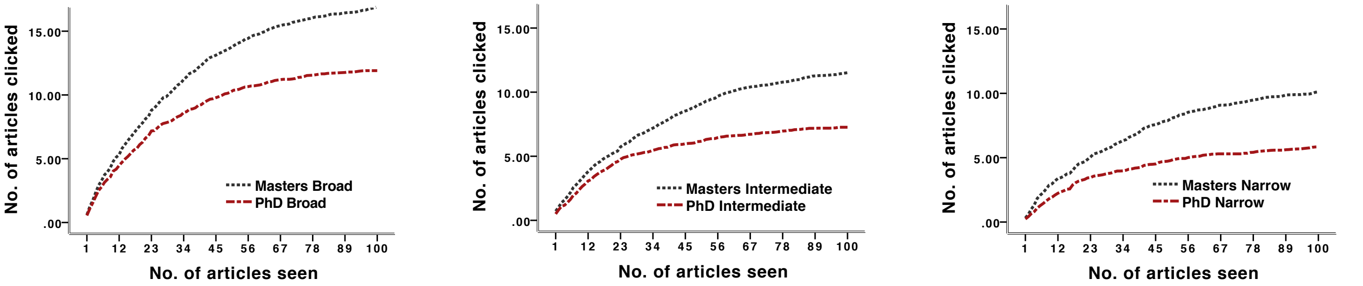


Figure 3: Seen–Clicked graphs for the broad, intermediate and narrow results for MSc and PhD students. MSc students have steeper gradients for all the three results compared to PhD students. Note the gradient of Seen–Clicked curve for narrow results of MSc students is at the same level as the gradient of Seen–Clicked curve for intermediate results of PhD students.

4.4.4 *In-Session Learning*

To investigate whether in-session learning have an effect on our model, we analyzed the Seen–Clicked graphs per permutation condition (Section 4.3). These permutations emulate transitions in results-specificity during a search session.

For each query permutation, Seen–Clicked graphs of the broad, intermediate and narrow results followed our model. As expected, when narrow results were considered after the broad ones, the gradients of Seen–Clicked graphs of the narrow results were greater than when narrow results were considered before broad results. A possible reason for this difference is an increase in user knowledge or in-session learning, which can be seen in Figure 2b and 2c. Table 6 shows the model prediction for these two scenarios. We can see that the gradient of the predicted model of the narrow results has increased from 1.9 to 2.3 when narrow is given after broad.

We conducted statistical test on the gradients of the models predicted for the three types of results given for the six tasks for every participant (3 queries × 6 tasks = 18 models/participant). The Friedman test shows no significant difference between the gradients for broad ($p = .051$) and intermediate curves ($p = .46$) among the six query permutations. However, the difference is significant for the gradients of the models predicted for narrow curves among the six query permutations ($\chi^2(5) = 19.4, p < .01$). In order to ensure this difference for narrow results was due to whether the broad results were before or after, we split the gradients predicted for narrow curves in to two groups; narrow results were presented before broad results (respectively *NIB*, *NBI*, *INB*) and vice-versa (*BIN*, *BNI*, *IBN*). A Friedman test showed no significant difference between the gradients of the narrow results in permutations *NIB*, *NBI*, and *INB* ($p = .86$). Similarly, the difference was not significant between *BIN*, *BNI*, and *IBN* permutations ($p = .95$).

These results suggest the gradients of the predicted models of narrow results change only if they are presented after the broad results. An explanation for this difference is that when results gradually become narrower, the user is likely to make better use of the narrow results than in the opposite direction. As a result of this behavior, when the narrow results are presented after the broad results, the number of articles clicked by the user increases and the effective information gain approaches that of the intermediate results. Further, the number of articles that overlap between the three lists is less than 4%. Therefore, we conclude that this difference is not due to the article overlap between results but rather a result of the learning effect.

Table 6: Logarithmic regression models and fit (R^2) for number of articles Seen–Clicked for broad, intermediate, and narrow results with regard to whether the Broad results were considered before (left) or after (right) the Narrow results.

	Broad $\succ$ Narrow		**Narrow $\succ$ Broad**	
Results	**Model**	R^2	**Model**	R^2
Broad	$4.1\ln(n){-}3.3$	0.97	$3.7\ln(n){-}3.9$	0.96
Intermediate	$2.3\ln(n){-}2.3$	0.96	$2.4\ln(n){-}1.8$	0.97
Narrow	$2.3\ln(n){-}1.5$	0.98	$1.9\ln(n){-}2.5$	0.95

Table 7: Cross-validation results for the broad, intermediate, and narrow models built for MSc and PhD students.

Results	**MSc (R^2)**	**PhD (R^2)**
Broad	0.71	0.88
Intermediate	0.86	0.95
Narrow	0.86	0.60

4.4.5 *Cross-validation*

In order to further validate the model, we report results obtained using leave-one-out cross-validation. Since the model is affected by the prior experience of the users, we split the data into two groups by experience (MSc and PhD). For each group we construct separate models per subjective specificity level, by leaving one participant out and fitting the model over the others. We then use this model to predict the Seen–Clicked curves for the left-out participant, calculating the model fit (R^2) with the actual Seen–Clicked curve of that participant. We iterate over all participants, and report the average R^2 per group. We obtained reasonably high R^2 values for both groups for the three subjective specificity levels as reported in Table 7.

4.4.6 *Classification*

Last, we perform a preliminary study evaluating the practical applicability of our model. To this end, we investigate how well the subjective specificity can be predicted based on the model gradient over the first 33 out of 100 articles. That is, we check whether the system can infer the subjective specificity *while* the user is still going over the full list, and can hence offer targeted assistance in doing so.

We use Weka [39] to train C4.5 decision trees [28] using 10-fold cross-validation. Despite our small training data, we already obtain 72.1% accuracy and an AUC of 0.687 when classifying between broad and narrow results. This means we beat the baseline, resp. 50% and 0.5 by a clear margin. When considering three classes of results, we obtain an accuracy of 48.1% and an AUC of 0.589 against a baseline of 33%, and 0.5, again a clear improvement. It is interesting to note that performance is stable between the first 33, 50, or all 100 articles. Given the stark differences in slopes seen (see, e.g., Fig. 3) it seems reasonable that with more training data and more advanced classifiers reliable calls can be made given only the first 10 or so articles.

4.5 Summary of Study 1 Findings

Overall, study 1 confirms that we can model information gain in exploratory search with a logarithmic function of the number of results seen by the user. It validates that the gradient of the Seen–Clicked curves decrease with an increase of the subjective specificity, hence we can estimate subjective specificity using our model. Further, the results suggest that our model is sensitive to both in-session learning and prior experience of the user. When the user has more experience with exploratory search and scientific information-seeking the gradient of the Seen–Clicked curve decreases, because she has a specific criteria for the type of information she needs. If a user gradually moves from broad to narrow results, then in-session knowledge gain would help the user to recognize more useful articles even from narrow results, increasing the gradient of the Seen–Clicked curve. These results suggest that our model could be used to predict when a user actually needs help with narrow results. Preliminary classification indicates the applicability of our model in a real IR system.

5. STUDY 2: FREE EXPLORATION

In order to validate our model in a more natural setting we conducted a second study involving ten computer science students exploring scientific articles for an actual information need. Participants of this study were not involved in our Study 1. Four were MSc students looking for scientific literature to include in their theses. The other participants have just finished their MScs and were exploring new research topics to prepare their PhD proposals. Google Scholar is the search tool they all use, therefore we implemented an interface similar to Google Scholar which enabled the participants to issue search queries and view results that we extracted from Google Scholar. We displayed 40 articles per page with same information as in Google result snippets and allowed every participant to conduct their natural exploration using our search interface for two hours. We did not impose any restrictions on the search process, and they could conduct search in the same way as with Google Scholar, i.e. click articles, read opened articles, and make notes. We logged their search queries, retrieved results, and clicked articles with time. We used experts in each search topic to assign the search results of every query in to one of the three categories: broad, intermediate, and narrow. The experts were either postdoctoral researchers or professors specializing in the search topic. Most of the experts (6/10) were supervisors of the participants and so had an idea about the level of knowledge of the participants' to predict the subjective specificity. To measure the quality of categorization, part of the assessments were conducted by two experts (6/10). We run Cohen Kappa test to measure the inter-annotator agreement between the experts. Kappa indicated a substantial agreement (Kappa = .67, p <.01).

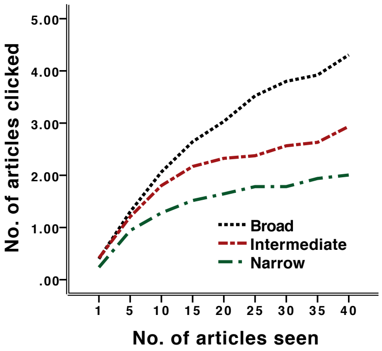


Figure 4: Seen–Clicked curves constructed by averaging over all the tasks performed by all the users with the broad, intermediate, and narrow search results in Study 2. Note that in Study 2, a SERP contained 40 articles unlike 100 articles in Study 1 which makes the Seen–Clicked curves of two studies slightly different.

5.1 Results of Study 2

In total all the participants have issued 142 search queries where 36% of them has retrieved broad search results, 37% has retrieved intermediate results, while 27% has retrieved narrow search results. All together there were 339 clicks.

We plotted the Seen–Clicked curves for broad, intermediate, and narrow results by taking the average over all the participants, as shown in Figure 4. As in Study 1, the gradients of the Seen–Clicked curves decrease predictably with increasing subjective specificity. We computed the model for each curve using logarithmic regression. As expected, the broad curve has the highest gradient ($\lambda = 1.20$) with model fit $R^2 = 0.96$, intermediate curve has the second highest gradient ($\lambda = 0.73, R^2 = 0.98$), while the narrow curve has the lowest gradient ($\lambda = 0.50, R^2 = 0.99$). To further confirm that the difference between the gradients of three curves are statistically significant we computed the model for individual Seen–Clicked curves for each participant and conducted Wilcoxon signed-ranked test. The test results indicate that predicted models of the Seen–Clicked curves of broad search results have a significantly higher gradient than that of both intermediate ($Z = -2.31$, p <.05) and narrow ($Z = -2.67$, p <.01) search results. However, the difference between the gradients of the predicted models of the intermediate and narrow search results was not significant ($Z = -1.836$, p =.06). We could expect the difference between the Seen–Clicked curves of intermediate and narrow search results to be small, because as the results of Study 1 indicate (see Section 4.4.4) users gain knowledge when they gradually transit from broad to narrow search results. Further, in Study 2 participants could browse through clicked articles, therefore they spent time reading them in addition to scanning results, which explains why the gradients of models in Study 2 are less than those in Study 1.

Overall, these results suggest that our model is applicable to natural exploratory search tasks.

5.1.1 Change Over Time

We also analyzed the percentage of search queries that retrieved broad, intermediate, and narrow search results with time to verify that search results become increasingly narrow over time. Most of the search queries (56%) that were issued within the first 20 to 40 minutes have retrieved broad results. Interestingly, some participants (10%) *started* their search session with queries retrieving already narrow results. After the first 1.5 hours, percentage of search queries that retrieve narrow results has increased to 42%. This confirms the common sense insight that during exploration over time users gradually narrow down their search queries using specific terms [36]. Our model helps an IR system to infer whether these broad or narrow search results correlate with what the user is actually expecting.

5.2 Summary of Study 2 Findings

Overall, study 2 confirms that our model can be used to predict subjective specificity in natural exploratory search tasks. The analysis of broad, intermediate, and narrow search results with time indicate that over time users issue narrower search queries. However, exploratory search may continue over several hours or even years and would involve offline learning through other media such as books, and social networks [25]. Therefore, it is not feasible to use time as a parameter to model subjective specificity. This study further verifies that the number of search results users click in exploratory search is affected by in-session learning. We postpone the classification experiments here, as these require us knowing the actual user information need for each query. Moreover, proper implementation requires a longitudinal study to build a reference model per user such that we can compare the Seen–Clicked curve against this reference model. This is beyond the scope of this paper.

6. DISCUSSION AND CONCLUSIONS

This paper has contributed a model for predicting the subjective specificity of search results. The model builds on earlier insights about exploratory search and Information Foraging Theory, assuming that for every individual there is an idiosyncratic baseline curve for information gain. Given this curve as a reference point, it predicts whether the current search results are too broad or narrow for the user's information need. We empirically validated this model in two studies which show that when search results become too narrow—or, high in subjective specificity,—the gradient of the Seen–Clicked graph decreases significantly.

We show that our model applies in both a controlled environment and in realistic open ended settings. Our classification results show that our model indeed captures valuable information about the subjective selectivity of results. Although the exercise is preliminary, the results are promising: ideally one would train over much more data, over more results, use a more advanced classifier, and, in particular, take timings between clicks into account. However, these results do tell us that our model could be employed within an IR system to quickly obtain an educated guess on how narrow/broad the current search results are with regard to the current state of the users' information seeking process—and adapt its behavior accordingly.

The model has valuable implications for exploratory search systems. For example, it has potential applications in systems that support exploratory search by making query suggestions [12], organizing information according to facets [40], directing search by predicting keywords [14], or providing visualizations and summaries of results [24]. These systems could use our model to predict whether the suggested results are broader or narrower than the information need of the user. Furthermore, our model could be used as a substitute for relevance feedback techniques that put the user through tedious feedback loops. Even though a hierarchical ontology such as Open Directory Project (ODP, www.dmoz.org) could be used to suggest whether a search query is referring to a broad/narrow topic, such an ontology cannot predict whether the search results are actually broad/narrow with respect to the current information need of the user. Our model could also be applied to reinforcement learning based solutions to predict the right balance between exploration and exploitation according to the subjective specificity [14]. For example, we could increase the level of exploration for novice researchers in a given field based on their current information need in order to expose them to a large area of the information space. On the other hand, we would decrease the level of exploration for more advanced researchers thus exposing them to narrower search results. This model could also be used by IR systems to real-time update search results and visualizations according to subjective specificity [23].

An important open challenge is to incorporate our model into a running IR system. Our preliminary classification study shows a system without extra sensors and using only a simple classifier can obtain informed estimates on the subjective specificity while the user is interacting with its results. By leveraging larger training data and more sophisticated classification/regression algorithms significant improvements can be expected—in particular when user click/view timings and history data are taken into account. In the future, we will collect longitudinal data from users to construct reference models and evaluate the predictive power of the model. The cross-validation results suggest we could build a common reference model for users with similar backgrounds, and hence we may be able to build reference models for a set of known background levels and apply it to new users without building individual reference models.

To conclude, our model is useful for the design of personalized exploratory search systems that adjust the search results according to the evolving information needs and knowledge of the user in a given topic.

7. ACKNOWLEDGEMENTS

A. Oulasvirta and J. Vreeken are supported by the Cluster of Excellence "Multimodal Computing and Interaction" within the Excellence Initiative of the German Federal Government. D. Głowacka was supported by The Finnish Funding Agency for Innovation (under projects Re:Know and D2I) and by the Academy of Finland (under the Finnish Centre of Excellence in Computational Inference). This work has been partly supported by MindSee (FP7 – ICT; Grant Agreement # 611570).

8. REFERENCES

[1] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *SIGIR*, pages 3–10, 2006.

[2] O. Alonso, R. Baeza-Yates, and M. Gertz. Exploratory search using timelines. In *CHI Workshop on Exploratory Search and HCI*, 2007.

[3] K. Athukorala, E. Hoggan, A. Lehtiö, T. Ruotsalo, and G. Jacucci. Information-seeking behaviors of computer scientists: Challenges for electronic literature search tools. In *ASIST*, 2013.

[4] F. Baskaya, H. Keskustalo, and K. Järvelin. Modeling behavioral factors in interactive information retrieval. In *CIKM*, pages 2297–2302, 2013.

[5] M. J. Bates. The design of browsing and berrypicking techniques for the online search interface. *Onl. Inf. Rev.*, 13(5):407–424, 1989.

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

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

[8] D. R. Cutting, D R Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *SIGIR*, pages 318–329, 1992.

[9] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *SIGIR*, 2010.

[10] P. M. Fitts and J. R. Peterson. Information capacity of discrete motor responses. *Journal of experimental psychology*, 67(2):103, 1964.

[11] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *TOIS*, 23(2):147–168, 2005.

[12] W.-T. Fu, T. G. Kannampallil, and R. Kang. Facilitating exploratory search by model-based navigational cues. In *IUI*, 2010.

[13] W.-T. Fu and P. Pirolli. Snif-act: A cognitive model of user navigation on the world wide web. *Hum.-Comp. Int.*, 22(4):355–412, 2007.

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

[15] R. L. Goldstone and B. C. Ashpole. Human foraging behavior in a virtual environment. *Psychonomic Bulletin & Review*, 11(3):508–514, 2004.

[16] A. Hassan and R. W. White. Task tours: helping users tackle complex search tasks. In *CIKM*, 2012.

[17] A. Hassan and R. W. White. Personalized models of search satisfaction. In *CIKM*, pages 2009–2018, 2013.

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

[19] M. Hearst. *Search user interfaces.* Cambridge University Press, 2009.

[20] I. Hsieh-Yee. Research on web search behavior. *Libr. Inf. Sci. Res.*, 23(2):167–185, 2001.

[21] M. Karimzadehgan and C. Zhai. Exploration exploitation tradeoff in interactive relevance feedback. In *CIKM*, pages 1397–1400, 2010.

[22] D. Kelly and X. Fu. Elicitation of term relevance feedback: an investigation of term source and context. In *SIGIR*, pages 453–460, 2006.

[23] J. Y. Kim, M. Cramer, J. Teevan, and D. Lagun. Understanding how people interact with web search results that change in real-time using implicit feedback. In *CIKM*, pages 2321–2326, 2013.

[24] B. Kules, M. Wilson, M. C. Schraefel, and B. Shneiderman. From keyword search to exploration: How result visualization aids discovery on the web. Technical Report HCIL-2008-06, U. Maryland, 2008.

[25] G. Marchionini. Exploratory search: from finding to understanding. *Comm. ACM*, 49(4):41–46, 2006.

[26] N. Phan, P. Bailey, and R. Wilkinson. Understanding the relationship of information need specificity to search query length. In *SIGIR*, 2007.

[27] P. Pirolli and S. Card. Information foraging. *Psych. rev.*, 106(4):643, 1999.

[28] J.R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan-Kaufmann, Los Altos, California, 1993.

[29] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *ICML*, pages 784–791, 2008.

[30] J-F. Rouet. *The skills of document use: From text comprehension to Web-based learning.* Routledge, 2013.

[31] A. G. Sutcliffe, M. Ennis, and S. J. Watkinson. Empirical studies of end-user information searching. *J. Assoc. Inf. Sci. Tech.*, 51(13):1211–1231, 2000.

[32] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *CHI*, pages 415–422, 2004.

[33] P. Vakkari. Task complexity, problem structure and information actions: integrating studies on information seeking and retrieval. *Information processing & management*, 35(6):819–837, 1999.

[34] P. Vakkari, M. Pennanen, and S. Serola. Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Inf. Proc. Manag.*, 39(3):445–463, 2003.

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

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

[37] B. M. Wildemuth. The effects of domain knowledge on search tactic formulation. *J. Assoc. Inf. Sci. Tech.*, 55(3):246–258, 2004.

[38] B. M Wildemuth and L. Freund. Assigning search tasks designed to elicit exploratory search behaviors. In *Symposium on HCI and IR*, 2012.

[39] I.H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, San Francisco, 2nd edition, 2005.

[40] K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In *CHI*, 2003.