Lessons from the Journey: A Query Log Analysis of Within-Session Learning

Carsten Eickhoff
Data Analytics Lab
ETH Zurich
Switzerland
c.eickhoff@acm.org

Jaime Teevan, Ryen White,
Susan Dumais
Microsoft Research, WA, USA
{sdumais, teevan, ryenw}
@microsoft.com

ABSTRACT

The Internet is the largest source of information in the world. Search engines help people navigate the huge space of available data in order to acquire new skills and knowledge. In this paper, we present an in-depth analysis of sessions in which people explicitly search for new knowledge on the Web based on the log files of a popular search engine. We investigate within-session and cross-session developments of expertise, focusing on how the language and search behavior of a user on a topic evolves over time. In this way, we identify those sessions and page visits that appear to significantly boost the learning process. Our experiments demonstrate a strong connection between clicks and several metrics related to expertise. Based on models of the user and their specific context, we present a method capable of automatically predicting, with good accuracy, which clicks will lead to enhanced learning. Our findings provide insight into how search engines might better help users learn as they search.

Categories and Subject Descriptors

H.3.3 [**Information Search & Retrieval**]: Search Process;
I.2.6 [**Learning**]: Knowledge Acquisition;
H.1.2 [**User-Machine Systems**]: Human Factors

Keywords

User Modeling, Domain Expertise, Information Search, Search Intent

1. INTRODUCTION

With size estimates of 30 to 50 billion indexed pages on commercial web search engines [10], the Internet is a very large collection of information. Empowered by affordable and easily available Internet connections, searching for information on-line has become a natural step in the knowledge acquisition process for modern society [12]. The wealth of available information makes tools such as Web search engines indispensable in identifying and accessing information.

In recent years, search systems have employed representations of the user's preferences, current context, and recent behavior to more accurately model user interests and intentions [33]. This allows systems to better address the specific needs of individual searchers rather than considering only the dominant search intent across all users. However, even with such advanced retrieval technology, exploratory and open-ended information needs can be challenging to satisfy, especially when the searcher is unfamiliar with the domain in question [11]. Previous work [36] has shown that searchers acquire domain knowledge by means of web search. Continued exposure to and interaction with information appears to influence the users domain expertise over time. This increase in expertise is a byproduct of the actual search process. This observation is of particular interest with regard to modern relevance models. State-of-the-art search algorithms attempt to maximise the relevance of a list of results retrieved for an expressed information need. Commercial search engines often assess relevance using explicit judgments and click-through statistics, considering long dwell times as a proxy for searcher satisfaction. Following this paradigm of the shortest path to the goal, the searcher might not be exposed to additional, relevant information [37].

In this paper, we study the development of expertise at the session level in order to better understand the value that the journey towards the final result, as opposed to just that result in isolation, holds for the searcher.

Our work makes three major contributions over the state of the art in knowledge acquisition: (1) We present an investigation of explicit knowledge seeking sessions dedicated to finding procedural or declarative information. Based on such sessions, we examine changes in domain expertise within search sessions and show how these changes are sustained across session boundaries. (2) We investigate factors related to changes in domain expertise by studying the connection between page visits and the subsequent development of expertise and searcher behavior. (3) Finally, we present an automatic means of predicting the potential for learning from page visits.

The remainder of this paper is structured as follows: Section 2 provides an overview of related work in the areas of Web search personalization, search intent frameworks, domain expertise, and exploratory search. In Section 3, we identify web search sessions in which people explicitly seek to acquire knowledge and characterize different knowledge acquisition intent types seen in the log files of a popular Web search engine. Section 4 studies how domain expertise develops within a session and how it is sustained across ses-

WSDM'14, February 24–28, 2014, New York, New York, USA.
Copyright 2014 ACM 978-1-4503-2351-2/14/02 ...$15.00.
http://dx.doi.org/10.1145/2556195.2556217.

sions. Section 5 investigates the reasons that underly the previously observed learning. In Section 6, based on page-level features, we predict which page visits are most likely to help the users in expanding their domain expertise. Sections 7 and 8 present a discussion of our findings and their implications on future research, and practical applications in Web search.

2. RELATED WORK

The work presented in this paper is situated at the intersection of several areas of related prior work: Search personalization, investigations of domain expertise, search intent classification schemes and exploratory search and their respective roles in information retrieval applications. We will discuss each of these in turn.

Search engine providers employ a wide array of personalization techniques and large-scale resources such as previous interactions or external resources to accurately determine the relevance of documents to a query for individual searchers [26, 34]. Personalization is often applied in a re-ranking step following the original global ranking. Examples of such efforts include topical user profiles based on previous browsing or search activity [29]. Teevan et al. [33] re-ordered the top 50 search results based on previous user profiles, finding that full-text models outperformed selective keyword models. Li et al. [20] proposed a dynamic graph adaptation scheme to account for changes in the user's short term browsing behavior. Bennett et al. [5] investigated trade-offs between long- and short-term usage histories for search personalization. Eickhoff et al. [11] investigated personalizing atypical sessions in which searchers seek unfamiliar topics that would not benefit from their general search history.

Recent work has considered the searcher's familiarity with the topic domain as a ranking criterion. White et al. [36] discovered differences in the Web search behavior of domain experts and novices. Experts displayed significantly higher search success for their topic of expertise, and found evidence that domain expertise increases as searchers spent several sessions on the same topic. White et al. identified several factors associated with differences in behaviors between experts and novices, including query terms, the diversity of sites visited, the time spent reading, and search strategies. Zhang et al. [40] used behavioral features in order to predict users' domain expertise. Liu et al. [21] studied the evolution of knowledge across multiple sessions of the same overarching task. Based on a lab-based user study, the authors find a general tendency for knowledge to increase with each successive search session. Wildemuth [38] studied the connection between domain expertise and search strategies, finding that domain novices converge towards the same search patterns as experts as they are exposed to the topic and learn more. Previous work primarily focused on the long-term, cross-session dynamics of domain expertise, typically measured over the course of an academic semester. To the best of our knowledge, there has not yet been an investigation of domain expertise at the session level. In this paper, we close this gap by studying the within-session development of domain expertise and its connection to sustained learning across sessions.

The complexity of content has also been shown to be related to domain expertise. Tan et al. [32] measure reading level and comprehensibility of content on the community question answering portal Yahoo! Answers. Depending on the user's domain expertise, simple vs. more technical answers, respectively, were ranked higher. Collins-Thompson et al. [9] factor readability into Web search personalization, showing that accommodating for users' preferences in document complexity can significantly improve rankings. Kim et al. [16] connect user, complexity and topic, and show that personalized readability models perform even better when considering the topic-specific distribution of complexity. Our work is inspired by these previous findings on expertise as a function of users, topics and content complexity. As such, we will further investigate expertise as a constantly evolving notion, rather than a static property that could be attributed to a user-topic pairing.

There is an extensive body of work on understanding and describing the intentions and motivations that underly the search process. Belkin et al. [4] describe the phenomenon of the *Anomalous State of Knowledge* (ASK) whereby searchers may lack a clear mental representation of their own information needs. Broder [7] proposed a ternary classification of informational, navigational and transactional queries. Following on Broder's work, a multitude of different search intent classification schemes have been presented. Rose and Levinson [27] expand Broder's classification with a number of sub-categories such as directed, undirected or advice intents under the informational needs category. Baeza-Yates et al. [3] identify two fundamental intent dimensions, topic and goal, each of which can take one of a number of different values. Alternative frameworks and criteria have been proposed, including for example the user's intent to buy products [2] or the concrete data types and formats [31] that were searched for. Our work investigates two different knowledge acquisition intents grounded in psychology and cognitive science literature on procedural and declarative memory in order to characterize different learning goals.

Exploratory search is closely related to the knowledge acquisition scenario that we investigate in this paper. Marchionini [22] describes exploratory search as an information seeking scenario where the primary goal is broadly learning about a given topic. As we will see in Section 3, this matches the definition of *declarative information needs*, one of the knowledge acquisition intent classes that we consider. The existing literature on exploratory search focuses primarily on alternative user interfaces and technologies to best support this interaction paradigm. Examples include the use of faceted search interfaces [18], collaborative and multi-user interfaces [13] and topic summarization systems [25]. In this work, we will focus on studying different knowledge acquisition intents and their characteristics based on the log files of a popular Web search engine. Rather than designing a search system to cater for one particular intent class, we aim at furthering our understanding of knowledge acquisition in general.

3. DATA SET & METHODOLOGY

Previous work [36] showed that searchers can, over time, acquire domain expertise as they are exposed to domain-related information. We believe that learning happens all the time when people interact with information items. However, for the sake of this study, we try to select the clearest and most explicit examples of sessions where learning is likely to occur.

3.1 Identifying Example Sessions

Jones et al. [15] suggest that search sessions are often task-driven and dedicated to a satisfying a single overarching information need. We follow the established approach of drawing session boundaries after an inactivity of at least 30 minutes and begin by identifying those web search sessions targeted at knowledge acquisition. To do this we draw on psychological literature which distinguishes between two fundamental types of knowledge: *procedural knowledge* and *declarative knowledge*. Procedural knowledge refers to knowing **how to do** something, whereas declarative knowledge refers to knowing **about** something [1]. In the context of web search we seek to identify search sessions in which people explicitly search for how to do something or to find out about something. These are sub-categories of Broder's informational sessions.

To identify procedural and declarative search sessions, we use a simple 4-step heuristic:

(1) We start by identifying popular procedural and declarative resources. We use `http://ehow.com`, a site that offers more than 2 million tutorial articles and videos to several million visitors daily, as the procedural source. And, we use `http://wikipedia.org`, a site contains more than 4 million encyclopedia pages in English, as the declarative information source. We then identify all queries, Q_p, for which the last click in a session is on `http://ehow.com`, and all those queries, Q_d, for which the last click in a session is on `http://wikipedia.org`.

(2) From these two query sets, we build variable length n-gram language models LM_p and LM_d as described by Niesler and Woodland [24]. We also build a general collection model, LM_G, which we describe in more detail below.

(3) Following the approach used by Tomokiyo and Hurst [35], we determine the point-wise KL-divergence between the general collection model and each of the knowledge acquisition intent-specific models. Equation 1 shows how we compute the KL divergence for each n-gram t. This method lets us identify those n-grams that occur more frequently in the knowledge acquisition intent query pools than in the overall query log. Table 1 shows the top 10 query terms for the two knowledge acquisition intents after removing domain-specific terms such as "wiki" or "ehow" from the candidate list. Intuitively, the term lists are reasonable and they reflect our intended separation between learning how to do something and learning about something.

(4) On the basis of these 10 term sets, we create experimental datasets D_{proc} and D_{decl}, defined as those sessions that contained at least one of the indicator terms. If terms from both lists were present in the session, we assigned the class with the greatest query term coverage. The remainder of our corpus that was not found to be either procedural or declarative will be addressed as D_{other}. Table 2 gives two typical examples of procedural and declarative search sessions obtained from our log files.

3.2 Dataset

Our investigation is based on the log files of a popular web search engine. The data sample used for this study covers the period between February 1st and 28th, 2013. In order to reduce variability introduced by the highly multilingual nature of the Internet and its users, we focus on the English-speaking US market. To concentrate on informational search intent, we remove all purely navigational sessions based on the output of a proprietary query classifier. After this step, we are left with 26.4 million sessions issued by 2.1 million unique users. About 3% of these sessions have an explicit knowledge acquisition intent, falling into either D_{proc} or D_{decl}.

Table 1: Knowledge acquisition intent cue words.

Rank	Procedural	Declarative
1	to	what
2	how	what is
3	how to	who
4	how do	list of
5	to make	syndrome
6	how to make	biography
7	how do I	what is a
8	computer	about
9	can you	is the
10	change	history of

Table 2: Procedural / declarative session examples.

Procedural		
#	Query	Clicked URL
1	Weak wireless signal	-
2	How to expand wifi range	http://www.repeaterstore.com/.../fg24008.php
3	Wifi repeater how to	-
4	Wifi repeater tutorial	http://forum.ubnt.com/showthread.php?t=13735
5	How to boost wifi signal	http://www.ehow.com/...boost-wifi-signal.html

Declarative		
#	Query	Clicked URL
1	What do sponges look like	a-z-animals.com/animals/sponge/
2	What do sponges feed on	tolweb.org/treehouses/?treehouse$_{i}d = 3431$
3	Sponges as pets	http://www.buzzle.com/articles/sponge-facts.html
4	How do sponges reproduce	http://answers.yahoo.com/...127140016AANRe91
5	Where do sponges live	http://en.wikipedia.org/wiki/Sponge

$$KL_{LM_i,LM_G}(t) = P_{LM_i}(t) log \frac{P_{LM_i}(t)}{P_{LM_G}(t)} \quad (1)$$

Closer inspection of the "knowledge acquisition intent" collections showed a high precision despite the simplicity of our method. We manually labeled a sample of 300 D_{proc} and D_{decl} sessions and found that 87% of the selected instances indeed identified procedural or declarative knowledge seeking intents. Since only about 3% of sessions have a clear knowledge acquisition intent, manual labelling efforts to identify additional knowledge acquisition intents were not practical. Instead, we used the existing high-precision collections as bootstrap resources and create expanded collections $D_{proc-ext}$ and $D_{decl-ext}$ by including all sessions in which the URLs clicked in the high-precision sets also appeared.

Table 3 gives an overview of key properties of these five data sets. D_{proc} and D_{decl} show a generally greater tendency for content exploration. For these sets, sessions are longer, and users issue more queries, dwell longer on each result, visit lower-ranked results, and move beyond the first result page more often. We also see a greater topical diversity than was observed for D_{other}. The expanded data sets closely follow the overall behavior of the background collection.

We conducted all experiments described in this paper with the two expanded collections $D_{proc-ext}$ and $D_{decl-ext}$ as well as the two high-precision collections D_{proc} and D_{decl}. The results for $D_{proc-ext}$ and $D_{decl-ext}$ were nearly identical to those for the general background collection. Thus our attempt to expand knowledge acquisition sessions seems to

Table 3: Comparison of experimental data sets.

Dataset	# sessions	queries per session	median dwell time	median session duration	topics/ SERP	max page no	lowest rank clicked	query length
D_{other}	25.6m	1.7	10 sec	43 sec	3.1	1.07	1.3	2.2
D_{proc}	443k	4.1	185 sec	603 sec	4.5	1.20	3.2	5.8
D_{decl}	355k	6.3	287 sec	1003 sec	6.2	1.28	4.1	4.5
$D_{proc-ex}$	3.8m	2.1	17 sec	123 sec	3.4	1.09	1.5	3.1
$D_{decl-ex}$	4.1m	2.4	20 sec	147 sec	3.7	1.09	1.7	2.8

have resulted in more noise than useful knowledge acquisition sessions. For the sake of brevity, we will not discuss these two collections further in this paper. Instead we focus on the two high-precision collections of procedural and declarative knowledge acquisition intents as well as D_{other}, the remainder of the original 26.4m sessions.

3.3 Metrics

As the starting point for our investigation, we consider a set of six metrics that have previously been used to describe domain expertise and search behavior:

Domain Count. White et al. [36] found that domain experts encountered different and more diverse domains than domain novices. They measured diversity by the number of unique domains present on the result pages (SERPs), and we use this measure as well. We also consider two measures of topical diversity (focus and entropy), as described below.

Focus. Focus is concerned with how narrow the topical space is that the user explores. We measure focus as the proportion of entries on the result page that fall into the most frequent observed topical category. We use the output of a text classifier that assigns ODP category labels based on the textual content of web pages. To achieve a degree of detail that has been previously found appropriate, we follow Shen et al. [28] and use the second level of the ODP taxonomy.

Entropy. Where focus describes the degree to which the SERP is covered by just a single topic, entropy is concerned with how diverse the entire topic distribution on the result page is. The higher the entropy, the more diverse is the range of topics on the result page.

Branchiness. Branchiness describes how often the searcher returns to a previously visited point (e.g., a search engine result page or other hub page) and follows a previously unexplored link from there on. We use branchiness to measure how broadly the user explores the available content. Previous work [36] has shown that domain experts have more branchy search paths than domain novices.

Display Time. The amount of time a user spends on average to read retrieved documents has previously been considered an indicator of domain expertise. White et al. [36] found that experts generally spend less time per retrieved web page than novices. They hypothesised that domain experts may be more adept at reading technical content and locating the desired information than domain novices.

Query Complexity. Content complexity is a traditional and intuitive indicator of domain knowledge which manifests in the form of technical jargon or other highly specialized vocabulary unknown to novices. Because our collection covers so many different technical domains it was impractical to collect domain-specific thesauri as others have done for restricted domains such as medicine. Instead we chose to use reading level metrics to characterize changes in the complexity of query terms. Traditional reading level metrics require longer coherent text samples and are not very effective when applied to very short texts. Kuperman et al. [19] compiled a listing of more than 30,000 English words along with the age at which native speakers typically learn the term. The higher this score, the harder and more specialized a term is assumed to be. In order to measure the complexity of queries, we report the maximum age of acquisition across all query terms.

4. WITHIN-SESSION LEARNING

The central goal of our work is to better understand how users acquire new knowledge while they are searching the Web. Previous work on domain expertise investigated long term developments of knowledge across the course of several months worth of search activity [36] or throughout the course of an academic semester during which students were exposed to lecture material [39]. In this section, we start to address a previously unstudied problem by investigating domain expertise at much finer granularity; at the session level.

4.1 Metric Changes Within Sessions

Figure 1 compares within-session changes of all six metrics for our three experimental corpora for 5-query sessions. The same tendencies hold for other lengths, but are omitted for space reasons. We partition the data by session length to rule out external effects such as different dynamics of short and long sessions. All scores are relative to the score observed for the first query in each session.

We observe that focus rises initially but tends to fall quickly for D_{proc} and D_{decl}, evidencing broader exploration than in D_{other} where scores level out. After initial increases, topical entropy decreases for all corpora. Especially for D_{proc} scores plummet dramatically, which we take as evidence for the user narrowing down the topical space when approaching the desired resource. This is further supported by display time scores which rise most dramatically towards the end of procedural sessions. This suggests that the user found the desired information and now studies a single document in depth, potentially replicating its instructions on the spot. Branchiness of sessions tends to fall, with only D_{decl} as an exception which shows the broadest exploration of content.

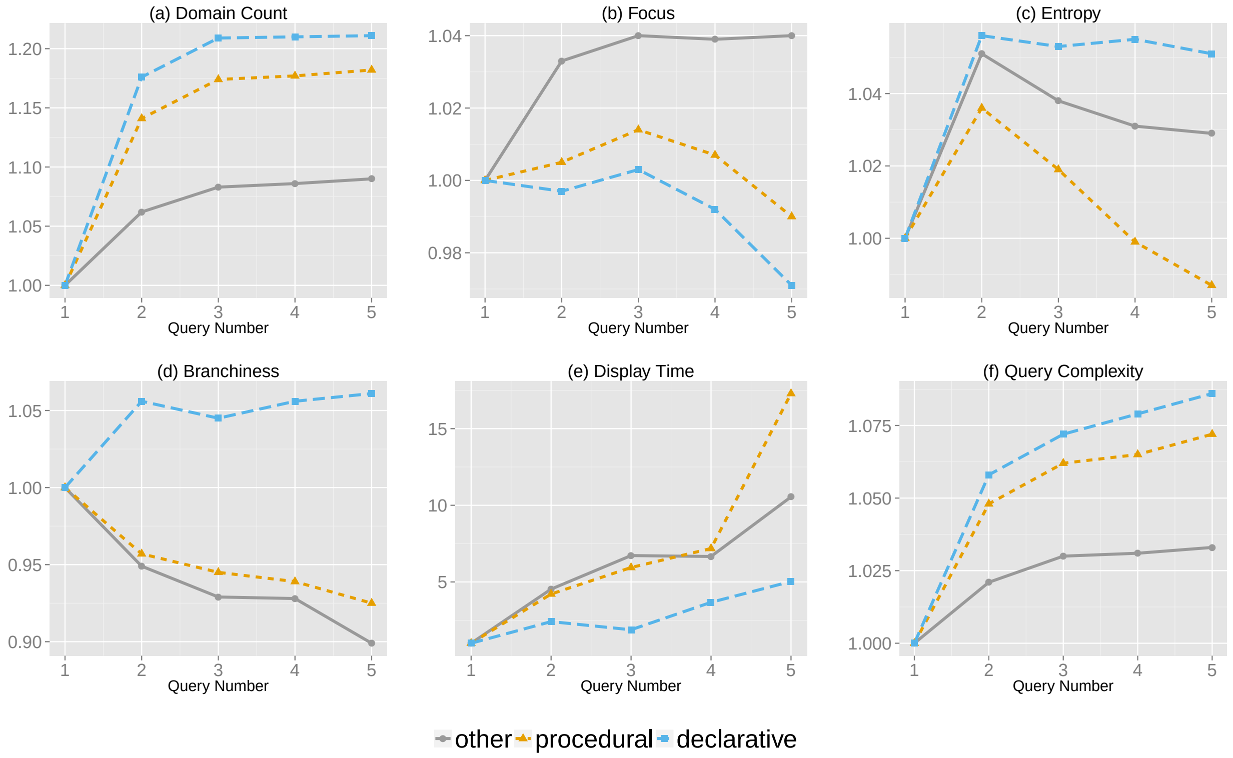

Figure 1: Within-session development of domain count (a), focus (b), entropy (c), branchiness (d), display time (e) and query complexity (f) for sessions of length 5. Scores are relative to the first query per session.

At the same time, the domain count rises for all session types with strongest increases among knowledge acquisition sessions. Query complexity gradually increases for all sessions but knowledge acquisition sessions show much steeper gradients than others. To summarize, for focus, domain count and complexity, knowledge acquisition sessions (D_{proc} and D_{decl}) show strong differences from the remaining collection. Consistently, declarative search sessions show tendencies towards broader, less focused exploration as evidenced by higher entropy, domain count and branchiness.

4.2 Persistence of Learning

Previously, we demonstrated how expertise development can be observed at the session level. We now study how domain expertise acquired within a session is sustained across session boundaries. As a first step in this direction, we investigate how much knowledge is carried over from within-session learning to the following session. We examine this by comparing sessions S_i based on whether or not the directly preceding session S_{i-1} showed a within-session increase in the respective metric. All scores in this comparison are relative to the user's previous average score for this information need. To rule out the effects of task-specific variance, we only compare those sessions that belong to the same task. This condition is ensured by requiring sessions to share at least a 50% overlap in query terms. E.g., if a user has 5 sessions in the task and we find a metric gain in Session 3, we compare the onset (first query) of Session 4 with the average across Sessions 1 - 3. Results of this analysis are shown in Figure 2. The coherence metrics (focus and entropy) show somewhat greater (but not statistically significant) gains for post-gain sessions. For domain count and complexity, the tendencies are more pronounced and we observe significantly greater increases for post-gain sessions. Statistical significance was determined by means of a Wilcoxon signed-rank test with $\alpha \leq 0.01$.

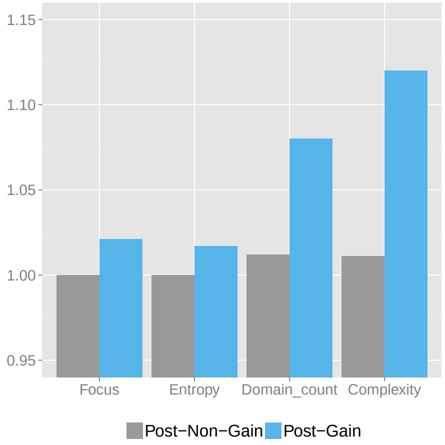

Figure 2: Comparison of domain expertise onset between post-gain sessions and post-non-gain sessions.

Finally, we move beyond the onset of post-gain sessions and investigate evidence of expertise across several sessions within the same overarching task. To do this, we study the cross-session development of expertise for sessions that follow a session in which within-session learning was detected. We contrast these sessions with a sample following on sessions without learning. Figure 3 shows the results of this comparison. As we found for session onsets, focus and entropy show small insignificant changes as a function of whether metrics increased in the previous session. Domain count and query complexity, however, show significantly larger increases for sessions on the same task after within-session increases.

5. PAGE VISITS AS LEARNING CATALYST

Previously, we observed within-session increases of domain expertise that were sustained across session boundaries. In this section, we try to find the cause for these developments.

5.1 Page Visits Change Behavior

To gain a better understanding of the effects and dependencies at play within a search session, we conduct a qualitative analysis of different sub-samples of our experimental corpus. For each of the six metrics, we select the 200 search sessions with the strongest change in the respective metric from the overall dataset D. These sets are contrasted with six sets of 200 randomly sampled sessions in which no significant gains or losses were observed. The first author manually analysed the query log entries for the 2400 selected sessions, trying to identify reasons for the observed changes. From this sample, we note that **(1)** Sessions with clear increases in branchiness contain many medical queries as well as queries related to construction and DIY. **(2)** Procedural sessions tend to show the most pronounced increases in dwell time. This can be explained by the particular mode of interaction that is typical for procedural information. The user searches for a useful resource (e.g., a how-to), once the right resource is located, they will follow the instructions on the page step by step, which may take considerable amounts of time. **(3)** Finally, we observed that there were two major reasons for changes in query complexity. The first is task switching. When the user tries to satisfy different information needs there can be large topic-dependent changes in complexity. More interestingly, however, we noted that within the confines of the same task, page visits often resulted in a change in the complexity of subsequent queries.

5.2 Page Visits and Query Reformulations

It appears that page visits have significant influence on the vocabulary of subsequent queries as users are exposed to new information which they use to modify their queries. To verify this hypothesis, we investigate the origin of newly added query terms. We inspect the proportion of query term additions that can be explained based on snippets or page content of clicked pages. Table 4 shows the proportion of new query terms that were found among the snippets shown on the previous result page or on clicked pages. It also reports the share of newly added terms that were not found in any of the previous resources. The numbers do not sum to 100% per row because the same term could occur in snippets as well as on clicked pages. SERP snippets account for approximately one quarter of all new additions. Nearly half of all newly added terms could be found on pages that the user visited directly before the query reformulation. Long page visits (dwell time $\geq$ 30 seconds) contain about 4 times as many term additions as short visits. Our results confirm the general findings of previous work by Stamou and Kozanidis [30]. In our dataset, the influence of page content is significantly higher than reported previously. This may be due to the fact that earlier work only considered visited pages (not snippet content), or that they considered only 18 selected users.

Table 4: The origin of newly added query terms. Measured in terms of recall of newly added query terms among different resources.

	Snippet	Short page visits	Long page visits ($t \geq 30$ sec)	None
D_{other}	0.27	0.13	0.41	0.52
D_{proc}	0.23	0.12	0.44	0.49
D_{decl}	0.26	0.10	0.45	0.49

5.3 Page Visits and Metrics

Finally, let us quantify the impact that page visits have on the various coherence and expertise metrics. At this point, we work on the basis of the full collections D_{proc}, D_{decl} and D_{other} rather than just the qualitative 200-session samples. For each metric m, we compute the posterior probability of its score increasing by more than one standard deviation $P_{m,+}(.)$, falling by more than one standard deviation $P_{m,-}(.)$ or staying stable $P_{m,=}(.)$. We partition the data by whether the previous result page $SERP_{i-1}$ received no clicks, only short clicks, or at least one long click, represented by its click condition $c \in \{no\ click, short, long\}$.

$$P_{m,+}(c) = \frac{Count(m(SERP_i) > m(SERP_{i-1}) + \sigma_m, c)}{Count(c)}$$

$$P_{m,-}(c) = \frac{Count(m(SERP_i) < m(SERP_{i-1}) - \sigma_m, c)}{Count(c)}$$

$$P_{m,+}(c) = \frac{Count(m(SERP_{i-1}) + \sigma_m \geq m(SERP_i) \geq m(SERP_{i-1}) - \sigma_m, c)}{Count(c)}$$

In order to reduce the influence of task switching during sessions, we require all result page pairs $(SERP_{i-1}, SERP_i)$ to be dedicated to the same overarching task. We measure this by requiring them to be adjacent in time (i.e., there was no other query in between any pair $SERP_{i-1}, SERP_i$) and to share at least a 50% overlap in query terms. Table 5 shows the results of this comparison. In order to interpret the scores, we have to compare analogous changes across click conditions. E.g., $P_{m,+}(c = non)$ to $P_{m,+}(c = short)$ in order to see how short clicks influence the likelihood of a subsequent increase in m as compared to unclicked SERPs. Statistically significant changes with respect to unclicked SERPs are denoted by the △ character for increases and the ▼ character for decreases. Statistical significance was tested by means of a Wilcoxon signed-rank test at $\alpha \leq 0.01$-level. We observe a number of fundamental tendencies: (1) Topical diversity tends to increase as the result of a click (P_+ falls while $P_=$ and P_- rise). Similarly, the entropy in the topic distribution increases (P_+ and $P_=$ increase while P_- falls with respect to unclicked search result pages). This suggests that exposure to new information offered on visited pages diversifies subsequent queries by introducing new influences. (2) The number of unique domains per SERP also tends to increase in response to a click (P_+ and $P_=$ increase, P_- falls). This finding supports the diversifying effect observed in topical space. (3) After a click, the complexity of

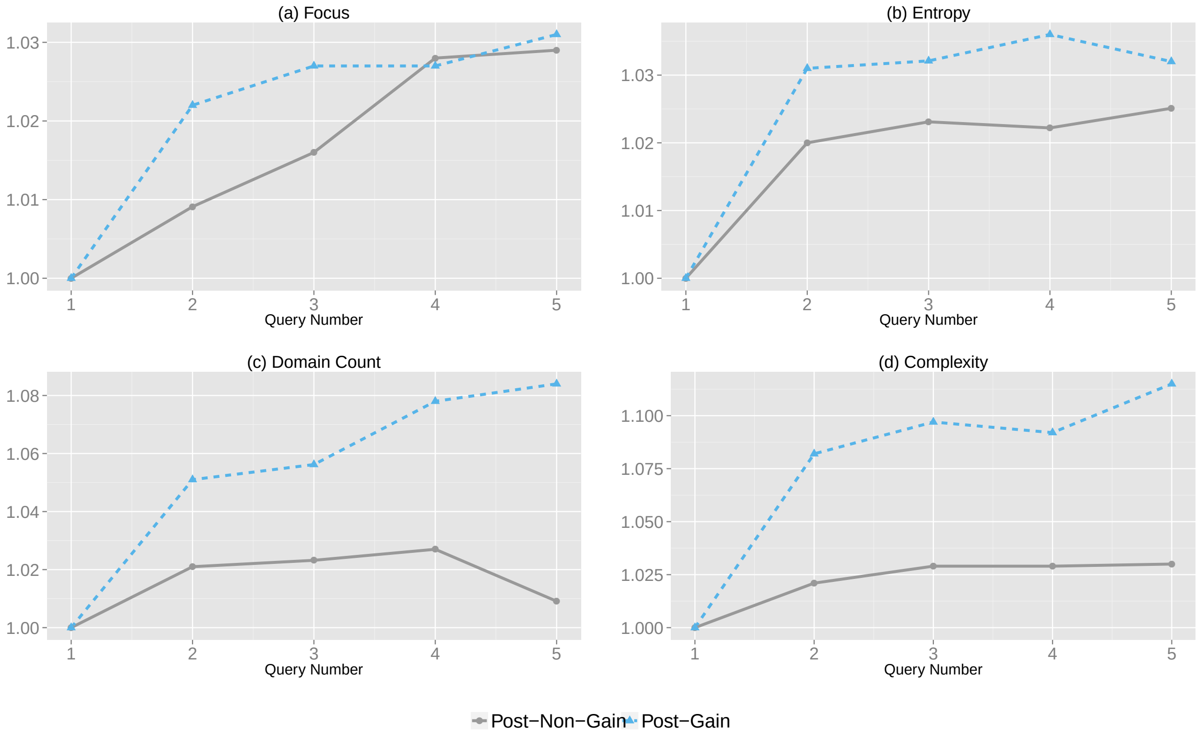

Figure 3: The development of focus (a), entropy (b), domain count (c) and query complexity (d) as a consequence of within-session learning. All numbers are relative to the first session per task.

subsequent queries increases more than it does for unclicked SERPs (P_+ and $P_=$ increase, P_- falls).

The findings are largely consistent across collections, with D_{other} showing less pronounced tendencies than D_{proc} and D_{decl}. For most metrics, there were no significant differences between long and short page visits. It seems that even short visits introduce sufficient amounts of new information to influence subsequent queries. It should be noted that the tendencies shown for branchiness and page display time (both increase most dramatically after an unclicked search result page) are necessarily biased since both metrics depend on clicks. E.g., there is no chance for a display time of 0 in the case of an unclicked result page to still fall.

At this point, our experiments can be expected to include a moderate amount of chance variance. Especially in D_{other} we observe seemingly arbitrary drops and increases in various metrics that influence the resulting aggregates in Table 5. This can be attributed to chance drops in metrics as a consequence of e.g., exchanging synonymous query terms or visiting near duplicates of previously seen results which blur our observations. In order to mitigate such effects, we further restrict our analysis by considering only sustained changes in metrics. This means, that only those sessions in which changes are sustained throughout the whole session, contribute to the computation of posterior probabilities. Take for example a query complexity increase from a score of 5.4 to 7.1 between $SERP_{i-1}$ and $SERP_i$. Where previously, we would have directly counted this towards P_+, now, we only do so if the respective score never falls below 7.1 for all further result pages $SERP_{i+1}, SERP_{i+2} \ldots R_I$. The normalization component $Count(c)$ is adjusted to this new definition accordingly. Table 6 shows the outcome of this altered experimental setup. We observe the same general tendencies, but the differences between clicked and unclicked SERPs are more pronounced.

6. PREDICTING CLICK IMPORTANCE

In the previous sections, we showed evidence of within-session learning and its continuation across session boundaries. In an effort to explain the observed domain expertise gains, we found a connection between page visits and subsequent learning. In the final experimental contribution of this paper, we now predict which page visits are most likely to advance a user's domain expertise. To this end, we begin with a brief description of the features used to represent Web documents:

Text length The overall amount of Web page text has been previously reported as an indicator for resource complexity [23]. We measure text length as the overall number of words on the page.

Sentence length Similarly, long sentences were found to indicate content complexity [23]. Sentence length is measured as the average number of words per sentence.

Term length Long words are an indicator for specific and complex vocabulary [23]. We measure term length as the average number of characters as well as the average number of syllables per term on the page.

Table 5: Post-click trends.

		Focus			Entropy			Branchiness			Display Time			Domain Count			Complexity		
		P_+	$P_=$	P_-	P_+	$P_=$	P_-	P_+	$P_=$	P_-	P_+	$P_=$	P_-	P_+	$P_=$	P_-	P_+	$P_=$	P_-
D_{other}	Non	0.25	0.45	0.30	0.24	0.47	0.29	0.44	0.56	0.00	0.65	0.35	0.00	0.29	0.36	0.35	0.17	0.66	0.17
	Short	$0.16^{\blacktriangledown}$	$0.60^{\triangle}$	$0.24^{\blacktriangledown}$	$0.29^{\triangle}$	$0.59^{\triangle}$	$0.12^{\blacktriangledown}$	$0.25^{\blacktriangledown}$	$0.24^{\blacktriangledown}$	$0.51^{\triangle}$	$0.31^{\blacktriangledown}$	0.35	$0.34^{\triangle}$	$0.22^{\blacktriangledown}$	$0.56^{\triangle}$	$0.22^{\blacktriangledown}$	$0.21^{\triangle}$	$0.60^{\blacktriangledown}$	$0.19^{\triangle}$
	Long	0.23	$0.49^{\triangle}$	0.28	$0.30^{\triangle}$	0.49	$0.21^{\blacktriangledown}$	$0.28^{\blacktriangledown}$	$0.25^{\blacktriangledown}$	$0.47^{\triangle}$	$0.18^{\blacktriangledown}$	0.34	$0.48^{\triangle}$	$0.33^{\triangle}$	0.37	$0.30^{\blacktriangledown}$	$0.23^{\triangle}$	0.67	$0.10^{\blacktriangledown}$
D_{proc}	Non	0.34	0.32	0.34	0.30	0.35	0.35	0.43	0.57	0.00	0.69	0.31	0.00	0.38	0.16	0.46	0.33	0.37	0.30
	Short	$0.30^{\blacktriangledown}$	$0.35^{\triangle}$	0.35	0.31	$0.39^{\triangle}$	$0.30^{\blacktriangledown}$	$0.24^{\blacktriangledown}$	$0.21^{\blacktriangledown}$	$0.55^{\triangle}$	$0.25^{\blacktriangledown}$	$0.35^{\triangle}$	$0.40^{\triangle}$	$0.43^{\triangle}$	0.17	$0.40^{\blacktriangledown}$	0.35	$0.42^{\triangle}$	$0.23^{\blacktriangledown}$
	Long	$0.31^{\blacktriangledown}$	$0.35^{\triangle}$	0.34	$0.33^{\triangle}$	0.37	$0.30^{\blacktriangledown}$	$0.21^{\blacktriangledown}$	$0.27^{\blacktriangledown}$	$0.52^{\triangle}$	$0.19^{\blacktriangledown}$	$0.37^{\triangle}$	$0.44^{\triangle}$	$0.46^{\triangle}$	0.15	$0.39^{\blacktriangledown}$	$0.38^{\triangle}$	0.37	$0.25^{\blacktriangledown}$
D_{decl}	Non	0.36	0.30	0.34	0.31	0.35	0.34	0.41	0.59	0.00	0.71	0.29	0.00	0.38	0.15	0.47	0.32	0.36	0.32
	Short	$0.31^{\blacktriangledown}$	$0.33^{\triangle}$	0.36	0.31	$0.38^{\triangle}$	$0.31^{\blacktriangledown}$	$0.24^{\blacktriangledown}$	$0.20^{\blacktriangledown}$	$0.56^{\triangle}$	$0.31^{\blacktriangledown}$	$0.35^{\triangle}$	$0.44^{\triangle}$	$0.43^{\triangle}$	$0.17^{\triangle}$	$0.40^{\blacktriangledown}$	$0.36^{\triangle}$	$0.40^{\triangle}$	$0.24^{\blacktriangledown}$
	Long	$0.30^{\blacktriangledown}$	$0.35^{\triangle}$	0.35	$0.34^{\triangle}$	$0.38^{\triangle}$	$0.28^{\blacktriangledown}$	$0.27^{\blacktriangledown}$	$0.25^{\blacktriangledown}$	$0.48^{\triangle}$	$0.18^{\blacktriangledown}$	$0.37^{\triangle}$	$0.45^{\triangle}$	$0.44^{\triangle}$	0.16	$0.40^{\blacktriangledown}$	$0.37^{\triangle}$	0.37	$0.26^{\blacktriangledown}$

Table 6: Sustained post-click trends.

		Focus			Entropy			Branchiness			Display Time			Domain Count			Complexity		
		P_+	$P_=$	P_-	P_+	$P_=$	P_-	P_+	$P_=$	P_-	P_+	$P_=$	P_-	P_+	$P_=$	P_-	P_+	$P_=$	P_-
D_{other}	Non	0.28	0.44	0.28	0.23	0.49	0.28	0.43	0.57	0.00	0.71	0.29	0.00	0.29	0.36	0.35	0.21	0.67	0.12
	Short	$0.11^{\blacktriangledown}$	$0.59^{\triangle}$	0.30	$0.29^{\triangle}$	$0.56^{\triangle}$	$0.15^{\blacktriangledown}$	$0.25^{\blacktriangledown}$	$0.22^{\blacktriangledown}$	$0.53^{\triangle}$	$0.28^{\blacktriangledown}$	$0.35^{\triangle}$	$0.37^{\triangle}$	$0.32^{\triangle}$	$0.50^{\triangle}$	$0.18^{\blacktriangledown}$	$0.25^{\triangle}$	$0.62^{\blacktriangledown}$	0.13
	Long	$0.20^{\blacktriangledown}$	0.46	$0.34^{\triangle}$	$0.32^{\triangle}$	0.49	$0.19^{\blacktriangledown}$	$0.28^{\blacktriangledown}$	$0.21^{\blacktriangledown}$	$0.51^{\triangle}$	$0.17^{\blacktriangledown}$	$0.32^{\triangle}$	$0.51^{\triangle}$	$0.37^{\triangle}$	$0.44^{\triangle}$	$0.19^{\blacktriangledown}$	$0.27^{\triangle}$	0.65	$0.08^{\blacktriangledown}$
D_{proc}	Non	0.35	0.33	0.32	0.29	0.34	0.37	0.44	0.56	0.00	0.74	0.26	0.00	0.42	0.18	0.40	0.31	0.39	0.30
	Short	$0.25^{\blacktriangledown}$	0.32	$0.43^{\triangle}$	$0.32^{\triangle}$	$0.37^{\triangle}$	$0.31^{\blacktriangledown}$	$0.21^{\blacktriangledown}$	$0.22^{\blacktriangledown}$	$0.57^{\triangle}$	$0.28^{\blacktriangledown}$	$0.34^{\triangle}$	$0.38^{\triangle}$	0.44	0.19	$0.37^{\blacktriangledown}$	$0.35^{\triangle}$	$0.42^{\triangle}$	$0.23^{\blacktriangledown}$
	Long	$0.18^{\blacktriangledown}$	$0.35^{\triangle}$	$0.47^{\triangle}$	$0.34^{\triangle}$	$0.39^{\triangle}$	$0.27^{\blacktriangledown}$	$0.22^{\blacktriangledown}$	$0.26^{\blacktriangledown}$	$0.52^{\triangle}$	$0.22^{\blacktriangledown}$	$0.34^{\triangle}$	$0.44^{\triangle}$	$0.46^{\triangle}$	0.19	$0.35^{\blacktriangledown}$	$0.36^{\triangle}$	$0.44^{\triangle}$	$0.20^{\blacktriangledown}$
D_{decl}	Non	0.36	0.30	0.34	0.34	0.36	0.30	0.47	0.53	0.00	0.68	0.32	0.00	0.41	0.16	0.43	0.31	0.38	0.31
	Short	$0.30^{\blacktriangledown}$	$0.33^{\triangle}$	$0.37^{\triangle}$	0.36	$0.38^{\triangle}$	$0.26^{\blacktriangledown}$	$0.24^{\blacktriangledown}$	$0.21^{\blacktriangledown}$	$0.55^{\triangle}$	$0.26^{\blacktriangledown}$	0.35	$0.39^{\triangle}$	$0.44^{\triangle}$	0.17	$0.39^{\blacktriangledown}$	$0.36^{\triangle}$	$0.40^{\triangle}$	$0.24^{\blacktriangledown}$
	Long	$0.27^{\blacktriangledown}$	$0.34^{\triangle}$	$0.39^{\triangle}$	$0.37^{\triangle}$	0.36	$0.27^{\blacktriangledown}$	$0.33^{\blacktriangledown}$	$0.27^{\blacktriangledown}$	$0.40^{\triangle}$	$0.18^{\blacktriangledown}$	$0.37^{\triangle}$	$0.45^{\triangle}$	$0.45^{\triangle}$	$0.19^{\triangle}$	$0.36^{\blacktriangledown}$	0.32	$0.42^{\triangle}$	$0.26^{\blacktriangledown}$

Coverage of query terms in title The presence or absence of query terms in the page title may give an insight in the nature of the page. I.e., does the page only mention the topic (low expected learning potential) or is it centrally concerned with the topic (high expected learning potential)?

Distribution of query terms Following a similar intuition, we measure the distribution of query terms on the page. Concentrated term occurrences may signal that the query topic is only one of many aspects of the page, whereas when query terms are spread across the whole page they may represent the page's core topic. To this end, we measure the proportion of page text that lies between the first and the last occurrence of query terms as well as the median distance between query terms.

POS distribution The syntactic structure of the page may give clues about its intention and complexity [8]. To reflect this, we include four features representing the relative share of page texts falling into the Parts-of-speech *noun*, *verb*, *adjective*, or *other*.

Page complexity One of the most direct indications of page complexity can be achieved by measuring the age of acquisition across all terms on the page. Similarly, Kuperman et al. [19] offer statistics of the percentage of adult native speakers that know each corpus term. The lower this number, the more specific and complex is the language that was used. We report page-wide averages for both figures.

Page complexity vs. query complexity All previous features treated page complexity in isolation. We assume, however, that the relative complexity with regard to the user's current state of knowledge will have strong influences on the learning potential. To account for this, we report the ratio of page complexity to query complexity.

Our investigation is based on a balanced sample of 50,000 data points. For 50% of the data subsequent learning was detected. The remaining half shows no evidence of learning. A stratified sample of 10,000 instances is withheld as a test set. Each instance represents a page visit and is annotated with whether or not the visit was immediately followed by an increase in both query complexity and domain count. We evaluated several state-of-the-art classifiers. The overall best performance of $F_1 = 0.76$ and an area under ROC curve of 0.81 is achieved by a Support Vector Machine (SVM) with Pearson Universal kernel ($\omega = 1.2, \sigma = 1.0, \epsilon = 10^{-12}, c = 1.0$). The parameter settings were determined by means of a greedy parameter sweep. To further understand the separation between those clicks that lead to an increase in expertise metrics and those that do not, we compare the mutual information between each of the above features and the class label. We find the strongest connection in features that contrast page and query complexity, effectively expressing how close the complexity of the user's produced vocabulary is to that of the page. We generally find the highest learning potential for pages that are slightly more complex than the user's observed active vocabulary.

On the basis of our classifier, we now analyze the distribution of learning potential across the ranks of search result pages. Ideally, for knowledge acquisition sessions we would require the documents with the highest likelihood of advancing the user's state of knowledge to be ranked highest. Note that we have to rely on predictor output for this experiment since log-based examples of learning can only be observed as a consequence of clicks. Clicks in turn have been frequently reported to be affected by position biases [14]. Figure 4 shows the distribution of predicted learning potential across SERP ranks. We note only a weak correlation between search result page ranks and learning potential. Integration of our classifier as a ranking criterion might help to optimize the learning potential of result lists.

7. DISCUSSION

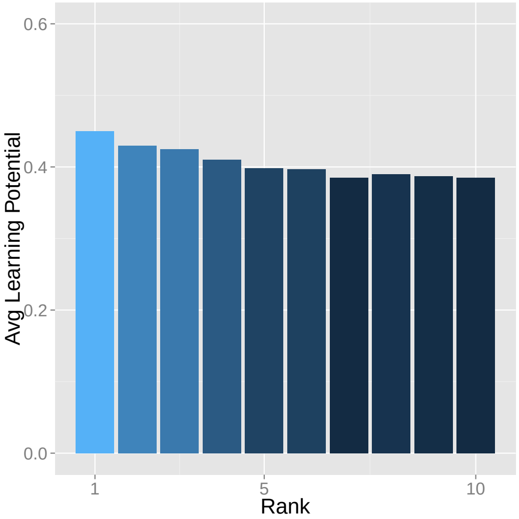

Figure 4: Distribution of predicted learning potential across the ranks of the search result page.

We have seen that people's behavior changes over the course of a search session in a way that suggests they learn as they search, and observed that what they learn appears to persist across sessions. Further, it is possible to predict which clicks and sessions will result in the greatest change in behavior. In this section we discuss the ways these findings can be used to support users and optimize the search experience to promote knowledge acquisition.

By recognizing the different stages of learning in a search session we can reinforce existing knowledge acquisition behavior. For example, we observe that the number of domains on a search result page increases over the course of a session, particularly for procedural and declarative queries. We could reinforce this outcome by encouraging greater search result diversity for queries later in a session when it does not happen naturally. Likewise, we see that the query terminology gets more sophisticated as a session progresses. The query suggestions offered to users could, when a user first embarks on a new topic, start out simple and increase in complexity over time. Because people often use terms from the search result snippets in their queries as they search, the best query suggestions may be ones that are biased towards including terms that occurred in the snippets of results that are visited and dwelt on.

These approaches could be applied to all search sessions, since many of the changes we observed were true for general queries. However, we also saw that it was possible to automatically identify queries that indicate procedural or declarative information needs, and that these sessions exhibit particularly strong and sometimes unique learning behavior. For example, while search result entropy increases for most queries, it drops sharply towards the end of a session for procedural queries. Rather than encouraging topical diversity later in a session for such queries, it may make sense instead to help the user converge and enforce greater consistency among the search results.

In addition to reinforcing observed behavior, our findings could also be used to promote faster learning, helping people get to the information that will change their search behavior earlier than they might otherwise. For example, we found that we could easily and accurately identify results that promote learning. This information could be used to re-rank the search results so that the top ranked results promote learning. This may be particularly useful early in a search session and for new topics that the searcher engages with, as an increase in knowledge acquired early on may encourage future learning in subsequent queries and sessions.

It may also be possible to promote faster learning by augmenting the search interface. We saw that it is possible to identify terms that will be important to the user because they appear in the clicked search result snippets. These terms could be defined on the search result page to help improve the vocabulary available to the searcher. Additionally, rather than requiring a user to actually visit a page that will help improve their ability to search on a topic, valuable information could be extracted from those pages and provided alongside the results. The approach used by Bernstein et al. [6] to extract short answer text for in-line search result answers could be applied in this case.

As search engines get better, and people are able to find what they are looking for with less need to hunt around, there is a risk that they will also learn less during the search process. Our findings can be used to ensure this does not happen. In addition to providing the end content, search engines can provide important, intermediate information like the clicked results studied in Section 4.

8. CONCLUSION

In this paper, we investigated evidence of users' within-session knowledge acquisition based on the log files of a popular Web search engine. During several qualitative and quantitative analyses, we made the following observations: **(1)** Information seeking sessions can be divided into procedural and declarative sessions, both of which show significant differences from sessions without explicit knowledge acquisition intent as well as from each other. Many information seeking sessions do not fall into either of these categories. They do, however, represent a high-precision set of cases in which people are explicitly seeking to acquire knowledge. As such, they allow us to observe knowledge acquisition via Web search in great detail. **(2)** Based on an automatically extracted set of procedural and declarative search sessions, we studied the development of domain expertise and find evidence of within-session learning that is sustained across sessions as well. **(3)** We gained initial insights into the reason for the observed learning. By tracking the origin of query reformulations, we show that significant proportions of newly added query terms had been previously present on result page snippets and recently visited pages. This suggests that the search process, not just the final result, contributes towards expanding the user's domain knowledge. **(4)** Finally, we develop a classifier based on document and session-level features that was able to accurately predict the knowledge acquisition potential of Web pages for a given user.

There are several promising directions for future work on Web search-driven knowledge acquisition. Firstly, it will be interesting to investigate the potential of using information about the user's knowledge acquisition intent for ranking purposes. Evidence such as the output of the predictor that was presented in Section 6 may be able to improve the ranking quality of knowledge search sessions. Similarly, we will investigate how confirmed knowledge acquisition intents can be supported on an interface level by providing dedicated tools appropriate for the intent in question. This could, for example, include providing high-level overviews and summaries of relevant concepts in the domain in the case of declarative information needs which are often exploratory in nature. Finally, in Section 4, we showed that future query terms are often present on previously visited pages. In this paper, we have no means of knowing whether the searcher actually saw the term on the page or whether it has another origin. A dedicated eye-tracking study of query reformulation can be expected to create much further insight into this interesting question.

9. REFERENCES

[1] John R Anderson. *Language, Memory, and Thought.* Earlbaum, 1976.

[2] Azin Ashkan, Charles LA Clarke, Eugene Agichtein, and Qi Guo. Classifying and characterizing query intent. In *ECIR 2009.*

[3] Ricardo Baeza-Yates, Liliana Calderón-Benavides, and Cristina González-Caro. The intention behind web queries. In *SPIRE 2006.*

[4] Nicholas J Belkin, Robert N Oddy, and Helen M Brooks. Ask for information retrieval: Part i. background and theory. *Journal of documentation*, 1982.

[5] Paul N Bennett, Ryen W White, Wei Chu, Susan T Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. Modeling the impact of short-and long-term behavior on search personalization. In *SIGIR 2012.*

[6] Michael S Bernstein, Jaime Teevan, Susan Dumais, Daniel Liebling, and Eric Horvitz. Direct answers for search queries in the long tail. In *SIGCHI 2012.*

[7] Andrei Broder. A taxonomy of web search. In *ACM SIGIR Forum 2002.*

[8] Jamie Callan and Maxine Eskenazi. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *NAACL HLT 2007.*

[9] Kevyn Collins-Thompson, Paul N Bennett, Ryen W White, Sebastian de la Chica, and David Sontag. Personalizing web search results by reading level. In *CIKM 2011.*

[10] Maurice de Kunder. Daily estimated size of the world wide web. http://www.worldwidewebsize.com/, 2013.

[11] Carsten Eickhoff, Kevyn Collins-Thompson, Paul N. Bennett, and Susan T. Dumais. Personalizing atypical web search sessions. In *WSDM 2013.*

[12] Carsten Eickhoff, Pieter Dekker, and Arjen P de Vries. Supporting children's web search in school environments. In *IIiX 2012.*

[13] Gene Golovchinsky, John Adcock, Jeremy Pickens, Pernilla Qvarfordt, and Maribeth Back. Cerchiamo: a collaborative exploratory search tool. *Proceedings of Computer Supported Cooperative Work (CSCW)*, 2008.

[14] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR 2005.*

[15] Rosie Jones and Kristina Lisa Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *CIKM 2008.*

[16] Jin Young Kim, Kevyn Collins-Thompson, Paul N Bennett, and Susan T Dumais. Characterizing web content, user interests, and search behavior by reading level and topic. In *WSDM 2012.*

[17] Aris Kosmopoulos, Eric Gaussier, Georgios Paliouras, and Sujeevan Aseervatham. The ecir 2010 large scale hierarchical classification workshop. In *ACM SIGIR Forum*, 2010.

[18] Bill Kules, Robert Capra, Matthew Banta, and Tito Sierra. What do exploratory searchers look at in a faceted search interface? In *2009 ACM/IEEE-CS joint conference on Digital libraries.*

[19] Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 2012.

[20] Lin Li, Zhenglu Yang, Botao Wang, and Masaru Kitsuregawa. Dynamic adaptation strategies for long-term and short-term user profile to personalize search. In *Advances in Data and Web Management.* 2007.

[21] Jingjing Liu, Nicholas J Belkin, Xiangmin Zhang, and Xiaojun Yuan. Examining usersï£¡ knowledge change in the task completion process. *Information Processing & Management*, 2012.

[22] Gary Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 2006.

[23] G Harry McLaughlin. Smog grading: A new readability formula. *Journal of Reading*, 1969.

[24] Thomas R Niesler and PC Woodland. A variable-length category-based n-gram language model. In *ICASSP 1996.*

[25] Brendan O'Connor, Michel Krieger, and David Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*, 2010.

[26] James Pitkow, Hinrich Schütze, Todd Cass, Rob Cooley, Don Turnbull, Andy Edmonds, Eytan Adar, and Thomas Breuel. Personalized search. *Communications of the ACM*, 2002.

[27] Daniel E Rose and Danny Levinson. Understanding user goals in web search. In *WWW 2004.*

[28] Dou Shen, Jian-Tao Sun, Qiang Yang, and Zheng Chen. A comparison of implicit and explicit links for web page classification. In *WWW 2006.*

[29] Mirco Speretta and Susan Gauch. Personalized search based on user search histories. In *2005 IEEE/WIC/ACM International Conference on Web Intelligence.*

[30] Sofia Stamou and Lefteris Kozanidis. Impact of search results on user queries. In *International Workshop on Web Information and Data Management 2009.*

[31] Shanu Sushmita, Benjamin Piwowarski, and Mounia Lalmas. Dynamics of genre and domain intents. In *Information Retrieval Technology.* 2010.

[32] Chenhao Tan, Evgeniy Gabrilovich, and Bo Pang. To each his own: personalized content selection based on text comprehensibility. In *WSDM 2012.*

[33] Jaime Teevan, Susan T Dumais, and Eric Horvitz. Personalizing search via automated analysis of interests and activities. In *SIGIR 2005.*

[34] Jaime Teevan, Susan T Dumais, and Eric Horvitz. Potential for personalization. *ACM TOCHI 2010.*

[35] Takashi Tomokiyo and Matthew Hurst. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18.*

[36] Ryen W. White, Susan T. Dumais, and Jaime Teevan. Characterizing the influence of domain expertise on web search behavior. In *WSDM 2009.*

[37] Ryen W White and Jeff Huang. Assessing the scenic route: measuring the value of search trails in web logs. In *SIGIR 2010.*

[38] Barbara M Wildemuth. The effects of domain knowledge on search tactic formulation. *JASIST 2004.*

[39] Zahide Yildirim, M Yasar Ozden, and Meral Aksu. Comparison of hypermedia learning and traditional instruction on knowledge acquisition and retention. *The Journal of Educational Research*, 2001.

[40] Xiangmin Zhang, Michael Cole, and Nicholas Belkin. Predicting users' domain knowledge from search behaviors. In *SIGIR 2011.*