Personalizing Atypical Web Search Sessions

Carsten Eickhoff
Delft University of Technology
Delft, The Netherlands
c.eickhoff@tudelft.nl

Kevyn Collins-Thompson
Paul N. Bennett
Susan Dumais
Microsoft Research
Redmond, WA, USA
{kevynct,pauben,sdumais}@microsoft.com

ABSTRACT

Most research in Web search personalization models users as static or slowly evolving entities with a given set of preferences defined by their past behavior. However, recent publications as well as empirical evidence suggest that for a significant number of search sessions, users diverge from their regular search profiles in order to satisfy atypical, limited-duration information needs. In this work, we conduct a large-scale inspection of real-life search sessions to further understand this scenario. Subsequently, we design an automatic means of detecting and supporting such atypical sessions. We demonstrate significant improvements over state-of-the-art Web search personalization techniques by accounting for the typicality of search sessions. The proposed method is evaluated based on Web-scale search session data spanning several months of user activity.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: Models and Principles—*Human Factors*; H.3 [**Information Systems**]: Information Storage and Retrieval

Keywords

Personalized search; user modeling; adaptive interfaces; domain expertise.

1. INTRODUCTION

In recent years, we have seen a strong emerging tendency towards personalizing users' Web search experiences in order to better account for the searcher's individual context [29]. Context, in this case, is often understood as the searcher's previous search history, geo-spatial position, topical interests, and language or literacy background. Most of these are static or slowly-evolving properties of an individual and are typically captured by means of query and interaction log analyses. While personalization functionality has been shown to benefit retrieval performance [29], there are significant situational factors that can influence performance and thus should be taken into account. Domain expertise is one such factor. Depending on the topic searched for, an individual can display significantly different search behavior based on their previous knowledge of the domain at hand [33]. In this work, we investigate instances of users straying from their search profiles to satisfy information needs outside their regular areas of interest. Such atypical information needs can often be triggered by external events (*e.g.*, pending medical treatments, financial deadlines, or upcoming vacations) that explain the unprecedented interest in a previously unseen domain. As an example, a user might in general favor easily-readable documents about sports and be confident in querying, selecting, and understanding this type of information. At the same time, they might display significantly different preferences and skills when pursuing a novel task, such as completing a particularly involved tax form. Due to static modeling of user profiles, atypical information needs are currently poorly represented by Web search engines. Personalizing atypical search sessions in the 'regular' way does not seem appropriate as it assumes topical and behavioral consistency with previous search sessions. Often, this is not the case for atypical searches.

This work advances the state of the art by answering the following three research questions: **(1)** What is the frequency, extent and success rate of users pursuing atypical information needs? Additionally, can we identify common types of information needs for which users diverge from their previous preferences? **(2)** How can we automatically distinguish atypical search sessions from typical ones? **(3)** Can we improve retrieval performance for atypical sessions by re-ranking search results in a typicality-aware fashion? Our investigations are based on manually-annotated log files of the Bing commercial Web search engine.

The remainder of this work is structured as follows: Section 2 gives an overview of previous work on search personalization and expertise modeling. Section 3 introduces the problem domain by investigating the frequency and extent of atypical Web search sessions. Section 4 introduces an automatic method for detecting atypical sessions based on query and interaction logs. Section 5 quantifies the usefulness of applying typicality-aware personalization based on a large-scale query log evaluation, showing significant performance gains at Web scale. In Section 6, we discuss potential extensions of our method and directions for future work. Finally, we conclude by summarizing our findings and their implications on state-of-the-art Web search in Section 7.

WSDM'13, February 4–8, 2012, Rome, Italy.

2. RELATED WORK

Previous related work can be grouped into three categories: **(1)** general search personalization efforts; **(2)** query log analysis with the goal of long-term user modeling; **(3)** investigations of user expertise and content readability.

2.1 Search personalization

A growing number of data sources, such as search history, manually or automatically created preference profiles, and social network information, are being exploited for personalizing the selection of results for users [24, 30]. Approaches to search personalization vary in types of features considered (*e.g.*, language models, topical categories, links or other metadata such as reading level), the time frame chosen (*e.g.*, short-term or long-term profiles), and how the profiles are used (*e.g.*, for ranking or recommendations). Several researchers have shown how profiles that consist of topical representations of users' search interests can be used to personalize search. Gauch and colleagues learned topical user profiles based on browsing history [11] or search history [25], and Ma *et al.* [21] used topical profiles that users specified explicitly. In all cases, user profiles were compared with those of search results and used to re-order search results for individuals. Bennett *et al.* [5] have recently shown how topics from the *Open Directory Project's* (ODP) topology of the Web[1] could be used to personalize search ranking for individuals. Expanding on the PageRank algorithm, Haveliwala [14] proposed a topic-sensitive modification to allow for a direct, more focused scoring of the Web graph given a query's topic or a user's topical preference. Queries and Web content were automatically categorized using the ODP hierarchy in order to facilitate topic-sensitive scoring. Sugiyama *et al.* [26] employed collaborative filtering on users' observed Web search histories for profile building. They compared their approach to exclusively using browsing history and implicit feedback mechanisms, finding significant merit in the use of profile expansion via their proposed method. Teevan *et al.* [29] investigated the potential of re-ranking the top 50 search engine results based on previous user profiles. In particular, the authors explored alternative document representations for search personalization, finding that full-text representations outperformed priority-driven selective keyword models. Li *et al.* [20] developed a dynamic graph-based adaptation scheme modeling a user's general preferences for search personalization, while accounting for changes of interest by incorporating short-term browsing information. In a recent study, Goel *et al.* [12] analyzed the U.S. market's large-scale consumption of movies, music, and Web search results in order to quantify the importance of the 'long tail' of items in those respective categories of popular media. The authors found that the majority of users largely displayed standard tastes in most categories but showed some degree of eccentricity in choices. In this work, we will investigate a related notion, namely, that of atypical search sessions: cases in which users occasionally stray from their personal previous mainstream.

2.2 Long-term user modeling

A special subclass of research on user modeling and search personalization is based on long-term profiles rather than focusing only on the user's immediate history. While being noisier than short-term profiles, this approach has the advantage of being able to detect niche interests or those that surface in long cycles. Matthijs and Radlinski [22] captured users' 3-month Web history across multiple search engines and sites via a browser plug-in. The full resulting log files were used for result re-ranking and showed significant performance improvements over the native ranking of popular search engines. Furthermore, long-term user profiles served as reliable general descriptors of a user's interests. Tan *et al.* [27] presented a language modeling approach that interpolated immediate search history and long-term user profiles in order to improve retrieval performance. They found that short-term profiles contained more useful clues as to the current query's intent, but that adequately-weighted long-term information introduced further performance gains. White *et al.* [32] investigated the usefulness of short-, mid-, and long-term profiles for the task of predicting user interest in Web sites. The authors demonstrated that, depending on the type of information being profiled as well as the type of information need, different profiling durations could be optimal. Finally, Bennett *et al.* [5] showed how long- and short-term profiles could be optimally combined for effective search personalization. They found that long-term models provided the most benefit at the beginning of a session, while short-term models became more important for longer sessions.

2.3 Expertise

Studies of Web search often distinguish between two types of expertise - search expertise (reflecting knowledge of the search process) and domain expertise (reflecting knowledge of the domain or topic of the information need). In one of the early comparisons of Internet information search behavior and success, Hölscher and Strube [15] examined both search and domain expertise. They reported that search experts displayed a richer set of skills, such as selection of tools, query formulation and relevance judgment than novice searchers. Also, experts were found to navigate search interfaces more efficiently. Beyond search skills, Thatcher [31] showed that experts and non-experts followed different strategies to obtain search results, depending on the task. White and Morris [34] conducted a large-scale log analysis of the differences in search behaviors and success of search experts and novices. The authors found that experts generated different types of queries, had shorter and less branchy post-search browse trails, and were generally more successful than novices. More recent work has tried to model strategies of successful searchers. Ageev *et al.* [1] exploited this expertise-dependent difference in search behavior by using a Markov chain approach to predict search success for a range of pre-defined search tasks based on the sequence of actions the searcher had undertaken in the session. One of their main findings was that searchers who are more successful are generally more active (*e.g.*, more queries issued and results clicked) in a given time window. Aula *et al.* [3] analyzed different characteristics of successful and unsuccessful search sessions. Based on a small qualitative lab study and a subsequent large-scale evaluation, they established a range of indicators for user frustration during search sessions that were not yielding the desired results. Most saliently, the authors report longer sessions, question-type queries, the use of advanced query operators and aimless scrolling on the results page for failing searches. We will revisit these findings in Section 4 to employ them for identifying atypical sessions.

[1]http://www.dmoz.org

Beyond the effect of general search expertise on success rates, other recent work has considered the searcher's familiarity with the search topic. Based on a large-scale query log analysis, White *et al.* [33] found significant differences between the search behavior of domain experts and non-experts within the domain of their expertise (but not outside of the domain). The authors found that domain experts generated longer queries with more technical terms, had longer search sessions with more branches, and had greater success in satisfying their information needs than novices. Collins-Thompson *et al.* [6] investigated the use of reading level metadata for search personalization, finding that search ranking could be improved by taking into consideration the user's previous reading level preferences as well as the reading level coherence between a Web page and its result snippet. Kim *et al.* [18] followed up in this direction by jointly modeling reading level and topic preferences to describe users. Their so-called RLT profiles were used to distinguish domain experts from non-experts as well as to identify occurrences of 'stretch' reading behavior, *i.e.*, when users go beyond their usual preferences to satisfy information needs. Their work is central to ours as it observes users temporarily diverging from their profiles to solve particular tasks. In this work, we focus on the in-depth analysis and support of such cases. In a similar effort, Tan *et al.* [28] exploit notions of reading level and text comprehensibility for ranking popular answers on the Web portal *Yahoo! Answers*. According to the searcher's degree of domain expertise, simple *vs.* more technical answers were ranked higher. The research we present in this paper extends previous results by: characterizing the extent to which searchers diverge from their long-term search profiles, and demonstrating how the ability to detect such atypical sessions can be used to improve search personalization.

3. DATA SET

As a starting point for our investigation, we begin with an analysis of real search sessions to get insights into the problem domain. Our data set originates from the proprietary log files of the commercial Web search engine Bing. Our analysis focuses on a 4-month period of query logs from January to April, 2012 submitted by English-speaking U.S. users. We refer to the respective time spans as M1 (January) through M4 (April). Throughout this paper, we will use M3 as our profiling period and M4 to test for atypical sessions. Later, in Section 4, we will also investigate the usefulness of prolonged profiling periods, using M2 and M1 in addition to M3. To gain a first, qualitative insight into the domain, we limit our scope to the 200 most active users. Together, they submitted a total of 679,808 queries in 67,812 sessions. Session boundaries are drawn based on a 30-minute threshold of user inactivity as suggested by several previous studies (*e.g.*, [9, 10]). Since this work is concerned with information seeking behavior, we exclude navigational queries from our inspection in order to get a clearer impression of the difference between normal and atypical *informational* queries. To this end, we employed a list of frequent navigational queries as well as structural heuristics to detect queries encoding domain names or URLs (*e.g.*, those starting with "www." or ending in ".[domain]"). After this pre-processing step, 370,844 queries and 44,059 sessions remained for investigation. On average, users submitted 464 queries in 55 sessions per month. Figure 1 shows the observed distribution of session counts across users for M1.

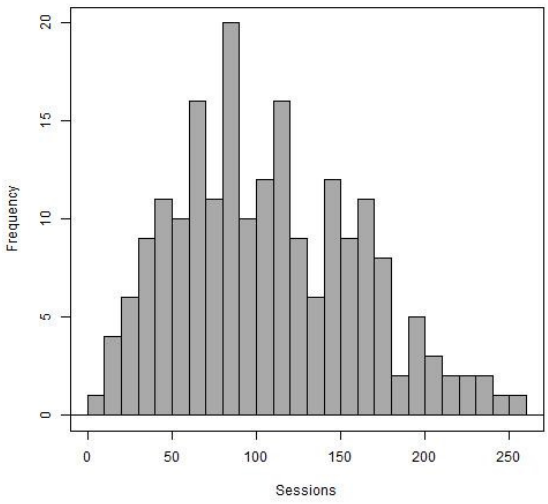

Figure 1: Distribution of session counts across unique users in January 2012.

3.1 Annotation

In order to get a deeper understanding of the extent and frequency of atypical search sessions, we collected manual labels for all search sessions from M4. Previous work [16, 19] suggests that sessions tend to be topically coherent, often serving a single information need. Based on this observation, we labeled typicality at the session level in order to preserve as much search context as possible. The labeling was facilitated by means of an internal crowdsourcing effort. Prior to this step, all sessions containing *personal identifiable information* (PII), such as names, phone numbers, addresses, social security numbers, *etc.*, as well as identical sessions for the same user, were removed in order to protect the users' privacy and anonymity.

At first, each annotator was shown a condensed profile representing the users' previous search history during the profiling duration (M3). For each clicked result, we automatically determined the topic distribution along the ODP's Web taxonomy. User profiles were summarized using the most common topics per user, *i.e.* those that comprised at least 5% of the overall click volume for that user. To give a more detailed overview of the user's interests, we also provided the three most frequently-issued queries per category. In addition to the topical domain, we incorporated information about the textual complexity of the accessed material. Following previous work [6], we estimated clicked pages' *reading level* (RL) and highlighted easy-to-read material ($RL \leq 4.0$) in green, moderately difficult material ($4.0 < RL < 9.0$) in blue, and advanced resources ($RL \geq 9.0$) in red. This allowed judges to make decisions about typicality using the distribution of topics, and the topic and reading level of individual resources. Figure 2 shows an example of a user profile as presented to the annotators. The judges were then presented with a single search session from M4. All queries, both in the profile as well as

55% Sports/Baseball ("ncaa baseball", "ectb baseball", "pg baseball")
14% Society/Religion_and_Spirituality ("pope benedict bio", "shamanistic travel", "sacred heart newton")
5% Reference/Education ("matlab student version", "umass email", "my math lab")
5% Sports/Hockey ("elmira pioneers", "umass lax", "necbl")

Figure 2: Example of Web search user profile with topical categories and color-encoded reading level.

in the session, were presented as Web search hyperlinks to enable the judges to quickly explore the types of content to which the search led. Analogously to the user profile, session queries were color-encoded by reading level. A short survey probed two main aspects of a session: 1) its typicality for the user, and 2) the degree of importance that the search task shows. The latter is interesting as we assume that atypical search sessions may often relate to important tasks or problems. The survey questions were:

```
1.  How typical is this session for the user whose
profile was shown above?

2.  Now go back to the list of queries in the
session above, and select all queries that support
your decision.

3.  For the queries you selected in #2, do you
think the desired information has high importance
to this user?  (e.g., likely to have lasting
value, or help solve an important problem.)
```

Questions 1) and 3) were rated on a 5-point scale. In addition to the main questions, the judges could give further feedback by means of a free text field. To account for subjectivity and inaccuracies of individual workers, each session was labeled by 5 independent judges. The final label was determined by averaging across the constituent judgments. One of the researchers broke ties that could occur when individual judgments were rejected as the worker had flagged their decision as 'unsure'. The task was offered as a privately sourced task on the crowdsourcing platform Clickworker[2] at a pay level of 5 cents per session, a rate comparable to those suggested in related studies [2]. A grand total of $500 was invested in label acquisition. We followed previous surveys on the design and quality control of crowdsourcing studies [17, 8] by employing a hand-labeled set of gold standard tasks as well as measuring agreement between judges in order to discourage low-quality submissions.

The results showed substantial agreement between workers for the typicality vote. The standard deviation between each individual crowdsourcing judge and the majority vote among all 5 judges was found to be less than one point (0.854). To give an indication of the general task difficulty, we asked 3 expert judges to create redundant annotations for a subset of 100 sessions in a lab-based study using the same interface as the crowdsourcing workers. Among experts, the standard deviation from majority votes was found to be even lower (0.495). Finally, we computed the overlap between majority votes from experts and those from crowdsourcing workers. In the vast majority of cases (82.6%) the two majority votes were identical. Details on this labeling task and construction of the user profile are available in a separate paper [7].

[2]http://www.clickworker.com

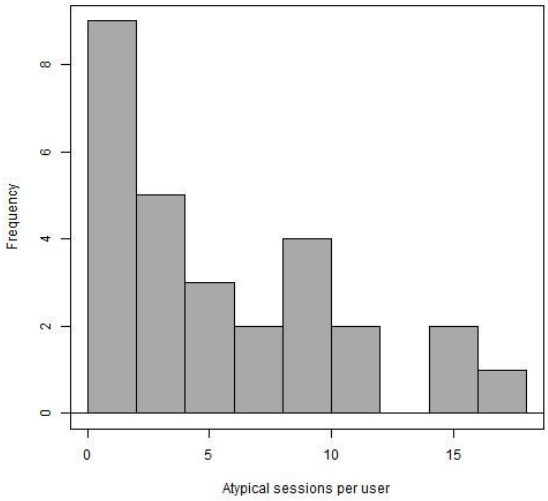

Figure 3: Distribution of atypical search session counts across unique users in April 2012 (M4).

3.2 Data analysis

Out of all 2790 informational search sessions labeled in M4, 166 were found to be atypical given the user's previous profile. Based on M3 profiles, 74% of all users showed at least one atypical search session in M4. On average, each user displayed 5.9 atypical sessions which comprised 7.5% of their overall monthly query volume. Figure 3 shows the distribution of atypical session counts across users. While atypical sessions can be observed for most users, their frequency differed across searchers. The average user was largely coherent in search behavior except for occasional atypical sessions, which is consistent with what Goel *et al.* [12] observed as well. Some, however, regularly explored different topics, making their search history very diverse topically.

As a next step, we compare typical and atypical sessions based on a number of session-level properties. Table 1 shows a juxtaposition across the whole user population. Statistically significant differences between session types are denoted by an asterisk. Significance was tested using a Wilcoxon signed-rank test ($\alpha < 0.05$). Both groups show comparable session lengths with only a slight increase in number of queries submitted in atypical sessions. As we turn to the queries, however, we observe significant differences. Atypical sessions show longer queries (5.23 vs 3.10 terms/query)

Table 1: Properties of typical and atypical sessions.

Type	Frequency	queries/session	terms/query	unique terms/session	SAT. dwell time	SAT. rank	RL	SAT. RL
typical	2624	6.26	3.10	8.93	209 sec	1.5	5.4	3.9
atypical	166	6.69	5.23*	16.07*	180 sec	1.8	5.8	5.3*

Table 2: Divergence from previous session statistics per user.

Type	Frequency	queries/session	terms/query	unique terms/session	SAT. dwell time	SAT. rank	RL	SAT. RL
typical	2624	-0.21	-0.11	-0.80	+15 sec	+0.19	-0.70	-0.40
atypical	166	+0.43	+1.70*	+1.55*	-21 sec	+0.49	-0.09	+1.80*

and also explore the result space more broadly by employing almost twice as many unique terms as regular sessions (16.07 vs 8.93 unique terms/session). Since explicit relevance judgments are not available, previous work frequently accepts clicked results on which the user dwells for at least a threshold amount of time, as relevant given the underlying information need. A commonly-used threshold is a dwell time of at least 30 seconds [9, 10]. We investigated these so-called *satisfied* (SAT) clicks in terms of their dwell time and the rank on which the user clicked. For typical sessions, such clicked pages show significantly longer dwell times (208 vs 180 sec.) and a better user satisfaction in the top ranks (average SAT click rank position of 1.5 vs 1.8) compared to atypical sessions. Finally, turning to textual complexity, we find comparable reading levels of clicked resources for both session types. However, SAT-clicked material in atypical sessions had significantly higher reading difficulty (average level 5.3) than in typical sessions (average level 3.9). In conclusion, we notice a less optimal search experience in atypical sessions along several dimensions - more complex queries, shorter SAT dwell times, lower SAT rank and higher SAT reading level.

In order to rule out the effect of individual users' querying style on these numbers, we now compare typical and atypical sessions per user. To enable this, we collected session statistics for all users in M3 and report the divergence from those observations for M4. Table 2 shows the population-wide averages of this comparison. Generally, typical sessions conform more closely with the user profile than atypical ones. While the differences are less marked than in the global comparison, we can note similar tendencies. For atypical sessions, queries are longer on average (by 1.7 terms) compared to the user's typical sessions, and use a wider selection of unique terms (an average of 1.55 more). Dwell times shrink and rankings are somewhat less optimal. Again, we note a higher textual complexity of documents that satisfy atypical information needs.

These distinct characteristics of typical and atypical search sessions are also reflected in the resulting retrieval performance. An indication of this trend can be found in the number of sessions that are abandoned without a single user click. This occurs 17% more frequently for atypical sessions than for typical ones.

Finally, we are interested in the content and cause of atypical search sessions. We manually grouped the 166 atypical sessions from M4 according to their high-level topic. Table 3 shows an overview of the most prominent resulting categories. Almost half of the atypical search sessions are concerned with health and medical information. Queries in these sessions are often dedicated to getting advice on healthy diets or finding information about causes and cures for certain medical conditions. Technical and computer queries are another major reason for atypical sessions. Typically computer problems such as viruses, or requests for help on diagnostic procedures, can be found in this group. The remaining 30% of atypical queries are distributed across a wide range of topics. Instructions for claiming taxes, preparing foreign recipes, or caring for pets were among the most prominent queries. To set these numbers into perspective, we contrast them with the overall frequencies of the respective categories in the same period. Interestingly enough, we note that many of the dominant topics in atypical sessions occur at significantly different frequencies than in the global collection. For example, medical and technical queries are significantly less common in the set of general queries. On the other hand, generally popular topics such as inquiries about celebrities, sports, or movies are rarely found in the set of atypical sessions.

Table 3: Prominent topics in atypical sessions.

Category	atypical freq.	typical freq.
Medical	49%	3%
Computers	21%	9%
Crafting	7%	3%
Cooking	5%	5%
Pets	4%	2%
Administrative	4%	2%
Travel	3%	7%
Other	7%	69%

With respect to our first research question, we conclude that atypical Web search sessions are events that affect the majority of users. Often, they occur when users seek advice on unfamiliar subjects outside of their topical area of expertise. As the users struggle with finding the appropriate keywords in the unknown domain, many affected queries are natural language questions, a class of queries that is known to often yield inferior result quality [4]. In Section 5, we will demonstrate that state-of-the-art personalization techniques achieve inferior results on atypical sessions.

4. IDENTIFYING ATYPICAL SESSIONS

The previous section summarized the concept and extent of atypical search sessions. In this section, we turn towards automatically identifying atypical search sessions and queries in order to appropriately react to the different nature of the information need. While, ultimately, it would be desirable to classify ongoing sessions to directly benefit re-

trieval performance, this first investigation of using typicality information for search personalization addresses sessions in a post-hoc fashion. To this end, we propose a two-step approach: First, we model user interests, preferences and querying style in the form of a profile, *e.g.* based on past sessions. Then, for new sessions, we measure the divergence from the existing profile.

4.1 Feature design

Based on the findings summarized in Section 3, as well as previous work, our classification scheme employs 2 distinct types of features: **(1)** Direct observations from the current search session. **(2)** Divergences of the current session from the user profile.

Session-level features

Session length Search sessions within a well-known topical area are typically shorter than those issued by users exploring a novel domain. In the latter case, frequent query reformulations can be expected as the user closes in on the desired information. To measure this effect, we consider the number of queries issued per session. In the previous section, this feature could not be confirmed to indicate session typicality when inspected in isolation. We include it in our classification scheme to test its validity in interplay with other features.

Query length Atypical queries, on average, were found to be longer than typical ones. We use the average number of terms per query as a feature.

Unique terms per session Previously we saw that there are significant differences in how deeply typical and atypical sessions explore a given topic by using a wide or narrow vocabulary. We measure the number of unique terms per session as an indicator of topic exploration *vs.* focused search.

Question query ratio In our qualitative analysis, we observed that many atypical sessions contain natural language questions. To account for this fact, we measure the ratio of queries per session that contain at least one of the following question words: *What, Where, When, Why, Who, How.*

Advanced operator ratio Previous work on the nature of unsuccessful and difficult search sessions [3], found that struggling searchers tend to make more use of otherwise often neglected advanced querying operators. We denote the ratio of queries per session employing at least one of the following advanced operators: *AND*, *OR*, *NOT* and *literal text matches* indicated by quotation marks.

Position of longest query The query editing history has been previously reported to hold information about the success rate of a search session [3]. Successful sessions tend to end in the longest query, as the user has sufficiently narrowed down the scope of the result set. On the other hand, unsuccessful sessions often see several iterations of specifications and generalizations before the search is finally abandoned. In the latter case the longest query can be found in the middle of the search session. We employ this observation by considering the relative position of the longest query as a feature (*i.e.*, the rank of the longest query divided by the overall number of queries in the session).

POS ratios Our analysis of atypical sessions showed a high number of natural language queries. In order to exploit this apparently different syntactic structure of regular and atypical queries, we apply *Part-of-Speech* (POS) tagging and note the relative frequencies of *nouns, verbs, adjectives*, and *miscellaneous constituents* (anything that could not be grouped into one of the previous categories). We assume that natural language queries will display a lower ratio of verbs and nouns but more of the 'syntactic glue', such as prepositions, that fall into the miscellaneous category.

Clicks per query Previous work [1, 33] found domain experts to be more active and to generally explore more results per query than non-experts. We measure the average number of clicks each query receives in order to account for different degrees of user activity and proficiency in the target domain.

SAT clicks per query Similarly to clicks per query, we consider the relative frequency with which SAT clicks (clicks with a dwell time of at least 30 seconds) occur. More frequent SAT clicks can indicate a better ability to formulate successful queries and identify relevant material in the result lists [15].

SAT click ratio In relation to the previous two features, we measure the relative number of satisfied clicks. A high ratio indicates efficient search behavior with targeted clicks on relevant material. Atypical search sessions are expected to display comparably lower ratios.

SAT click dwell time In Section 3, we saw shorter dwell times on the results of atypical sessions. In order to measure the degree to which the current result list satisfies the user, we report the average dwell time of all satisfied clicks in the session.

Median SAT clicked rank Previously, we observed a difference in ranking quality for regular and atypical sessions. For the latter, the user was more often forced to visit lower ranks of the result list. We account for this difference by measuring the median rank per session on which a SAT click was registered.

Reading level When faced with an unfamiliar problem, users are not always able to maintain their usual preferences for (typically lower) textual complexity. Due to the novel domain, they might lack the necessary knowledge for finding adequate, yet easy-to-understand material. Alternatively, the domain might inherently be of a more complex nature. We follow up on this notion by measuring the average reading level (as estimated using the classifier described in [6]) of all clicked search results per session.

SAT-clicked RL Similarly to the previous feature, here we only consider SAT-clicked pages. This distinction has been used in the previous section and was observed to separate regular and atypical search sessions better than considering all clicks.

Topical flags In the previous section, we saw that certain topics are more dominant in the group of atypical queries than others. To reflect this, we include a signal that indicates whether the current session serves, *e.g.* a medical information need. Since the actual distribution of topics underlying the user's information need is unknown, we employ topical classification of clicked results and note the relative frequency at which we observe the following categories: *Medical, Computers, Crafting, Cooking, Pets, Administrative, Travel.*

Unique topics per session Exploratory sessions in a novel topical domain tend to be more diverse than regular ones. Again, we classify all clicked search results into ODP categories and report the number of unique categories per session as a measure of coherence and focus.

Profile-based features

For each session-based feature above, we compute a corresponding divergence feature with respect to the user's profile. More specifically, we compute the difference between the session feature for the current session, and its historical average value across a user's previous sessions. For example, the length of the current session, minus the average session length across the profiling duration, will give the session length divergence feature. Additionally, two new feature types are considered.

Query term divergence For each user, we collect frequency counts of all query terms during the profiling duration. For each new session we do the same. Both profile and session can now be projected into a vector space with one dimension per unique term and frequency counts as components. We measure vocabulary coherence in terms of cosine distance between previous query terms and the current session.

Topic divergence Analogously, we also measure coherence in terms of topics, using cosine distance across topical vectors to account for sudden changes in the general subject domain.

For a high-level understanding of the problem domain, we computed estimates of the informativeness of all previously presented features. Table 4 shows a ranking of the 10 strongest features (out of 34 total) according to *Information Gain* (IG) and a χ^2 test. The feature rankings produced by the respective methods are largely consistent, with a few swaps at lower ranks. The strongest feature overall was the difference in query length from the user's previous profile (query length divergence), followed by the absolute query length. A majority (7 out of 10) of the high-ranking features are directly based on query information. Additional important signals are those based on the reading level of SAT-clicked pages (SAT RL) as well as the estimated difference in page topics (topic divergence). We note a balanced mixture of session- and profile-based features in the top 10.

Table 4: 10 strongest features for identifying atypical search sessions by information gain and χ^2.

Feature	Rank by IG	Rank by χ^2
query length divergence	1	1
query length	2	2
question ratio	3	4
verb ratio divergence	4	3
topic divergence	5	5
longest query position	6	8
SAT RL	7	6
SAT RL divergence	8	7
adjective ratio divergence	9	9
noun ratio	10	10

4.2 Classification

We applied the above features to the binary classification task of predicting whether a session was typical or atypical, relative to a given user's profile. We compared several different classification models, including support vector machines and various regression methods using the Weka toolkit [13]. The data set was split into distinct stratified training (90%) and test (10%) sets, such that no unique user's sessions were present in both sets. The session labels were obtained as described in Sec. 3.1.

Consistently, the best results on the training set were achieved by a logistic regression classifier that reached a final performance of $F_1 = 0.84$ ($P = 0.82$ and $R = 0.86$). When moving from the cross-validation setting on the training set to the previously unseen test set, we observed a final score of $F_1 = 0.74$ ($P = 0.8$ and $R = 0.68$). This number is comparable to a human annotator's accuracy of agreeing with the annotator majority vote label.

We investigated how the amount of previous search history used to compute features affected the classification performance in finding atypical sessions. Figure 4 shows cross-validation performance of the logistic regression classifier as a function of the number of search sessions per user that were used for building profiles. While performance rises quickly across the first sessions per user, scores level out between 18 and 20 sessions. At this point, no significant differences from the previously observed overall performance of $F_1 = 0.84$ can be observed. 95% of our users issued this threshold amount of 20 sessions across 14 days. Using longer search histories beyond this two-week threshold (*e.g.*, from M2 and M1) for profiling did not result in statistically significant performance changes.

With respect to our second research question, we conclude that automatic classification methods based on direct session-level features and divergence-from-profile features can be effective at estimating a search session's degree of typicality. Additionally, we note that a few weeks of query logs (typically between 1 and 2 weeks) were sufficient to make reliable typicality decisions for a given user.

5. RETRIEVAL PERFORMANCE

After examining properties of atypical search sessions (Section 3) and describing an automatic scheme for identifying them (Section 4), we now turn towards improving retrieval performance for atypical search sessions. Previously, we conducted our investigations on a sample of the 200 most active users. Now, we apply those insights to Web-scale retrieval tasks on a much larger dataset, as described next.

5.1 Method

Our study in this section closely follows recent research by Bennett *et al.* [5], and so we summarize that work briefly. In Bennett *et al.*, the authors examined a rich family of mod-

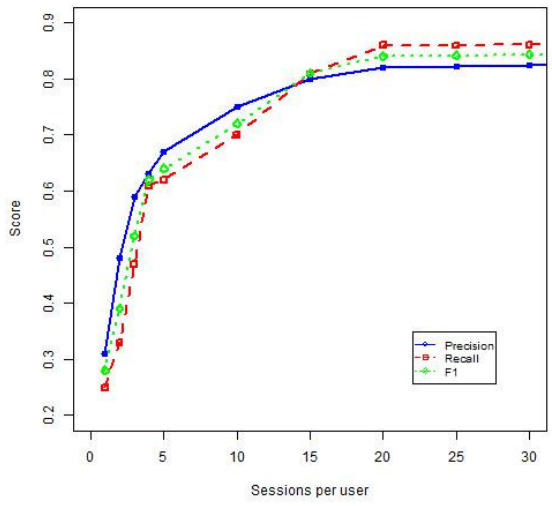

Figure 4: Classification performance as a function of the number of sessions per user in training set.

eling techniques for search personalization, looking at how different scopes of search and interaction history affected search personalization performance. They defined three key scope types: **(1) Session** considers all previous queries and actions from the current session. **(2) Historic** includes all previous actions and queries except for those in the current session. **(3) Aggregate** considers all previous actions before the current query. In addition to temporal extent, they considered several other factors used in earlier personalization research, including related queries and results in the profile. For each related query a score based on the weight of related queries, the similarity of the related results to the current result and the action taken on related results (*e.g.*, click, skip) was computed and used to re-rank the current results. Then, they used a unified re-ranking framework that comprised the following components:

rel(q, u) For each new query q issued by user u, this function returns a set of related past queries q_r from the same user.

view(u) Denotes the temporal view on the user u. Different temporal views can limit the scope of $rel(q, u)$. Possible settings for $view(u)$ are: *session*, *historic* and *aggregate*. Their concrete definitions will be given later in this section.

w(q_r, q, u) Each related query q_r is assigned a relatedness score. We abbreviate it as w_{q_r}.

results(q_r) For each related query q_r, a set of previously returned results d_{q_r} is stored.

sim(d_{q_r}, d) This function determines the similarity between a current result for query q and past results for related query q_r.

action(q_r, d_{q_r}) The action the user took for the previously seen result d_{q_r} for query q_r. Actions, such as SAT clicks, can indicate document relevance [9, 10].

Given the above components, many commonly studied personalization features can be represented as triples of $< q, d, u >$. The feature score $f(q, d, u)$ presented by [5] is:

$$f(q, d, u) = \sum_{q_r \in R} w_{q_r} \sum_{d_{q_r}} sim(d_{q_r}, d) action(q_r, d_{q_r}) \quad (1)$$

where $R = \{q : q \in rel(q, u) \,|\, q \in view(u)\}$.

A key finding in [5] was that the *aggregate* context scope achieved better overall improvements in MAP of satisfied clicks than either *session* or *historic*. The authors noted that some sessions showed performance losses, which might be attributable to sessions in which users look for very different material than what they are usually interested in. We hypothesize that the atypical sessions studied in this paper are examples of this class of sessions. Thus, as a first step towards personalization of atypical search needs, we investigate the performance of the above personalization framework on typical and atypical search sessions.

5.2 Experiments

To ensure comparability of results, we identically replicated the original experiment setting used in Bennett *et al.*, with the same underlying dataset [5]. Our experiments ranged over an 8-week period based on logs collected in July and August 2011. The selection covers 155,000 unique users and 10.4 million sessions, with an average of 174.4 queries per user and 2.61 queries per session. All reported results are mean values across 5 stratified experiment folds.

The features derived from the above framework were used to train a LambdaMART learning algorithm [35] for re-ranking the top 10 returned results. The goal was to produce an optimized ranking, and following [9], positive judgments were assigned to *satisfied result clicks* (SAT clicks). We estimated session typicality with the logistic regression classifier used in Section 4.

Table 5 compares the MAP re-ranking performance gains using *session*, *historic* and *aggregate* profiles over the original search engine ranking. We are also interested in the proportion of searches that were affected by the re-ranking. Consequently, we report the ratio of sessions whose MAP scores improve to those whose MAP score worsened. MAP scores are computed as the mean of average precision across the top 10 retrieved results. Cases in which the performance on atypical sessions differs significantly from that of typical ones are marked with an asterisk (determined via Wilcoxon signed-rank test at $\alpha < 0.05$-level). We confirmed the previous finding of Bennett *et al.* [5] that aggregate profiles lead to the highest overall performance gains for typical sessions. However, atypical sessions show a very different trend. Session-level information yields the strongest gains, followed by aggregate information. Interestingly, re-ranking using historic (pre-session) profiles is worse than the original ranking for atypical sessions. All performance differences between different information sources for the same class of sessions (*e.g.*, historic vs. aggregate information for typical sessions) are statistically significant.

We now address this difference between personalization performance for typical and atypical search sessions. Rather than uniformly applying one type of search history for personalization, we propose a hybrid approach that uses an initial classification step to predict whether the user is enacting a typical vs atypical session. Then, for all typical

Table 5: Personalization for atypical search sessions.

δ_{MAP}

	session	historic	aggregate
typical	0.0023	0.0047	0.0064
atypical	0.0067*	-0.001*	0.0059*

improved / # worsened

	session	historic	aggregate
typical	1.56	1.26	1.48
atypical	1.79*	0.91*	1.5

sessions, we apply historic personalization, and for atypical sessions, we rely exclusively on session-level information. Table 6 shows the overall performance gains of the proposed hybrid approach compared to both constituent methods in isolation. Significant improvements over **both** constituent methods are marked with an asterisk. Despite the relatively low frequency of atypical sessions, there are substantial gains in overall performance over the original search engine ranking. This tendency is also reflected in the case-based improvement and loss ratios. Atypical sessions see significantly more performance losses than gains when exclusively using historic profiles.

Finally, we conduct the analogous experiment for hybrid session-level and aggregate personalization. Table 7 shows the result of this alternative setup. We can note that the improvements over the original search engine ranking are consistently higher than in the previous case. The gain of the hybrid method over uniform application of aggregate personalization shrinks, yet remains significant. This makes intuitive sense as aggregate histories already inherently contain session-level information. The ratio of improvements and performance losses remains largely stable.

With respect to our third research question, we were able to obtain significant personalization improvements when identifying atypical sessions first and treating them differently from typical ones during re-ranking, by applying short-term session-level personalization rather than the historic or aggregated versions.

6. DISCUSSION AND FUTURE WORK

One important implication of our study is that a user's motivation to succeed at a search, and the corresponding utility they place on finding the information, might be estimated in part *by the effort or risk they are willing to take to get the information*: an application of the classic von Neumann-Morganstern definition of economic utility. By 'risk' we have in mind a compound quantity that captures both a) the uncertainty of relevance for the information sources the user is accessing, as measured by proxy quantities such as how 'unfamiliar' or 'new' a source is for that user, and b) the opportunity cost that the user perceives from accessing these unknown information sources with uncertain payoff compared to accessing a known source with more certain payoff. This is a different dimension of user effort than is captured by existing behavior-oriented measures like user frustration, since it accounts for content-based factors such as the unfamiliarity and difficulty of the material being retrieved, and the quality of alternatives that may be available. We believe these connections to economic utility theory as well as related work on information foraging [23] could be a rich area for further exploration.

Our initial study and modeling of atypical searches could be extended in a number of directions, of which we mention two here. First, over time, certain information needs or topics that were initially atypical may become recurring and thus part of a user's typical profile; the classic example is the search for information on medical conditions that follow initial symptoms, through diagnosis and treatment. More generally, however, it would be important to broaden the class of tasks or needs for which we can model such variability. Second, the findings from our post-hoc analysis of atypical sessions could be applied to online prediction of atypicality, in which we make dynamic predictions of typicality as a user progresses through a session.

7. CONCLUSION

While previous work on search personalization has focused on the problem of matching content to a user profile, we characterize and predict *atypical searches*. For such sessions, matching against the user's existing profile may *not* be accurate or desirable. Atypical searches are particularly interesting because in many cases they correspond to high-motivation needs in which the user exhibits a willingness to stretch their own boundaries for what is familiar or easy. Based on human labeling of 'typical' vs. 'atypical' sessions from several months of commercial search logs, we analyzed topic, reading level, and session-level properties of atypical sessions. We found significant differences between typical and atypical sessions: certain topics such as medical information and technical support were much more likely to arise in atypical sessions, along with query features such as increased term count, more unique terms, and more natural language-type terms. We showed how atypical sessions could be successfully identified using a classification approach that combined session-level and profile-based features. Finally, we showed that the ability to identify atypical sessions results in significant performance gains for search personalization based on short- and long-term user profiles.

8. REFERENCES

[1] M. Ageev, Q. Guo, D. Lagun, and E. Agichtein. Find it if you can: A game for modeling different types of web search success using interaction data. In *SIGIR 2011*. ACM.

[2] O. Alonso and R. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. *ECIR 2011*, pages 153–164.

[3] A. Aula, R.M. Khan, and Z. Guan. How does search behavior change as search becomes more difficult? In *CHI 2010*, pages 35–44. ACM.

[4] M. Bendersky and W.B. Croft. Analysis of long queries in a large scale search log. In *Proceedings of the 2009 workshop on Web Search Click Data*, pages 8–14. ACM.

[5] P.N. Bennett, R.W. White, W. Chu, S.T. Dumais, P. Bailey, F. Borisyuk, and X. Cui. Modeling the impact of short-and long-term behavior on search personalization. In *SIGIR 2012*. ACM.

[6] K. Collins-Thompson, P.N. Bennett, R.W. White, S. de la Chica, and D. Sontag. Personalizing web search results by reading level. In *CIKM 2011*, pages 403–412. ACM.

Table 6: Session and historic information for search personalization.

	% improved	% worsened	# improved / # worsened	δ_{MAP}
session	3.32%	2.1%	1.58	0.00247
historic	3.53%	2.83%	1.25	0.00454
session/historic hybrid	4.11%*	2.6%	1.58	0.0055*

Table 7: Session and aggregate information for search personalization.

	% improved	% worsened	# improved / # worsened	δ_{MAP}
session	3.32%	2.1%	1.58	0.00247
aggregate	4.9%	3.31%	1.48	0.00637
session/aggregate	4.83%	3.19%	1.52	0.00639*

[7] C. Eickhoff, K. Collins-Thompson, P.N. Bennett, and S. Dumais. Designing human-readable user profiles for search evaluation. *ECIR 2013*.

[8] C. Eickhoff and A.P. de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*, pages 1–17, 2012.

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

[10] J. Gao, W. Yuan, X. Li, K. Deng, and J.Y. Nie. Smoothing clickthrough data for web search ranking. In *SIGIR 2009*, pages 355–362. ACM.

[11] S. Gauch, J. Chaffee, and A. Pretschner. Ontology-based user profiles for search and browsing. *WI-IAT 2003*, pages 219–234.

[12] S. Goel, A. Broder, E. Gabrilovich, and B. Pang. Anatomy of the long tail: ordinary people with extraordinary tastes. In *WSDM 2010*, pages 201–210. ACM.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The Weka data mining software: an update. *SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[14] T.H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *TKDE 2003*, 15(4):784–796.

[15] C. Hölscher and G. Strube. Web search behavior of internet experts and newbies. *Computer networks*, 33(1):337–346, 2000.

[16] R. Jones and K.L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *CIKM 2008*, pages 699–708. ACM.

[17] G. Kazai. In search of quality in crowdsourcing for search engine evaluation. *ECIR 2011*, pages 165–176.

[18] J.Y. Kim, K. Collins-Thompson, P.N. Bennett, and S.T. Dumais. Characterizing web content, user interests, and search behavior by reading level and topic. In *WSDM 2012*, pages 213–222. ACM.

[19] A. Kotov, P.N. Bennett, R.W. White, S.T. Dumais, and J. Teevan. Modeling and analysis of cross-session search tasks. In *SIGIR 2011*, pages 5–14. ACM.

[20] L. Li, Z. Yang, B. Wang, and M. Kitsuregawa. Dynamic adaptation strategies for long-term and short-term user profile to personalize search. *Advances in Data and Web Management*, pages 228–240, 2007.

[21] Z. Ma, G. Pant, and O.R.L. Sheng. Interest-based personalized search. *TOIS*, 25(1):5, 2007.

[22] N. Matthijs and F. Radlinski. Personalizing web search using long-term browsing history. In *WSDM 2011*, pages 25–34. ACM.

[23] Peter Pirolli and Stuart Card. Information foraging in information access environments. In *CHI '95*, pages 51–58.

[24] J. Pitkow, H. Schütze, et al. Personalized search. *Communications of the ACM*, 9(45):50–55, 2002.

[25] M. Speretta and S. Gauch. Personalized search based on user search histories. In *WI-IAT 2005*, pages 622–628. IEEE.

[26] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *WWW 2004*, pages 675–684. ACM.

[27] B. Tan, X. Shen, and C.X. Zhai. Mining long-term search history to improve search accuracy. In *SIGKDD 2006*, pages 718–723. ACM.

[28] C. Tan, E. Gabrilovich, and B. Pang. To each his own: personalized content selection based on text comprehensibility. In *WSDM 2012*, pages 233–242. ACM.

[29] J. Teevan, S.T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *SIGIR 2005*, pages 449–456. ACM.

[30] J. Teevan, S.T. Dumais, and E. Horvitz. Potential for personalization. *TOCHI*, 17(1):4, 2010.

[31] A. Thatcher. Web search strategies: The influence of web experience and task type. *Information Processing & Management*, 44(3):1308–1329, 2008.

[32] R.W. White, P. Bailey, and L. Chen. Predicting user interests from contextual information. In *SIGIR 2009*, pages 363–370. ACM.

[33] R.W. White, S.T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In *WSDM 2009*, pages 132–141. ACM.

[34] R.W. White and D. Morris. Investigating the querying and browsing behavior of advanced search engine users. In *SIGIR 2007*, pages 255–262. ACM.

[35] Q. Wu, C.J.C. Burges, K.M. Svore, and J. Gao. Ranking, boosting, and model adaptation. *Technical Report, MSR-TR-2008-109*, 2008.