

Characterizing the Influence of Domain Expertise on Web Search Behavior

Ryen W. White, Susan T. Dumais, and Jaime Teevan

Microsoft Research

One Microsoft Way

Redmond, WA 98052

{ryenw, sdumais, teevan}@microsoft.com

ABSTRACT

Domain experts search differently than people with little or no domain knowledge. Previous research suggests that domain experts employ different search strategies and are more successful in finding what they are looking for than non-experts. In this paper we present a large-scale, longitudinal, log-based analysis of the effect of domain expertise on web search behavior in four different domains (medicine, finance, law, and computer science). We characterize the nature of the queries, search sessions, web sites visited, and search success for users identified as experts and non-experts within these domains. Large-scale analysis of real-world interactions allows us to understand how expertise relates to vocabulary, resource use, and search task under more realistic search conditions than has been possible in previous small-scale studies. Building upon our analysis we develop a model to predict expertise based on search behavior, and describe how knowledge about domain expertise can be used to present better results and query suggestions to users and to help non-experts gain expertise.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search & Retrieval—*query formulation, search process, selection process.*

General Terms

Experimentation, Human Factors.

Keywords

Domain expertise, Web search.

1. INTRODUCTION

Web searchers differ from each other in many ways that can greatly influence their ability to carry out successful searches. One way in which they can differ is in their knowledge of a subject or topic area. Domain expertise is not the same as search expertise [22], as it concerns knowledge of the subject or topic of the information need, rather than knowledge of the search process. Domain expertise has been studied extensively in the information science community (see [23] for a review). Studies of domain expertise have highlighted several differences between experts and non-experts, including: site selection and sequencing [4], task completion time [3], vocabulary and search expression [2], the number and length of queries, and search effectiveness [24]. These studies involved small numbers of subjects with carefully controlled tasks, making it difficult to generalize their findings.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WSDM '09, February 9-12, 2009, Barcelona, Spain

Copyright 2009 ACM 978-1-60558-390-7...\$5.00.

In this paper we build on this previous research via a large-scale log-based study of web search behavior. We contrast the search strategies of domain experts with those of non-experts through analysis of naturalistic interaction log data over a three-month period of time. This large-scale analysis allows us to identify greater diversity in vocabulary, web site visits, and user tasks than is possible with smaller-scale laboratory studies. In addition we develop methods for identifying domain experts using online search interaction patterns rather than offline tests of expertise. We focus on four domains – medicine, finance, law, and computer science – with complex subject matter and a large potential benefit to non-experts in identifying effective search strategies.

In addition to highlighting differences in the search behavior of experts and non-experts, we describe the possible benefits of being able to identify domain experts and leverage their querying strategies and source selection abilities. Search tools currently provide the same experience to users regardless of expertise. A cardiologist searching for the latest research studies on heart disease gets the same search results for the query “heart disease” as a newly-diagnosed patient with little background in the area. We believe that by understanding how people’s domain expertise affects their search behavior, we can better support interactions at the appropriate level, and help non-experts gain expertise.

The remainder of this paper is structured as follows. Section 2 presents related work on domain expertise. Section 3 describes the search logs, and Section 4 the approach used to identify experts within them. In Section 5 we discuss differences and commonalities in the interaction behavior of domain experts and domain non-experts. Section 6 presents a classifier that can identify users, actions, and sessions as expert or non-expert based on observable behavior, and discusses how such a classifier can be used to improve the Web search experience for people of varying domain expertise. We conclude in Section 7.

2. RELATED WORK

Research on domain expertise has examined differences between experts and non-experts in three main classes of search behavior: query attributes (choice of search terms, query length and syntax), search strategies and tactics (resource selection, sequence of actions, mix of querying and browsing), and search outcomes (accuracy, time). Many of these studies were conducted in the context of library systems rather than the general web search, and involved small numbers of participants in laboratory settings.

Allen [2] examined the relationship between topic knowledge (in the area of Voyager explorations of Neptune) and search behavior in an early online library catalog. He found the searchers with high domain knowledge had greater familiarity with the vocabulary for the topic and found more items. Hsieh-Yee [13] found that library science students used more of their own search

terms on a library science topic, and used more terms suggested by external sources (thesaurus and synonym list) for a topic in which they had little domain knowledge. They also spent more time preparing queries and examining results when they had little domain knowledge. More recently, Hembrooke et al. [11] investigated the effects of domain knowledge on users' search term selection and reformulation strategies for web searches. They found that self-rated domain experts issued longer and more complex queries than novices. Further, experts used elaborations as a reformulation strategy more often compared with the simple stemming and backtracking modifications used by novices.

Wildemuth [23] examined the search behavior of medical students in microbiology. In this experiment, students were observed at three points of time (at the beginning of the course, at the end of the course, and six months after the course), under the assumption that domain expertise changes during a semester. Some search strategies, most notably the gradual narrowing of the results through iterative query modification, were the same throughout the observation period. Other strategies varied over time as individuals gained domain knowledge. Novices were less efficient in selecting concepts to include in search and less accurate in their tactics for modifying searches. Vakkari et al. [19] also examined students at multiple points in time, as they were developing their thesis proposal. One important change in behavior was the use of a more varied and more specific vocabulary as students learned more about their research topic. An important methodological issue with studies that examine changes in expertise of an individual over time is that individuals also acquire many other kinds of knowledge during that time (e.g., in Wildemuth's study students also acquired search knowledge of the database of microbiology facts that was used in the study), so it is difficult to isolate domain expertise from search expertise.

Bhavnani [3,4] conducted several studies examining the search strategies of domain experts and novices. In these studies, five healthcare experts and five shopping experts performed web search tasks in their domain of expertise as well as in the other domain. Important differences were identified in site selection and knowledge of goal sequencing. Domain experts knew about key resources for their domain and often went directly to these sites rather than starting with general web search engines. In addition, domain experts had a general strategy for performing tasks – e.g., in the shopping tasks experts visited sites with detailed product reviews, comparative shopping sites, and discount sites. Novices, in contrast, went to a general search engine and examined a few items in the results list, often terminating their session before identifying good sites or thorough information. Using these insights they developed a search system, Strategy Hubs, to provide support for novices to identify comprehensive information from high-quality sites in the medical domain [5]. Hölscher and Strube [12] examined both search and domain expertise. Search expertise was determined by interview and a pre-test, and domain experts were undergraduates in economics (the topic of the search task). They found that both variables affected search behavior. Individuals with both domain and search expertise accessed web sites directly, but others always used search engines. Search novices were more likely to reformulate their queries, especially when they were also domain novices. Search experts were more likely to use richer query syntax and shorter queries. Kelly and Cool [14] investigated the relationship between topic familiarity and information search behavior (in the form of document reading time and efficacy).

They showed that as searcher's topic familiarity increases, search efficacy increases and reading time decreases. They claim that it may be possible to infer topic familiarity from search behavior.

Marchionini et al. [15] compared the performance of search specialists, domain specialists, and novices (students) using a full-text hypertext system. Both types of experts were more successful than novices in completing search tasks. Zhang et al. [24] examined relationships between engineering domain knowledge, search behavior, and search success. Expertise was assessed by having undergraduate and graduate students rate their familiarity with two hundred terms from the Heat and Thermodynamics category in the Engineering Information Thesaurus. Experts found slightly more relevant documents. Experts issued more queries per task and longer queries, and their vocabulary overlapped somewhat more with thesaurus entries, although these differences were not reliable statistically. In a recent study, Duggan and Payne [9] looked at web search performance for individuals with varying expertise in the domains for music and football. Domain knowledge was assessed using answers to 30 simple fact-based questions. Search performance was assessed for these same questions when a web search engine was available to searchers. For the music domain, there was little effect of domain expertise, perhaps because of a relatively narrow range of expertise scores. For the football domain, several interesting associations with domain knowledge were observed. Expertise was positively correlated with search accuracy (even for questions that they did not know the answer to), and negatively correlated with time spent on web pages and mean query length. Experts could process pages related to their domain more quickly, which is to be expected. However, that experts issued shorter queries is inconsistent with previous work and not well explained by the authors. Query length may vary depending on the nature of the task (chosen by participants vs. assigned by experimenters) or on the content source (web vs. domain-specific resources).

Freund and Toms [10] reported some interesting differences between software engineering consultants performing work-task scenarios and general web search behavior. They found that the software engineers issued longer queries on average (4.4 words), and used technical terminology and acronyms in 66% of their searches. Although this work does not explicitly compare domain experts and novices, it does suggest that experts performing realistic work-related tasks exhibit different search behavior than is reported in general web log analyses (e.g., [17]).

As summarized above, previous research has shown differences in search queries, strategies, and search outcomes as a function of domain expertise. However, much of this research examined search behavior in controlled laboratory settings using small numbers of searchers, experimenter-specified tasks, and required the explicit measurement of domain expertise. Our research generalizes this along several dimensions, with the goal of developing methods that can be broadly applied to understanding and supporting domain experts in naturalistic task environments. We use a large-scale log analysis of web search behavior allowing us to identify greater diversity in search vocabulary, interaction patterns, and tasks than have previously been reported. We study experts in four different domains (medicine, finance, law, and computer science), enabling us to identify similarities and differences in interaction patterns across domains. And, we develop a model for predicting domain expertise based on these patterns (rather than requiring tests to assess domain knowledge), and describe how these predictions can be used at web scale.

3. DATA SOURCES

To perform this study, we examined the querying and browsing behavior of many searchers in the four domains of interest. We obtained fully-anonymized logs of URLs visited by users who opted in to provide data through a widely distributed browser toolbar. The information contained in these log entries includes a unique identifier for the user, a timestamp for each page view, a unique browser window identifier (to resolve ambiguities in determining which browser a page was viewed), and the URL of the web page visited. Intranet and secure (https) URL visits were excluded at the source. In order to remove variability caused by geographic and linguistic variation in search behavior, we include only entries generated in the English speaking United States ISO locale. The results described in this paper are based on a sample of URL visits during a three-month period from May 2007 to July 2007 inclusive, representing more than 10 billion URL visits from more than 500 thousand unique users.

From these logs we extracted around 900 million *browser trails* and around 90 million *search sessions*, as defined by [20]. Browser trails consist of a temporally-ordered sequence of URLs comprising all pages viewed per web browser instance or browser tab. Search sessions are a subset of these browser trails, which begin with a query to a search engine such as Google, Yahoo!, or Live Search, and terminate with a period of user inactivity of 30 or more minutes. This threshold has been used previously to demarcate search sessions in logs [8,20]. These sessions are used as the basis of the comparison between experts and non-experts. We compare the groups based on aspects of querying, navigation, overall search success, and changes in expertise, within topical areas. In the next section we describe in more detail how experts and non-experts were identified.

4. IDENTIFYING DOMAIN EXPERTS

We selected four domains for our investigation: medicine, finance, law, and computer science. We selected these domains because, in addition to being of general interest to the population at large, there are large professional groups in each area whose domain expert members commonly use the web as a source of information. The selection of computer science was advantageous since the authors are domain experts and could manually verify the validity of the queries issued and web sites visited.

In our analysis, we first identified a set of people in the logs that appeared interested in each of the four domains, regardless of their expertise. This ensured that all of the behavior studied related to people interested in the domain and helped control for topical differences. From this group of interested people we then separated experts from non-experts based on whether they visited specialist web sites. This simple yet broadly-applicable method for identifying expertise allows us to extend on a large-scale the understanding of interaction patterns previous researchers have developed in laboratory studies to real-world situations. We now describe how we identify domain experts in more detail.

4.1 Identifying Users with Topical Interest

To identify users interested in each domain we classified pages in the browser trails into the topical hierarchy from a popular web directory, the Open Directory Project (ODP). Given the large number of pages involved, the classification needed to be automatic. Our classifier assigned labels to pages based on the ODP in a similar way to Shen et al. [16], by starting with URLs that were in the ODP and backing-off to cover other URLs. Using this classifier we calculated the proportion of pages that each user

visited that were related to each domain via the classification. We used the following ODP categories for each domain:

- Medicine: Health/Medicine
- Finance: Business/Financial_Services
- Legal: Society/Law/Legal_Information
- Computer Science: Computers/Computer_Science

Although the objective was to identify people with some degree of interest in each of the topics, we also wanted to remove outliers who viewed only a few pages. Thus, we selected people who viewed 100 or more pages of any type over the three-month duration of the study, and whose page views contained 1% or more of domain-related pages. Table 1 shows the number of selected users, the total number of search sessions, and the number of search sessions in each domain from these users.

Table 1. Number of users, sessions, and in-domain sessions.

Domain	# users	# sessions	# in-domain sessions
Medicine	45,214	1,918,722	94,036
Finance	194,409	6,489,674	279,471
Legal	25,141	1,010,868	36,418
CS	2,427	113,037	3,706

There are around 40 search sessions per user and around 5% of search sessions were on each of the four domains of interest. These selected users and their in-domain and out-of-domain sessions extracted from the logs form the basis for the analysis in the remainder of this paper.

4.2 Separating Experts from Non-Experts

Researchers have identified domain experts using user surveys or the completion of an academic course in the domain in question. These techniques can suffer from low participation rates and small sample sizes. To conduct a large-scale study of domain expertise it is necessary to identify experts and non-experts based solely on observable behavior. For this reason, we divided users based on whether they had ever visited one or more specific web sites. These expert sites (also referred to as expert URL *filters*) were identified through discussion with domain experts in each of the four subject areas. They were as follows:

Medicine: Visits to the U.S. National Library of Medicine’s web-based PubMed service. PubMed is used primarily by medical researchers and physicians, and provides access to citations and abstracts of biomedical research articles. A similar filter was used in earlier work to separate medical experts from non-experts [21].

Finance: Visits to online financial services Bloomberg, Hoovers, Edgar Online, and the U.S. Securities and Exchange Commission (SEC) that are used by financial professionals such as investment bankers and accountants. These sites provide market reports, analysis, and financial regulatory information.

Legal: Visits to major online legal research services Westlaw and LexisNexis that are used by lawyers and other members of the legal services profession. These services assist in finding, understanding, and applying the law and legal concepts.

Computer Science: Visits to the Association for Computing Machinery’s (ACM) Digital Library. The digital library contains the full-text repository of scientific papers that have been published by the ACM. These publications are typically read by computer science researchers in academia and industry, as well as students in computer science programs.

With the exception of PubMed and SEC, these sites require user subscriptions that can be prohibitively expensive for the general population. While some of the sites also contain valuable free content, it is likely that even those people who visit them for the free content have greater than average domain knowledge. Table 2 presents the expert filters used and the number of experts and non-experts identified in each domain using this technique.

Table 2. Expert URL filters and user group sizes.

Domain	Expert URL filters	Expert	Non-expert
<i>Medicine</i>	ncbi.nlm.nih.gov/pubmed pubmedcentral.nih.gov	7,971 (17.6%)	37,243 (82.4%)
<i>Finance</i>	bloomberg.com egdar-online.com hoovers.com sec.gov	8,850 (4.6%)	185,559 (95.4%)
<i>Legal</i>	lexis.com westlaw.com	2,501 (9.9%)	22,640 (90.1%)
<i>CS</i>	acm.org/dl portal.acm.org	949 (39.1%)	1,478 (60.9%)

As we can see from the table, the use of expert filters resulted in a large number of experts in each of the four domains. These large samples allow us to examine a wide range of naturally-occurring search sessions within each of these domains. This is important if we want to generalize to the breadth of search tasks and strategies that domain experts exhibit. There may be alternative methods for identifying in-domain sessions and domain experts, and it would be easy for us to apply them in a similar framework.

In addition to dividing users into expert and non-expert groups we also classified their search sessions as in-domain and out-of-domain. We performed this classification for each session based on whether it contained a page tagged with the domain ODP label listed in Section 4.1 by our classifier (e.g., search sessions containing at least one page classified as “Health/Medicine” were regarded as medical). This provided us with domain experts and non-experts, engaged in in-domain and out-of-domain sessions. In the next section we characterize the search behavior and related attributes of experts and non-experts in all four domains.

5. CHARACTERIZING EXPERTISE

In this section we compare and contrast the search behavior of experts and non-experts both within their domains of expertise, and outside of them. We analyze several characteristics of search behavior: querying (attributes of textual queries that are submitted to search engines), sessions (attributes of users’ interaction behavior during search sessions), and source selection (attributes of the pages that users’ visit). We then look at how expertise within a domain affects search success and how domain expertise develops over time. Finally, since task differences could account for some of the observed variation we compare behavior across several common tasks.

Our findings on search behavior are summarized in Table 4 and discussed in greater detail below. In Table 4 we present the means (\bar{M}) and the standard deviations (SD) of the query and session attributes for experts and non-experts separately for sessions in their domain of expertise and those outside the domain. Given the large sample sizes, most of the observed differences in the means between user groups were statistically significant with independent measures t -tests (at $p < 0.01$). We used Cohen’s d tests to determine the effect size of each between-group comparison [6]. Table 4 also shows the obtained d -values.

In general, differences between domain experts and non-experts are much larger inside the domain than outside of it. This means that the differences observed for the in-domain sessions reflect domain expertise and not general individual differences or search skills. We now describe the results in more detail.

5.1 Queries

We begin with a discussion of the differences in the queries issued by domain experts and non-experts. The queries submitted to search engines may provide clues about the expertise of users. For each of the four domains we examined features of the query length (in tokens and characters) and the query vocabulary. Based on previous laboratory studies (e.g., [10,11]) we conjectured that those with expertise within a particular domain would issue longer queries and use a more technical query vocabulary. To quantify query vocabulary we obtained a domain-specific lexicon for each of the four domains. In Table 3 we define the lexicons used for each domain and state the number of entries in each.

Table 3. Domain lexicons and number of entries.

Domain	Lexicon	# entries
<i>Medicine</i>	MedlinePlus medical encyclopedia (available from http://nlm.nih.gov/medlineplus/encyclopedia.html)	3,535
<i>Finance</i>	Financial dictionary (available from http://anz.com/edna/dictionary.asp)	2,476
<i>Legal</i>	Legal dictionary (available from http://law.dictionary.com)	2,463
<i>CS</i>	1998 ACM Computing Classification System (available from http://www.acm.org/class/1998)	1,361

For sessions in their domain of expertise, experts’ queries, as hypothesized, were longer and contained more of the vocabulary from the domain-specific lexicon. The vocabulary feature was consistently among the most important features, as measured by Cohen’s d . Experts generated queries containing words from these lexicons fifty percent more often than non-experts (37% vs. 26% of queries). In addition to being able to generate more technically-sophisticated queries, experts also generated longer queries in terms of tokens and characters. It may be that because domain experts are more familiar with the domain vocabulary, it was easier for them to self-generate content to include in the query. The magnitude of the differences was fairly consistent across domains, except for the legal domain where the differences were smaller. It is possible that legal terminology is used comfortably by both legal experts and non-experts.

For sessions outside their domain of expertise, the differences between experts and non-experts were much smaller. The Cohen’s d values were generally less than .10. Thus, the differences observed for the in-domain sessions reflect domain expertise and not general individual differences or search skills.

5.2 Search Sessions

We conjectured that experts and non-experts would exhibit differences in their search behavior. To compare groups we examined the following features of in-domain search sessions:

1. Session length (pages): The number of web pages visited in the session, including search engine home pages and search engine result pages.
2. Session length (queries): The number of web search engine queries issued in a session.
3. Session length (seconds): The total amount of time spent in the search session.

4. **Branchiness:** The number of re-visits to previous pages in the session that were then followed by a forward motion to a previously unvisited page in the session.
5. **Number of unique (non-search engine) domains:** The number of unique non-search engine domains in a session gives a sense of the breadth of coverage.
6. **Average page display time (seconds):** The average length of time for which a web page is viewed during a session.
7. **Ratio of querying to browsing:** The proportion of the session that is devoted to querying versus browsing pages retrieved by the search engine or linked to from search results. A high number (much greater than one) means that the session was query-intensive. In contrast, a low number (much less than one) means that the session was browse-intensive.

These features offer insight into how the different user groups interacted with web search engines and the web pages they visited during their search sessions.

In Table 4 we show the results of the session analysis for each of the four domains. For in-domain sessions, the search behavior of experts in all four domains differed from non-experts on most measures. And, for sessions outside the domain of expertise, the differences between experts and non-experts are smaller. The d -values reported in Table 4 for the out-of-domain comparisons suggest that the magnitude of the treatment effects is small.

The most noticeable differences were in the length of the sessions, the number of unique domains visited, and the “branchiness” of the session. The sessions conducted by domain experts were generally longer than non-expert sessions. Domain experts consistently visited more pages in a session, and in three of the four domains they spent more time and issued more queries. This could indicate a greater investment in the topics by experts than non-experts. The information being sought may be more important to the experts, making them more likely to spend time and effort on the task.

An interesting aspect of the experts’ sessions is that they also appear to be more diverse than non-expert sessions, with experts exhibiting more branchiness and visiting more unique domains in all cases (as also in [4]). It may be that experts have developed strategies to explore the space more broadly than non-experts.

The smallest differences were consistently in page display time, and in two cases (medicine and computer science) domain experts were faster than non-experts. This suggests that domain experts are more adept at reading domain-relevant pages, as others have sometimes observed [9].

5.3 Source Selection

Given the variation in session behavior, we investigated the nature of the web sites visited by experts and non-experts within each domain. To do this formally we analyzed the nature of visited page URLs and collected human judgments of the expertise level of popular pages visited in each of the domains. This section describes the results of our analysis.

5.3.1 URL-based

We first examined the features of the URLs of web pages visited by experts and non-experts in each domain. The objective was to determine whether there were any noticeable differences in the type of pages the user groups visited. Table 5 shows the top-ten most popular top-level web domains visited by users, and the distribution of domain name extensions (e.g., .com, .gov, .edu).

Table 5. Most popular domains and distribution of URLs over top-level domains. Top-level domains with differences < 2% are grouped into the “other” category.

Domain	Expert		Non-expert		
<i>Medicine</i>	nih.gov medicinenet.com mayoclinic.com medscape.com emedicine.com healthline.com rxlist.com nejm.org cdc.gov americanheart.org		medicinenet.com mayoclinic.com implantinfo.com about.com locateadoc.com emedicinehealth.com drugs.com plasticsurgery.org justbreastimplants.com webmd.com		
	Extension	% of pages	Extension	% of pages	
	com	46%	com	61%	
	org	26%	org	23%	
	gov	14%	gov	6%	
	edu	8%	edu	5%	
	other	6%	other	5%	
<i>Finance</i>	citibank.com americanexpress.com ml.com gs.com citigroup.com jpmorgan.com ms.com wachovia.com visa.com dnb.com		capitalone.com citibank.com americanexpress.com sovereignbank.com discovercard.com nationwide.com visa.com scotiabank.com bankofamerica.com wachovia.com		
	Extension	% of pages	Extension	% of pages	
	com	89%	com	87%	
	other	11%	other	13%	
	<i>Legal</i>	findlaw.com uspto.gov hhs.gov lawguru.com lexisone.com laborlawtalk.com eec.gov alllaw.com expertlaw.com ilrg.com		findlaw.com uspto.gov freeadvice.com freepatentsonline.com lawguru.com nolo.com divorcenet.com workerscompensation.com alllaw.com lectlaw.com	
		Extension	% of pages	Extension	% of pages
com		48%	com	62%	
org		6%	org	8%	
gov		37%	gov	23%	
other		9%	other	7%	
<i>CS</i>	acm.org ieee.org nist.gov sigmod.org columbia.edu cornell.edu cmu.edu msdn.com computer.org codeplex.com		microsoft.com download.com msdn.com codeproject.com nist.gov sun.com codeplex.com dell.com w3schools.com adobe.com		
	Extension	% of pages	Extension	% of pages	
	com	57%	com	71%	
	org	22%	org	9%	
	edu	11%	edu	8%	
	other	10%	other	12%	

The differences between experts and non-experts in URL selection appears to vary by domain. Legal experts are more likely to visit government pages than non-experts, which may reflect the direct use of laws and statutes by legal experts. Medical experts visit more government and educational sites, reflecting a preponderance of public data available at those locations. Computer scientists visit a relatively large proportion of organizational and educational sites, reflecting visits to conference

web sites (that typically have .org domain name extensions), major U.S. computing societies (ACM and IEEE), and the large U.S. computer science academic community.

In all of the preceding cases, experts visited fewer commercial sites. They appeared to focus on technical detail, while non-experts focused on more consumer-oriented or advisory aspects. In contrast, financial experts visit approximately the same proportion of commercial web sites as non-experts. This may reflect the commercial nature of finance.

In addition to looking at the domains visited by each of the user groups, we also investigated the domains unique to each group. The findings showed clear differences in the types of web sites that are unique to experts and non-experts within each domain. From the findings, we hypothesize that medical experts visited websites containing information on specific conditions relevant to their specialty (e.g., acc.org, a resource offering benefits and services for cardiologists). In contrast, it seems that medical non-experts visited sites related to conditions or medical procedures that were relevant to them (e.g., obesity, breast augmentation). Finance experts visited sites on funds, investments, and securities, whereas non-experts visited credit unions and credit advisory sites. Legal experts appeared concerned with regulations and legal precedents, while non-experts were interested in particular legal scenarios such as speeding tickets and rented accommodation. Computer science experts visited sites that were specific to a programming language (e.g., lyx, smalltalk). In contrast, computer science non-experts appeared interested in the customization of their operating system, their desktop environment, or protecting their personal computer from viruses.

From this analysis it appears that the web sites visited by domain experts were more technical in nature. However, analyzing URLs does not allow us to formally compare the content of the web sites that domain experts and non-experts visited when searching within their domain of interest. For this reason, in addition to studying web sites URLs, we also evaluated the content of popular pages visited by users in within each of the four domains.

5.3.2 Content

The goal of analyzing page content was to see if human judgments of the technical depth of pages were associated with our automatic methods of identifying experts. To do this we identified the top 150 most popular URLs within each domain visited by experts and non-experts, for a total of 600 pages. Two of the authors of this paper independently judged the technical detail of the pages, rating each page as either expert or not. Although these authors were domain experts in computer science, they judged all domains in the interests of judgment consistency.

The judged URLs contained a 20% overlap in pages to determine inter-rater reliability. There was substantial agreement between raters (Cohen’s Kappa $\kappa=0.72$), with the agreement ranging some according to domain from moderate (Finance, $\kappa=0.45$) to almost perfect (Computer Science, $\kappa=0.89$). Given that the authors are domain experts in the area of computer science, it is not surprising that this is the area of greatest overlap.

Table 6 shows the number of popular pages from each domain that were visited by either experts or non-experts, broken down by whether the pages were rated as expert or not. Pages visited by users in both groups are represented in both the expert and non-expert columns. Pages for which the raters disagreed or for which no judgment could be obtained (e.g., the page did not load) are excluded from analysis.

Table 6. Number (and percent) of pages rated as being a resource for experts or non-experts, broken down by group.

Domain	Expert visitor		Non-expert visitor		Cohen’s Kappa
	Rated expert	Rated non-expert	Rated expert	Rated non-expert	
Medicine	100 (68%)	46 (32%)	66 (46%)	79 (54%)	0.75
Finance	30 (22%)	108 (78%)	10 (8%)	121 (92%)	0.45
Legal	79 (56%)	63 (44%)	65 (47%)	72 (53%)	0.64
CS	107 (89%)	13 (11%)	18 (21%)	66 (79%)	0.89

In all cases, domain experts visited more web sites rated expert than domain non-experts did. Overall, 58% of the common pages visited by experts were judged expert, and 42% were judged non-expert. In contrast, only 32% of the pages commonly visited by non-experts were classified as expert, while 68% were judged non-expert. The trend is most pronounced for medicine and computer science. For those domains the inter-rater reliability is also highest (as measured by Cohen’s Kappa). The trend may be less pronounced for the finance and legal domains because experts visit many sites also visited by non-experts, and the explicit judgments in those domains are of lower quality since the human judges were not domain experts.

5.4 Search Success

In the previous sections we observed several important differences in the queries experts and non-experts use, the search session behavior, and the resource selection. In this section we investigate how successful users from both groups appeared to be when searching in domain and out of domain.

Since our study was log-based we could not control user task or confirm whether searchers had been successful in their search session. Instead we had to approximate success heuristically. Search result click-through data has been used previously to develop models of user preferences [1]. We used our logs and looked at the final action in a session. If the final event in a search session was a URL click we scored the session as a success, and if the final action was a query we score the session as a failure. In Table 7 we present the proportion of search sessions that were deemed successful with this metric.

Table 7. Percentage of successful sessions, by expertise. The Pearson’s R with domain expertise is reported.

Domain	In domain		Out of domain		Pearson’s correlation coefficient (R)
	Expert	Non-expert	Expert	Non-expert	
Medicine	84.9%	76.8%	75.4%	78.6%	0.55
Finance	84.6%	81.4%	80.1%	82.1%	0.38
Legal	82.3%	79.9%	79.3%	81.1%	0.49
CS	83.6%	69.1%	72.8%	71.0%	0.66
Average	83.9%	76.9%	76.8%	78.2%	0.52

Although this was only an approximation for search success, any overestimation should affect all groups equally. Findings showed that experts were more successful than non-experts when searching within their domain of expertise. When searching in out-of-domain sessions, experts and non-experts had comparable levels of search success, with non-experts being somewhat more successful on average. It is interesting that for each domain, experts performed better in domain than out of domain, while non-experts actually performed worse in domain than out of domain. Non-experts exhibited clear interest in the domain, with

at least 1% of all of the pages they visited falling within the domain, but still appeared to have trouble working in that domain. It may be that they lacked the technical expertise necessary to succeed in their searches.

To further probe the relationship between search success and domain expertise we examined the correlation between the two variables for in-domain sessions, as suggested by [23]. For each user, we calculated a search success and domain expertise score. Search success was defined as the proportion of successful sessions, and domain expertise was represented as the proportion of query terms with domain-specific vocabulary (since this was the most important variable for distinguishing experts and non-experts). For each domain we computed Pearson’s correlation coefficient (R) between the two measures. The values for R are reported in the last column of Table 7. They suggest that there is a good correlation between level of expertise and degree of search success. That is, the more expert a user is the more likely they are to be successful when searching in their domain of interest.

5.5 Development of Expertise

The development of people’s domain expertise over time has been observed in prior longitudinal studies [19,23]. Our logs afforded us the opportunity to track an individual user over the three-month duration of our investigation. While users may have domain knowledge extending outside the window, we wished to determine whether there was any evidence of users’ domain expertise developing over time. The presence of such evidence would suggest that we could remotely estimate aspects of user learning which may be useful in offering tailored search support.

To investigate domain expertise development, we examined the proportion of queries containing domain-specific vocabulary (the strongest predictor of search expertise that emerged from our earlier analysis) at a series of time points across the three months. To begin, we divided the three months into 13 one-week periods. To give us sufficient data we restricted our analysis to users for whom we had five or more weeks of data in the duration of the study. We then computed the proportion of queries with in-domain sessions that contained domain-specific vocabulary in each week. To gain a sense for whether query vocabulary was expanding we computed the Pearson’s R across all available data points for each user. This offered a sense for whether their query vocabulary was increasing, decreasing, or remaining constant across the five or more weeks we studied. In Table 8 we present the proportion of users whose query vocabulary usage trended downwards, those where it remained the same (i.e., R lies between -0.1 and 0.1), and those where it increased.

Table 8. Percentage of users with increase (\uparrow), decrease (\downarrow), or no change (\leftrightarrow) in query vocabulary.

Domain	Experts			Non-expert		
	\downarrow	\leftrightarrow	\uparrow	\downarrow	\leftrightarrow	\uparrow
<i>Medicine</i>	9.8%	74.9%	15.3%	8.3%	43.1%	48.6%
<i>Finance</i>	10.1%	75.8%	14.3%	9.9%	52.0%	38.1%
<i>Legal</i>	13.2%	73.2%	13.6%	15.2%	54.1%	30.7%
<i>CS</i>	9.4%	72.1%	18.5%	11.1%	51.7%	37.2%

The results show that experts’ use of domain-specific vocabulary changes only slightly over the duration of the study. However, many non-expert users exhibit an increase in their usage of domain-specific vocabulary. This provides evidence that suggests that non-expert domain expertise may be developing over time.

However, such welcome developments in users’ domain expertise also demonstrate the difficulty in monitoring domain non-experts over a period of time; domain expertise is dynamic.

5.6 Common Tasks

A possible confound in the above analysis is that the observed differences may be a function of task differences rather than expertise differences. Our observations would be the same if experts had inherently different tasks than non-experts, and exhibited the same behavior as non-experts for the same tasks.

To address this concern we developed two methods to identify comparable tasks: (i) we identified search sessions that began with the same query, and (ii) we identified sessions that ended with the same URL. The number of sessions with the same initial query or same last URL varied depending on the domain, from 20-30 for computer science to around ten thousand for finance.

For these linked sessions we extracted the same set of features of the queries and sessions as were shown in Table 4. Our analysis of these sessions showed that for matched queries and sessions, the between-group differences noted earlier in this paper held true. It seems that task differences do not significantly impact user interaction patterns. However, more work is necessary to study the effects of user intent on search behavior, especially given the recent findings of Teevan et al. [18].

6. SEARCHING BETTER VIA EXPERTISE

We have seen in our study that domain experts employ different search strategies and are more successful than non-experts in four different domains. Given these differences, we believe it is possible to help people better find what they are looking for by taking into account their domain expertise.

One way our findings could be used to improve the search experience would be tailor the results shown or search aids such as query suggestions to match the expertise of the searcher. Alternatively, the search strategies employed by domain experts could be used to support non-experts in learning more about domain resources and vocabulary.

Regardless of the specific strategies employed, any search system that takes advantage of domain expertise needs to be able to identify whether a user is an expert or a non-expert, and then modify the experience accordingly. For this reason, in this section we explore how well domain expertise can be automatically predicted. We focus on techniques that work using only observable search behavior and history as input, since these can be deployed widely. We then discuss how we can use prediction.

6.1 Predicting Domain Expertise

To predict domain expertise we developed a classifier based on the features of users’ interaction behavior listed in Table 4 within each of the domains. We employed a maximum-margin averaged perceptron [7] as the classifier, since it was appropriate for our binary classification task and has previously shown excellent empirical performance in many domains from natural language to vision. A separate version of the averaged perceptron was trained for each of the four domains. We focused on three prediction challenges: (i) whether an in-domain session was conducted by a domain expert, (ii) whether we could identify domain experts *during* the course of a session by predicting after successive actions (queries or page visits), and (iii) whether a user was a domain expert given multiple in-domain sessions. We now describe each in turn.

6.1.1 Post-Session Expertise Prediction

To train the classifier to predict if an observed in-domain session was conducted by a domain expert or not, we treated in-domain search sessions performed by domain experts as positive examples and search sessions performed by non-experts as negative examples. All of the query and session attributes shown in Table 4 were used as features to train the classifier. We performed a five-fold cross-validation experiment across ten runs, and generated precision-recall curves that summarize the performance of the classifier trained on each domain (Figure 1).

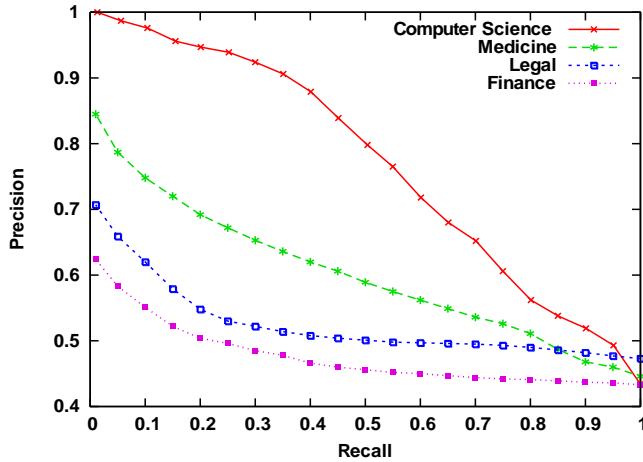


Figure 1. Prediction accuracy for different domains.

The precision-recall curves show the number of search sessions accurately classified as expert or non-expert at different recall levels. The curves illustrate that it is possible to accurately predict whether an observed session was performed by a domain expert or a domain non-expert, especially at low recall levels. Recall from Table 2 that anywhere from 4.6% to 39.1% of all people involved in domain-specific sessions were experts. The classification accuracy is highest for computer science, and this most likely reflects the fact that that was also the domain with the highest percentage of identified experts.

6.1.2 Within-Session Expertise Prediction

Post-session prediction has limited utility since the session must be complete before a prediction can be made. An attractive alternative is to predict expertise on-the-fly *during* the course of a session. While a web browser or client-side application may indeed know all of the interactions, search engines and other web applications typically do not. We trained the classifier using all of the features in Table 4. In addition, we also explored using only those features related to the query, derived at each query iteration, and using only those features of pages visited (e.g., page display time, number of unique domains), derived at each web page view.

To examine within-session predictions, we selected the 2181 CS sessions (59% of the total in-domain sessions) that contained at least five queries and at least five non-result page visits. These sessions enabled us to study accuracy for a sizeable number of actions over the same set of sessions. We used five-fold cross validation over ten experimental runs to train our classifier and evaluate accuracy, computing prediction accuracy after each action. The average accuracy across the ten runs is reported in Table 9, for all features and for query and web page features separately. Prediction accuracy after observing a full session with each feature set is also included for reference.

Table 9. Mean prediction accuracy for all / queries / pages. Significant p -values from t -tests comparing accuracy with baseline (.566) are shown with * = $p < .05$ and ** = $p < .01$.

Action type	Action number					Full session
	1	2	3	4	5	
All	.616*	.625*	.639**	.651**	.660**	.718**
Queries	.616*	.635**	.651**	.668**	.683**	.710**
Pages	.578	.590*	.608*	.617*	.634**	.661**

The findings show that we can generally predict CS domain expertise within a session after only a few user actions, compared with a maximal margin baseline that always predicts non-expert (accuracy = .566). Our findings show that predictions based on queries yielded a higher accuracy than page-based predictions or predictions from all features given few observations. As observed earlier, queries are a good source of domain expertise evidence. Across full sessions, the use of all features slightly outperforms query features. These findings hold for the other domains studied. Querying activity is readily available to search engines and could be classified immediately to rapidly tailor the search experience.

6.1.3 User Expertise Prediction

We also explored how well we could predict whether a user is a domain expert given interaction history across multiple sessions. To do this we selected users from each domain with at least five sessions in the interaction logs and monitored the improvement obtained by using additional sessions. For each domain we incremented the number of sessions used to compute the features and recorded the accuracy obtained at each iteration, up to at most five sessions. Sessions were used in chronological order to mimic how they would arrive in an operational setting. In addition to the single-session features, we also used several inter-session features including the time between in-domain sessions and average number of observed in-domain sessions per day.

Figure 2 shows the learning curves for each of the four domains generated from five-fold cross validation experiments across ten runs. The curves plot accuracy (i.e., the proportion of times the classifier accurately determines whether a user is an expert or a non expert) against the number of search sessions presented, averaged across all runs. The findings show that prediction accuracy for all classifiers improves with additional search sessions, but the marginal improvement decreases.

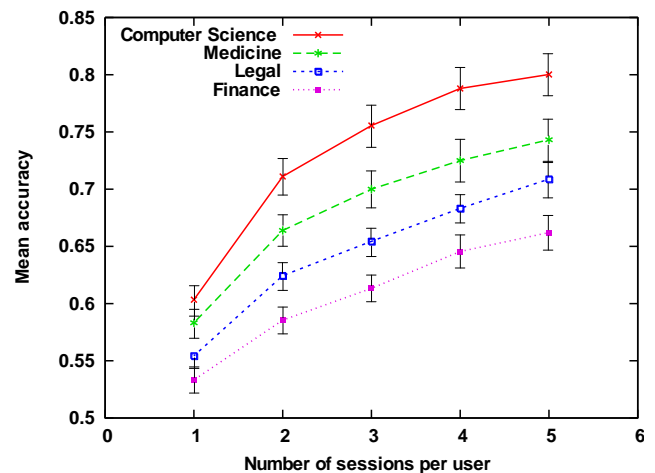


Figure 2. Accuracy given more search sessions per user (\pm SE).

Clearly having a lot of information about a user's web search activity is valuable for predicting domain expertise. A web browser, Internet service provider, or large domain-specific web site may be able to collect session-based information as described above and use it to tailor the experience.

However, such rich behavior information is often not available to popular search tools like online commercial search engines. To understand how performance of the classifier would be affected if interactions were limited to only those visible from a search engine, we trained a classifier using only queries and result clicks. The findings show a dip in accuracy of 5-10% across all domains, but the general trends identified in this section remain.

6.2 Improving the Search Experience

Knowledge of an individual's expertise level can be used to support the search process in many ways. For example, a search engine (or client-side application) could bias its results towards the web sites that people with similar expertise prefer, and provide query suggestions or query re-writing that use level-appropriate terminology. A challenge with this approach is that it reinforces behavior rather than encouraging people to learn over time. That is, non-experts are encouraged to search more like non-experts, rather than to gain expertise. This approach stands in contrast to the development of expertise over time observed in Section 5.5.

For this reason, it may be worthwhile for search tools to consider how they can better help domain non-experts become domain experts over the course of time. Bhavnani et al.'s work on Strategy Hubs [5] is an example of a system that supports and educates non-experts by providing critical search procedures and associated high-quality links. We believe that providing such support could be extended more broadly. One way a search engine could do this would be to provide non-expert definitions for related expert terms when a person searches. The results for a search for "cancer," for example, may include a definition of "malignancy," which would, in turn, help the non-expert better understand the technical vocabulary and the complete information space. Non-experts could also be taught to identify reliable, expert sites as they gain the necessary knowledge to understand them better, or to examine the broader range of information that experts do. By helping non-experts move from tutorial information to more detailed information, search tools can help them develop domain expertise through their search experience.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have described a large-scale, log-based study of the web search behavior of domain experts and non-experts. Our findings demonstrate that, within their domain of expertise, experts search differently than non-experts in terms of the sites they visit, the query vocabulary they use, their patterns of search behavior, and their search success. These differences were obtained in naturalistic web search sessions in four domains, thus extending previous lab studies in terms of breadth and scale.

We have also developed models to predict domain expertise using characteristics of search interactions. By focusing on attributes that are readily available from web search behavior, we can apply our models in real-time as part of a web search engine.

The identification of domain experts can allow us to provide expert query suggestions and site recommendations to non-expert users, and personalize results or suggestions based on expertise. Future work will involve developing such applications, testing them with human subjects, and deploying them at web scale.

8. REFERENCES

- [1] Agichtein, E., Brill, E., Dumais, S. & Ragno, R. (2006). Learning user interaction models for predicting web search result preferences. *Proc. SIGIR*, 3-10.
- [2] Allen, B.L. (1991). Topic knowledge and online catalog search formulation. *Lib. Quart.*, 61(2), 188-213.
- [3] Bhavnani, S.K. (2001). Important cognitive components of domain-specific search knowledge. *Proc. TREC*, 571-578.
- [4] Bhavnani, S.K. (2002). Domain-specific search strategies for the effective retrieval of healthcare and shopping information. *Proc. SIGCHI*, 610-611.
- [5] Bhavnani, S.K. et al. (2005). Strategy Hubs: Domain portals to help find comprehensive information. *JASIST*, 57(1), 4-24.
- [6] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.
- [7] Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. *Proc. EMNLP*, 1-8.
- [8] Downey, D., Dumais, S. & Horvitz, E. (2007). Models of searching and browsing: Languages, studies and application. *Proc. IJCAI*, 2740-2747.
- [9] Duggan, G.B. & Payne, S.J. (2008). Knowledge in the head and on the web: Using topic expertise to aid search. *Proc. SIGCHI*, 39-48.
- [10] Freund, L. & Toms, E.G. (2006). Enterprise search behaviour of software engineers. *Proc. SIGIR*, 645-646.
- [11] Hembrooke, H.A., Gay, G.K. & Granka, L.A. (2005). The effects of expertise and feedback on search term selection and subsequent learning. *JASIST*, 56(8), 861-871.
- [12] Hölscher, C. & Strube, G. (2000). Web search behavior of internet experts and newbies. *Comp. Net.*, 33(1), 337-346.
- [13] Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *JASIST*, 44(3), 161-174.
- [14] Kelly, D. & Cool, C. (2002). The effects of topic familiarity on information search behavior. *Proc. JCDL*, 74-75.
- [15] Marchionini, G., Lin, X. & Dwiggins, S. (1990). Effects of search and subject expertise on information seeking in a hypertext environment. *Proc. ASIS*, 129-142.
- [16] Shen, X., Dumais, S. & Horvitz, E. (2005). Analysis of topic dynamics in web search. *Proc. WWW*, 1102-1103.
- [17] Spink, A. & Jensen, B. (2004). *Web search: Public searching of the web*. Kluwer Academic Publishers.
- [18] Teevan, J., Dumais, S.T. & Liebling, D.J. (2008). To personalize or not to personalize: Modeling queries with variation in user intent. *Proc. SIGIR*, 163-170.
- [19] Vakkari, P., Pennanen, M. & Serola, S. (2003). Changes in search terms and tactics while writing a research proposal: A longitudinal case study. *IP&M*, 39(3), 445-463.
- [20] White, R.W. & Drucker, S.M. (2007). Investigating behavioral variability in web search. *Proc. WWW*, 21-30.
- [21] White, R.W., Dumais, S. & Teevan, J. (2008). How medical expertise influences web search interaction. *Proc. SIGIR* 791-792.
- [22] White, R.W. & Morris, D. (2007). Investigating the querying and browsing behavior of advanced search engine users. *Proc. SIGIR*, 255-262.
- [23] Wildemuth, B.M. (2004). The effects of domain knowledge on search tactic formulation. *JASIST*, 55(3), 246-258.
- [24] Zhang, X., Angheliescu, H.G.B. & Yuan, X. (2005). Domain knowledge, search behavior, and search effectiveness of engineering and science students. *Inf. Res.*, 10(2), 217.