Personalized Models of Search Satisfaction

Ahmed Hassan
Microsoft Research
Redmond, WA 98052 USA
hassanam@microsoft.com

Ryen W. White
Microsoft Research
Redmond, WA 98052 USA
ryenw@microsoft.com

ABSTRACT

Search engines need to model user satisfaction to improve their services. Since it is not practical to request feedback on searchers' perceptions and search outcomes directly from users, search engines must estimate satisfaction from behavioral signals such as query refinement, result clicks, and dwell times. This analysis of behavior in the aggregate leads to the development of global metrics such as satisfied result clickthrough (typically operationalized as result-page clicks with dwell time exceeding a particular threshold) that are then applied to all searchers' behavior to estimate satisfaction levels. However, satisfaction is a personal belief and how users behave when they are satisfied can also differ. In this paper we verify that searcher behavior when satisfied and dissatisfied is indeed different among individual searchers along a number of dimensions. As a result, we introduce and evaluate learned models of satisfaction for individual searchers and searcher cohorts. Through experimentation via logs from a large commercial Web search engine, we show that our proposed models can predict search satisfaction more accurately than a global baseline that applies the same satisfaction model across all users. Our findings have implications for the study and application of user satisfaction in search systems.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search Process

Keywords

Personalized search satisfaction; User cohorts; Satisfaction models.

1. INTRODUCTION

Evaluation is a central component of information retrieval (IR). Search system designers wish to better understand the performance of the systems they develop so that they can work to improve them. Metrics such as mean average precision (MAP) and normalized discounted cumulative gain (NDCG) have been developed by the IR community to evaluate search engine performance [20]. However, these methods require relevance judgments, which can be limiting and costly to obtain. Metrics are also computed at the query level, meaning that they cannot fully capture the holistic performance of search engines in their primary function, satisfying users.

Search behavior mined from log data can be used to provide implicit feedback from which the search engine can learn searcher preferences [3] and identify which results are relevant for particular queries [21]. Behavioral data can also provide insight into search engine performance during carefully controlled experiments involving comparisons between search result rankings [27] or compare the effectiveness of different interface treatments [25]. Given logs containing the interactions with a search engine, models of searcher satisfaction can be developed at the level of search engine result pages (SERPs), or complete sessions or tasks. In recent years there has been a growing interest in developing such models from search behavior, and various model refinements have been proposed [1][15][16][17]. The parameters settings in these models can be learned from in-situ judgments provided by users following events such as SERP departure [9], engine switching [14], or session termination [13][16], as well as other forms such as games and third-party labeling [1][15]. These methods have been used to develop models of searcher satisfaction applicable globally across all users. However, this fails to acknowledge the personal nature of satisfaction and research is needed to: (i) understand individual differences in the behaviors associated with search satisfaction, and (ii) develop tailored models to predict satisfaction for individual searchers and cohorts comprising searchers similar along one or more dimensions (e.g., topical interest, search expertise).

In this paper, we address this shortcoming by introducing and evaluating models of search satisfaction that learn signals from behavioral data gathered from a particular searcher or cohorts of similar searchers. Cohorts have utility in addressing data sparseness when we have insufficient data from an individual. We demonstrate the potential value of tailored models of search satisfaction by showing large individual differences in behavioral signals (SERP abandonment, dwell times, query refinements) typically associated with search satisfaction judgments provided directly by searchers in-situ. Using large-scale log analysis and estimates of dissatisfaction events mined from those data (approximated via automatically-labeled engine switching events) we compare the performance of our methods with a global baseline which estimates satisfaction across all users. We experiment with three cohorts: (i) users with similar topical interests, (ii) users with similar search expertise, and (iii) users exhibiting preference for one engine, all of which we thought could be related to satisfaction. Our results show increases in the accuracy with which we can predict satisfaction levels when using the tailored models. Modeling satisfaction at finer levels of granularity lets search engines more accurately estimate satisfaction with results, allowing them to compute better performance estimates.

The remainder of this paper is structured as follows. Section 2 describes related work in search engine evaluation and the application of behavioral signals to model search satisfaction. Section 3 further motivates this research by illustrating the extent of individual differences in commonly-used measures of satisfaction: abandonment, dwell time, and query refinement. Section 4 describes the models developed and the features they use. Section 5 describes the experiments performed to evaluate model effectiveness. Section 6 discusses findings and implications, and we conclude in Section 7.

2. RELATED WORK

There are a number of areas of related work relevant to the research described in this paper. These include methods and metrics for the evaluation of search systems, and inferring satisfaction and result relevance from observed search behavior, including individual actions and connected sequences of search behavior.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

Search systems are traditionally evaluated using the classical methodology involving a collection of documents, a set of pre-defined queries, and relevance judgments provided by human judges for subsets of the collection with respect to the queries [6][30]. The performance of search systems in retrieving relevant content from the collection and ranking it appropriately is determined using retrieval metrics such as MAP and NDCG [20]. These metrics employ a user model of how searchers inspect the result sets presented to them and compute estimates of relevance and relevance gain at different rank positions. These metrics are query based and ignore the connection between multiple queries occurring in a search session. The relevance judgments that these methods use are also expensive to collect and potentially noisy given that the third-party judges have limited knowledge of users' underlying search intent. It is therefore preferable to explore other methods of measuring engine performance, especially those that have lower cost, are more scalable, and are sourced from searchers not third-party judges.

Early research on implicit feedback demonstrated its utility for estimating search result relevance from behavioral signals [24]. Initial work on implicit feedback focused on client-side monitoring of events such as document retention (e.g., saving and copying), as well as dwell time estimates associating the amount of time spent examining a document with that document's relevance [23]. Moving from laboratory settings to Web-scale experimentation, implicit feedback also has utility in providing training data for learning-to-rank algorithms [2][21] and inferring search preferences [3]. Radlinski et al. [27] showed that interleaving the results of two ranking functions and presenting the interleaved results to users can serve as a good predictor of relative search engine performance. In their analysis, they discovered that metrics including abandonment and reformulation did not predict relative performance as accurately as interleaving. Using aggregated features extracted from user behavior is certainly correlated with satisfaction. However, it has been shown that modeling *transitions* between user actions during search is a stronger predictor of satisfaction [15].

Measures such as frustration and satisfaction can be estimated from sequences of searcher behavior on individual pages and also on search sessions. Fox et al. [13] found that a strong correlation exists between search log features and user satisfaction labels gathered in-situ via a browser plugin. They modeled explicit satisfaction ratings using features including clickthrough rate, dwell time, and features associated with session termination. In the absence of clicks, Diriye et al. [9] studied the relationship between SERP abandonment and satisfaction. They captured in-situ judgments of abandonment rationales and showed that the reason behind observed abandonments could be accurately inferred from behavior on SERPs, using among other things, cursor modeling, and that performance improved if preceding and succeeding search behaviors were also considered. Huffman and Hochster [19] studied the correlation between user satisfaction and simple relevance metrics. They reported a relatively strong correlation between user satisfaction and linear models encompassing the query URL relevance of the first three results for the first query in the search task.

Other research has shown that it is possible to infer other properties of the user, the task, of their current state of mind from search behavior. White and Morris [34] modeled the behavioral differences between novice and expert Web searchers, identifying the latter group as being those who used advanced query operators. They demonstrated differences in the search behavior of the two groups. Aula et al. [4] studied how user behavior changes in difficult search tasks, allowing inferences about the nature of the task to be made directly from search behavior. They performed a user study to understand how users behave with difficult search tasks. They found out that when faced with a difficult search tasks, users tend to use more diverse queries, use advanced operators, and spend more time on the search results page. Feild et al. [12] constructed models of user frustration using patterns of search interaction, but also input from physiological sensors which could measure signals such as heart rate and galvanic skin response. These studies provide interesting observations about user behavior, but do not model or predict search satisfaction as we do here.

Most relevant to the work described here are recent developments in behavioral modeling focused on the use of search interaction data to construct models of searcher satisfaction [1][15][16][17]. Hassan et al. [15] showed that modeling search satisfaction using action sequences of user behavior yields better performance compared to models derived from the query-URL relevance of top-ranked results for the first query in a task. The primary reason is that there can be different motivations for the same query and that the first query in a search task provides only limited insight into task satisfaction. A follow-up user study where satisfaction ratings were collected in-situ from users was presented in [16]. Ageev and colleagues [1] augmented this approach with additional search features. They also used a game like strategy for collecting labeled data where they ask participants to find answers to questions using Web search. Piwowarski et al. [26] have used models of user behavior of interactions to identify search behavior patterns and use those to predict query relevance without document content. To reduce the reliance of labeled data, that can limit the generalizability of the models, Hassan [17] proposed a semi-supervised approach to modeling Web search satisfaction. The proposed model uses a combination of labeled and unlabeled data to construct models of searcher satisfaction that outperform previous methods. A drawback of all of the methods described in this paragraph is that even though they gather judgments for each query instance, and hence for each user, all judgments are pooled to create a global satisfaction model. There are advantages of doing this, including more training data for machine-learned models. However, behavioral indicators of satisfaction differ between users (as we show in the next section), there is an opportunity to develop more tailored models of search satisfaction for each searcher or cohorts of similar searchers. The latter (cohorts) lets us balance additional focus with the need for sufficient training data for learning algorithms.

This work described in this paper extends previous work in a number of ways. First, as motivation for our research, using labeled satisfaction instances gathered directly from users, we demonstrate the existence of large individual differences in a number of behavior signals traditionally associated with satisfaction. Second, we propose and evaluate tailored models of search satisfaction, mined from behavioral data, and focused specifically on particular searchers and search cohorts rather than all users as has been proposed in previous work. Third, we show that these models outperform global satisfaction models where parameter values are learned across all users. Finally, we perform additional experiments on the effect of combining the tailored and global models, and show that the combination leads to performance improvements.

3. INDIVIDUAL DIFFERENCES IN SATISFACTION AND DISSATISFACTION

At the outset of our studies, we wanted to understand the variance in behavior traditionally associated with search satisfaction. Although there are a number of possible behaviors, we focused on the following three since they are commonly used [9][13][17], and we had access to labeled data gathered direct from users: (i) *SERP abandonment*, where users do not click on any of the search results returned for a query, (ii) *query refinement*, the number of queries the user issued during a search task, and (iii) *dwell time*, average duration of non-SERP page visits during the task. Abandonment data was gathered during one study, and the refinement and dwell time data was gathered during a separate study. We now describe the analysis that we performed of each search satisfaction signal.

3.1 Abandonment

To study abandonment rationales, we needed a way to capture them in-situ, at abandonment time. To do this we obtained data from the authors of the study described in [9] which contained <userid, abandonment, label> tuples. Labels provided the motivation for the observed abandonment and were assigned by the searcher at abandonment time. To gather these judgments they deployed a plugin within their organization during December 2011. Over 900 users installed the plugin. While users had the plugin installed, when they did not click on any of the returned results and abandoned the SERP, the plugin displayed a popup survey allowing them to indicate whether they were satisfied or dissatisfied with the SERP. Abandonment was initiated with actions such as closing a tab, manually entering a URL in the address bar, reformulating a query, or being inactive on the SERP for a prolonged period (30 minutes or more), all without clicking on result links. Searchers could be satisfied if the search engine provided a special instant answer or if they saw the answer to their question directly in the snippet. They could be dissatisfied if they could not find any results worth clicking or did not obtain the answer from the SERP directly. In total, 7,274 judgments were gathered from 928 users. Since we were interested in individual differences, we selected the users who provided at least 10 judgments in our data. Of the 5,294 judgments that these users provided, 4,264 (80.5%) were either satisfaction or dissatisfaction, and the remaining judgments (19.5%) were for other reasons including unintentional abandonment. Focusing on satisfaction and dissatisfaction given their relevance here, of those judgments 54% were satisfaction-related, 46% were related to dissatisfaction.

To understand the distribution of rationales across users, we plotted the number of abandonment instances associated with satisfaction (SAT) against the number of instances associated with dissatisfaction (DSAT). Figure 1 shows the distribution across all users in our dataset. The figure includes the average number of SAT and DSAT abandonments over all users in the set, marked with a large red circle (SAT=12.9 instances, DSAT=9.7 instances). The figure shows that the distributions for both SAT and DSAT are highly variable between users, with some users always being satisfied when abandoning and some always being dissatisfied. These findings clearly demonstrate the risks involved in making generalizations about particular behaviors (e.g., that abandonment is always good or bad), when it is clear that there are large differences between users.

3.2 Query Refinement

In addition to studying satisfaction and dissatisfaction associated with individual SERP instances, we can also consider sequences of interactions across the duration of a search session or task. Query refinement can either be interpreted as measure of user dissatisfaction, suggesting they struggled to find information, or satisfaction, in that they were engaged and able to complete multiple aspects of their search task. We obtained labeled search tasks from the authors from the study described in [16]. Each task was labeled as either satisfied or not by the user performing the search. To gather this data, they deployed a plugin that detected when a user submits a query to any of the three major search engines (Google, Bing, and Yahoo!). Users were instructed to submit a satisfaction rating at the end of their search task, where a search task is defined as an atomic information need that may result in one or more queries [22].

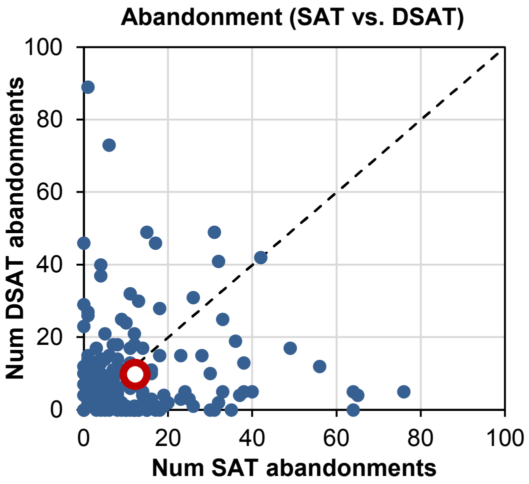

Figure 1. SAT and DSAT abandonment distributions across users for SERP abandonment dataset. Global SAT/DSAT value marked with red circle.

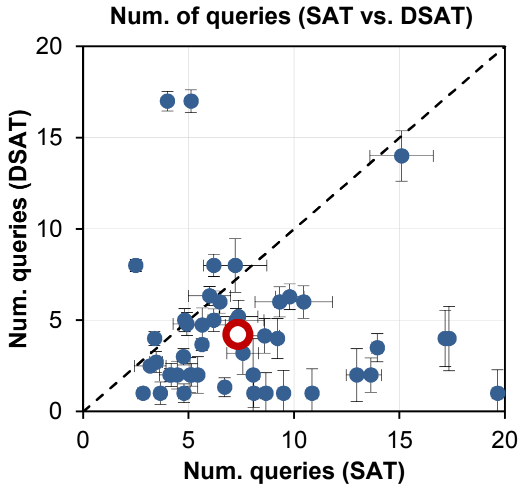

Figure 2. Average number of queries for SAT and DSAT tasks (±SEM). Global SAT/DSAT value marked with red circle.

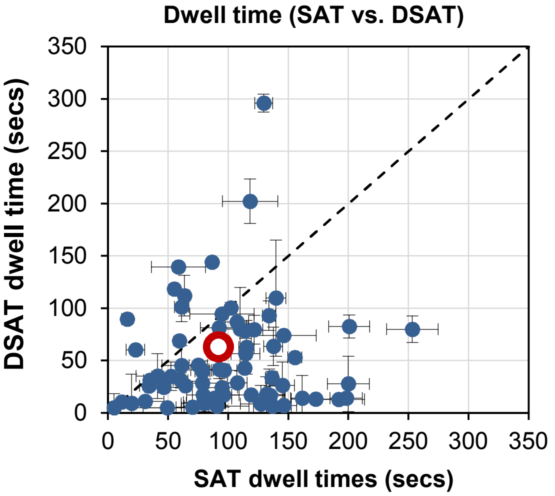

Figure 3. Average non-SERP dwells for SAT and DSAT tasks (±SEM). Global SAT/DSAT value marked with red circle.

The data gathered during that study provided in-situ judgments of satisfaction direct from searchers at the point of task termination. This meant that we were then able to compute the average number of queries performed by each user when they were satisfied or dissatisfied at the task level. Figure 2 presents a visual representation of the results for users from whom we saw at least 10 labeled search tasks overall and where we had evidence of both SAT and DSAT search tasks. Each dot on the scatterplot represents a single user. The overall average (7.34 queries (SAT) and 4.23 queries (DSAT)) is also marked on the figure across all users (large red circle).

Once again, the scatterplot clearly shows that there are large individual differences in the types of satisfaction associated with query reformulation. In a similar way to abandonment, there was significant deviation from the overall values. The small error bars (reflecting standard error of the mean (SEM)), also suggest that querying behavior is fairly consistent within each user. For some users, large amounts of query refinement within a task signaled overall satisfaction, whereas for others it signaled dissatisfaction. Important for demonstrating the extent of individual differences, some users exhibited behavior that was the complete reverse of what the overall values for SAT and DSAT suggest, i.e., issued fewer queries when satisfied and more queries when dissatisfied. Users above the diagonal in Figure 2 were more likely to behave in this manner.

3.3 Dwell Time

Using the same data set as the previous section we were able to also compute the average dwell time during a search task. Dwell time has been used previously to estimate satisfaction from search behavior [13]. For each user we can compute their average dwell time on non-SERP pages across the search task. The per-user values are shown in the scatterplot in Figure 3. Each dot in the plot represents a single user's SAT and DSAT dwell times. The overall time is 93s for satisfaction and 62s for dissatisfaction (large red circle).

In a similar way to abandonment and query refinement, the figure shows that there are large individual differences in the time duration they view non-SERP pages when they are ultimately satisfied or dissatisfied. Once again the within-user dwell time was consistent (low SEM). For most users, dwell time when satisfied exceeded dwell time when dissatisfied. This agrees with the findings of previous work [13], the overall values 93/62 reported above, and is widely accepted in the research community. However, there are users in the figure (those in the upper region above the diagonal) for whom DSAT dwell times far exceeded the average SAT dwell.

3.4 Summary

In this section we have shown that there are large individual differences associated with satisfaction and dissatisfaction for three commonly-studied behavioral signals—SERP abandonment, query refinements, and dwell time—associated with satisfaction. Our findings illustrate that to use these behaviors effectively as part of estimating satisfaction levels, we may need to also consider individual differences rather than making global assumptions about the causes for behaviors or assigning specific thresholds (e.g., a 30-second dwell time threshold for SAT [13]) across all users. These findings provide motivation for the analysis that we perform in this study. Note that within each user there are likely also effects on search behavior that can be attributed to the search *task* that they are attempting. In prior studies, search task has been shown to have a significant effect on users' search behavior [7][10]. Such effects are largely removed by averaging across all search behavior for a particular user. Although we do not consider satisfaction models for tasks or user-task pairs in this study, it is an interesting and important avenue for future work.

4. MODELS

In this section, we describe the methods that we use to predict dissatisfaction for searchers and searcher cohorts. These models are then applied to the task of predicting search dissatisfaction given recent search interactions. We start by describing the features we use for learning dissatisfaction models. These features can be either used to learn global models for all users, more personalized models for subset of users similar along some dimension (a user *cohort*), or models personalized for individual users. Next, we describe how we can build cohorts of searchers based on topics of interest, search expertise level, and engine preference. Finally, we describe several models that use those features to learn global models for all users, personalized models for individual users and user cohorts, or models combining the global and personalized data.

4.1 Features to Predict User Dissatisfaction

We use a large set of behavioral and SERP content features adopted and extended from previous work [14][17]. The set includes:

Query Features:

- Query length in terms of number of characters
- Query length in terms of number of words
- Query frequency estimated by counting the number of impressions of that query from a one-month worth of logs from a commercial search engine
- Query click-through rate (from the same month of log data)

Session Features:

- Number of queries
- Number of pairs of queries with word overlap (stop words not considered)
- Total number of clicks
- Number of algorithmic, sponsored and answer clicks
- Number of spelling suggestion clicks
- Number of related queries clicks
- Time in session so far
- Time to the first click
- Average, maximum, and minimum dwell time
- Average, maximum, and minimum time between queries
- Number of abandoned queries
- Average number of clicks per query

Search Engine Results Page (SERP) Features:

- Number of instant answers on the SERP
- Number of advertisements on the SERP
- Diversity of the results (number of unique Web domains and number of unique categories from the Open Directory Project (ODP, dmoz.org) that can be assigned to the results using an approach similar to [33]. URLs that exist in the directory were classified according to the corresponding categories. Missing URLs were incrementally pruned one path level at a time until a match was found or a miss declared.

Pattern of Behavior Features:

Previous work has shown that action sequences of user behavior are a strong predictor of user satisfaction [15]. Following previous work, we represent each search session as a sequence of actions. Actions include query submission, a related search click, a search result click, a sponsored result click, an answer clicks, and ending the search session. Given this set of actions A, we use the number of transitions between every action pair $a_i \rightarrow a_j$ for every $a_i \in A$, as well as the average time difference between every pair of actions. Additionally, we include action sequence features for the number of transitions between triplets of actions $a_i \rightarrow a_j \rightarrow a_k$. Previous work followed a first order Markov assumption by limiting features to pairs of transitions due to the limited number of data points in the training data [15]. The relatively large size of the dataset we use in this study (described in the next section), allows us to extend our feature set to longer sequences of user actions.

4.2 Cohorts of Searchers

In the previous section, we showed that there are large individual differences associated with satisfaction and dissatisfaction for several important behavioral signals often used to model dissatisfaction. This suggests that customizing dissatisfaction models to individuals may be of significant benefit. Unfortunately, when we try to customize models to individuals, we may run into a data sparseness problem (and we show later that this is indeed the case). Cohorts of similar searchers may be very useful for addressing this issue. A cohort is simply a group of searchers who share a common characteristic. While we may not have enough information to generate a personalized model for an individual user, we could use the profile for others who are similar to that user. In the remainder of this section, we present three ways of creating cohorts: (i) users with similar search expertise, (ii) users with similar topical interests, and (iii) users exhibiting preference for one engine. We selected these cohorts because they could be created from behavioral data and represented factors that could correlate with satisfaction.

4.2.1 Expertise Cohort

Previous work [34] has studied the behavioral differences between novice and expert searchers. Many differences between the behaviors of the two sets of users have been demonstrated. We hypothesize that these differences will also affect the signals correlated with search dissatisfaction. We identify expert users as those who use advanced query syntax. The correlation between expertise and the user of operators has been supported by the studies in [18][34].

Following [30], we used the following four operators, which are common to most search engines, as advanced syntax:

- **+ (plus):** which is used to match the search term exactly,
- **− (minus):** which can be added before a word to exclude all results that include that word,
- **“ ” (double quotes):** denotes exact match for a phrase, and
- **“site:”:** used to restrict the search to a domain or Web page.

Searchers who used any of this syntax in their queries were labeled as search experts and became part of the expertise cohort.

4.2.2 Topical Interests Cohort

Previous work has shown that people with topic knowledge are more efficient and effective in completing their search tasks [34]. We hypothesized that searchers with interest in different topics will have similar patterns of behavior. Hence, we could improve satisfaction prediction by using cohorts grouped by common topical interests. This allowed us to learn behavioral patterns that are specific to groups of users sharing common interests. To identify users with significant interests in different topics, we had to assign topic labels to different queries. We assign these labels by examining the top 10 URLs returned by the search engine given these queries.

With many millions of pages in our dataset, it was impractical to download and use the page content. Conversely, we could have used URLs or domains directly but that would be limited due to data sparseness (e.g., many URLs would only appear a few times in our dataset). To address this challenge, we used the ODP category labels for the URLs that users visit. ODP is an open Web directory maintained by a community of volunteer editors. It uses a hierarchical scheme for organizing URLs into categories and subcategories. Many previous studies have used ODP to assign topical categories to URLs (e.g., [29][33]). We focus on the top level categories (e.g., *Health*, *Computers*, *Shopping*, *Sports*, etc.) in our analysis since they were sufficiently distinct to distinguish users but also broad enough to contain enough users for cohort modeling.

Given the large number of URLs in our set we needed to label them automatically. We performed automatic classification of URLs into ODP categories using an approach described in [33]. We excluded the *Regional* and the *World* top level categories since they are typically uninformative in interest models. Queries were assigned the plurality label of the top 10 result URLs from the search engine.

We use one week worth of log data, not overlapping with the data used for training and testing our models, to assign topical interests to users. A user U is deemed significantly interested in topic P if the following conditions are satisfied:

- **Activity**: The number of queries submitted by U is at least 10% more than the average number of queries per user.
- **Topic Interest**: The percentage of queries ϵP submitted by U is at least 10% more than the average percentage of queries ϵP submitted by all users.

If these conditions are met for more than one topic for a user, they are assigned the topic with the highest percentage of their queries.

4.2.3 Engine Preference Cohort

We also create groups of users based on the search engine of preference (e.g., Bing, Google, or Yahoo!). To determine the engine of preference for each user we used one week of log data, not overlapping with the data used for training and testing the satisfaction models, as before. An engine E is designated as the engine of preference for a user U if U has the toolbar of E installed and uses E for the plurality of their searches. We determine whether a user has installed the toolbar for a particular search engine by matching certain patterns in the URLs of the search pages (e.g., one of the engines had a particular code in the URL if a query came from their toolbar). These patterns were obtained by issuing queries to different toolbars and observing the URLs submitted to the browser.

The hypothesis with this cohort is that different search engines may respond differently to queries and hence loyal users may adapt their behaviors to the engine. This can affect how they act when dissatisfied. Additionally, users of different search engines have been shown to have different demographics, as shown by ComScore reports[1], and demographics can also affect search behavior [31].

4.3 Methods

In this section, we present several methods that can be used to predict searcher satisfaction, using global data, personalized data, or a combination thereof. Notice that we can either build personalized models at the individual searcher level, or using cohorts of searchers. The latter method has the advantage that it overcomes concerns regarding data sparseness which are likely to be faced for individual models irrespective of how training data is captured. Before presenting the methods, we start by introducing the notation that will be used throughout this section.

Let X be the input space (typically $X = R^n$) and Y be the output space (DSAT or Not). For any given searcher or cohort of searchers, let us assume that we have two different distributions. One distribution over the searcher or cohort examples and another over the rest of the examples (i.e., global distribution). We assume we have

[1] http://blog.pmdigital.com/2010/08/who-uses-google-yahoo-and-bing

access to a sample of global examples D^G, and a sample of individual/cohort examples D^P. Let N be the size of the global dataset and M be the size of the personal dataset, where typically $N \gg M$. Our objective is to learn a classifier that would map X to Y while maximizing performance on personal examples.

GlobalOnly: In this method, we ignore all the personal data and train a single model using global data only (D^G) and the list of features from Section 4.1. Learning a single model across all users is what is traditionally done in satisfaction modeling [1][15][17] and this model serves as a baseline in our study.

PersonalOnly: This is the other extreme, where we train a single model for every searcher or searcher cohort (using D^P only) and completely ignore the global data. The same set of features is used. Remember that the word "personal" here may refer to an individual searcher or a cohort of searchers with a common characteristic as described in the previous section.

All: Here, we simply train a standard learning algorithm on the union of the global and personal data ($D^G \cup D^P$). The performance of this method is not expected to differ from the *GlobalOnly* method. The reason is that typically the global data is much larger in size than the personal data ($N \gg M$). Hence, any effect for the personal data will be probably offset by the global data.

Weighted: To alleviate the problem resulting from the difference in size between D^G and D^P , we reduce the weight of the examples from the global dataset and leave the weight of the personal data examples intact. For example if D^G is ten times the size of D^P then we can weight each example of the global data with 0.1.

Re-Classify: This is a cascade approach where the output of the global classifier is used as a feature for training a personal classifier. We start by building a classifier as described in the *GlobalOnly* method above. This classifier is applied to the training and testing data. The predictions made by the *GlobalOnly* classifier are used as an additional feature added to the personal data. Then we train a new classifier on the personal data along with the *GlobalOnly* feature. To classify new instances, we first obtain the *GlobalOnly* prediction and use that as input to the second classifier.

Prior (feature augmentation): The problem we are trying to solve is similar to domain adaptation problems that have been extensively studied especially in the context of natural language processing (NLP) applications. Domain adaptation tries to handle the mismatch that arises when models are trained using data from one domain (e.g., newswire) but are applied data from another domain (e.g., biomedical documents). Domain adaptation techniques have been successfully applied to several NLP problems including speech recognition, language modeling, and named entity recognition. This is similar to our problem where we are trying to handle the mismatch that arises due to the individual differences between searchers or cohorts of searchers as discussed in Section 3.

Several models have been proposed in the literature to handle domain adaptation scenarios with varying degrees of complexity. Chelba and Acero [5] introduced a model which uses the weights learned using the source data as prior on the weights for a second model trained on the target data only. The model in [5] was presented within the context of the maximum entropy and the maximum entropy Markov models. It has been show in [8] that it can be easily extended to other learning algorithms (e.g. Support Vector Machines, Naïve Bayes, etc.). This can be done by replacing the default regularization term $||w||_2^2$, with the regularization term $\lambda||w - w^s||_2^2$, where w is the weight vector being learned and w^s is the weight vector learned from the original classifier. This forces the learning algorithm to prefer the weights learned from the source classifier unless otherwise is demanded by the data. An alternative way for implementing this technique was presented in [8], where the two sets of weights are optimized jointly rather than sequentially by defining an augmented feature space. This is done by defining a mapping $\Phi^s(x) = \langle x, x \rangle$, and a mapping $\Phi^t(x) = \langle x, 0 \rangle$ on the source and target data respectively and then joining them into a single space. We use this feature augmentation method with global data (D^G) used as the source domain and personal data (D^P) used as the target domain.

Given each of the dissatisfaction models derived from search behavior and SERP content, the specific prediction task was to predict, given queries and the observed actions in a search session so far, whether the user was dissatisfied, when dissatisfaction was defined for the purposes of this study as a DSAT search engine switch as the next action in the session. Note that within the broader research area of satisfaction modeling, we focus specifically on predicting dissatisfaction, something that may be particularly important to search engines since it could lessen usage and revenue.

5. EVALUATION

We compare the performance of the models described in the previous section. In this section we describe the data we gathered and the results from the experiments that we performed.

5.1 Data

We analyzed a total of five weeks of interaction logs from October and November 2012, obtained from hundreds of thousands of consenting users through a widely-distributed browser toolbar. These log entries include a unique identifier for the user, a timestamp for each page view, a unique browser window identifier, and the URL of the Web page visited. Intranet and secure (https) URL visits were excluded at the source. Any personally identifiable information was removed from the logs prior to analysis. In order to remove variability caused by geographic and linguistic variation in search behavior, we only include entries generated in the English speaking United States locale. From these logs we extracted search sessions. Every session began with a query issued to Google, Bing, or Yahoo! and could contain further queries or Web page visits. A session ended if the user was idle for more than 30 minutes. Similar criteria have been used in previous work to demarcate search sessions, e.g., [10][32]. These sessions were later segmented into search tasks using the model presented in [22]. Jones and Klinkner [22] showed that many search sessions consist of multiple tasks, where a search task is a single information need that may result in one or more queries. We use the terms search session and search tasks to refer to search tasks throughout the paper.

A major challenge facing research on modeling search satisfaction is the lack of labeled data to train effective learning algorithms. Hence, most previous work has either been limited to small-scale studies [12], or large-scale analysis of unlabeled data [4]. Some methods have been proposed to overcome the limited availability of labeled data by collecting this data directly from users using toolbars [16], or using games [1]. The outcome of these studies is typically in the order of several hundred labeled instances. While this may be sufficient for studying the aggregated behavior of all users, it is certainly insufficient for any level of personalization either at the searcher level or even at the searcher cohort level.

Given our objective of predicting whether a user is dissatisfied or not, we identify dissatisfaction instances using engine switches. Search engine switching is the voluntary transition from one Web search engine to another. A search engine switching event is a pair of consecutive queries that are issued on different search engines within a session. Note that in identifying pre-switch queries, if the

Table 1. Cohort statistics.

Type	*Cohort Name*	*Number of Users*	*Number of Sessions*
Expertise	Experts	3,513	8,797
Search Engine Preference	A	8,255	19,924
	B	5,290	13,357
	C	2,591	6,442
Topical Interests	Arts	1,433	3,069
	Business	4,528	9,389
	Computers	1,564	3,351
	Shopping	1,674	4,042
	Society	1,323	2,778

user issued a navigational query for a target search engine (e.g., search for "yahoo" on Google), this query is regarded as part of the switch and the preceding query in the pre-switch engine is used as the "pre-switch" query. Search engine switching is an important event that has been shown to correlate with dissatisfaction in prior research [14][32]. Additionally, using search engine switching allows us to identify dissatisfied searchers without looking at any other behavioral signals (e.g., clickthrough, abandonment, quick backs) used in our prediction models.

Previous work has shown that users may switch from one search engine to another for various reasons [14][32]. One of the most common reasons is dissatisfaction with the results they received from the pre-switch engine [14], which accounts for around 60% of engine switches. Guo et al. [14] proposed a method for predicting the cause behind an observed switch. They gathered ground truth data through the deployment of a plugin to around 200 Web searchers. The plugin captured switching rationales in-situ when a switch occurred and also logged search behavior before and after the switch. The cause of the switch could be one of: (i) *dissatisfaction* (the searcher was unhappy with the pre-switch engine), (ii) *coverage* (the searcher wanted to check the information they had found on the other search engine), (iii) *preferences* (they usually used the target engine or the target engine was better for the current task type), (iv) *unintentional* (browser defaults or homepage settings), and (v) *other*. They found that they could accurately predict when the reason for an observed engine switch was dissatisfaction-related using only features from the *destination* search engines (i.e., post-switch behavior) with an F-score of 78.99. We obtained that classifier from the authors and applied it to distinguish dissatisfaction related switches from all other switches. It is critical that we use only features of the destination search engine for this task, because this ensures that search behavior before dissatisfaction (i.e., at the source engine) is completely hidden from the label generation.

We applied this method to the log data described in Section 5.1, and identified all DSAT-related switches. After removing all users with less than five instances per month, we collected approximately 60,500 dissatisfaction instances from over 25,000 users. We randomly sampled the same number of non-switching search sessions from the same set of users to act as our negative class, assuming that the switching instances represent our positive class.

Interested readers can reproduce this work by using any search log data that has dissatisfaction labels (either by judges, in-situ or using signals like engine switching). Most recently a large dataset has been released in the Yandex Switching Challenge as part of the WSCD workshop collocated with WSDM 2013 [28]. The main challenge with using that data for this task is that the queries and URLs needed to construct the searcher cohorts have been replaced by unique identifiers. Nevertheless, cohorts could be built by clustering users with similar behavior, or by similar queries, where query similarity can be estimated by measuring similarity between clicked results (e.g., if two queries frequently lead to clicks on same results, then they likely express similar information needs).

5.2 Experiments and Results

We perform several experiments to evaluate the ideas proposed earlier. We start by performing a personalization experiment at the individual searcher's level, then we repeat the experiment on the cohort levels with nine different cohorts. Finally, we study the relation between the cohort size and type and the performance again. Evaluation was performed using a temporal split in the data. We used the last week of November 2012 for testing and the first 23 days for training. We assign users to cohorts based on one week's data from October 2012. All models train a logistic regression classifier using the features described in Section 4.1. Statistical significance was evaluated using the McNemar's test [11]. We evaluate all techniques using the accuracy and the F-measure. The accuracy of the best performing technique is always bolded (as are all techniques whose performance is not statistically significantly different at the 95% confidence level). Techniques whose performance is significantly different at $p < 0.05$ from *GlobalOnly* are marked with an asterisk.

5.2.1 Individual Personalization

We apply the methods presented in Section 4.3 to create a personalized classifier for 100 different users selected randomly from all users who had 50 or more sessions in the dataset. Table 2 shows the average accuracy, and F-measure for the different methods. We notice from the table that *PersonalOnly* method performs worse than *GlobalOnly*. Our hypothesis is that the limited number of data points in the individual cases is the reason why the classifiers personalized on the individual level perform so poorly. Expectedly, the difference between the *All* method and the *GlobalOnly* method is not statically significant; as the small size of the personal data eliminates its effect when combined with the global data. Among the three methods that combine the global and the personal, the *ReClassify* method performs best with a small, yet statistically significant ($p < 0.05$), gain over the *GlobalOnly* and the *All* methods. This shows that we can achieve a reasonable performance gain by training an individual classifier for each user. The gain however is limited because of data sparseness.

5.2.2 Cohort-based Personalization

The previous experiment showed the potential of personalized models to predict user dissatisfaction. However, it also showed that the gain is rather limited due to the data insufficiencies. In this section, we describe the experiment we performed to evaluate the performance of the dissatisfaction predictors using cohorts of searchers instead of individual searchers for personalization.

We experimented with nine cohorts of three different types (Expertise: expert searchers, Search Engine Preference: Engines A, B, and C, Topical Interests: *Arts*, *Business*, *Computers*, *Shopping*, and *Society*). The engine cohorts represented the three main search engines, and the categories chosen for the topical interests' cohort are the largest categories in ODP. Statistics about the number of users and the number of sessions for each cohort are in Table 1.

The results for the expertise, search engine preference, and topical interests' cohorts are shown in Tables 3, 4, and 5 respectively. As before, we report accuracy, and F-measure for the different methods. Starting with the expertise cohort, we notice that the *PersonalOnly* methods achieves limited gain over the *GlobalOnly* method. When the global and the personal data are combined, the effect of the small personal data is offset by the much larger global data. However, when we combine the two datasets in ways that favor the personal data, we achieve much higher performance gains. The best performing methods are the *Re-Classify* and the *Prior* methods. The latter achieves 7% improvement in accuracy over *GlobalOnly*, showing that cohorts can yield strong gains.

Considering the engine preference cohorts, we notice similar trends where *GlobalOnly* and *All* achieve almost identical results, then *PersonalOnly* with a considerable gain, and then the three other methods with even larger gains. Among the last three methods that combine global and personal data, *Prior* and *Re-Classify* perform best, there does not seem to be a clear winner among them, and then *Weighted* in last position. Intelligently incorporating cohort signals (rather than just merging them) seems to lead to better gains, likely because the signal is not overshadowed by the other data. Also, methods where we treat global data as an input and then tailor to personal seem to perform better than other combinations because they may correctly balance the two sources.

Very similar trends also exist in Table 5, where the results of the topical interest cohorts are shown. We notice that the gain of using personal data over global data is limited here though. The reason behind this might be the smaller size of the topical interests cohorts compared to the expertise and search engine preference cohorts. We explore these differences in more detail in the next subsection. Another possible explanation might be that the difference between user behavior in the global domain and the topical interests cohorts may not be as noticeable as the difference in the case of the expertise and the engine preference cohorts (i.e., different interests may not translate to significantly different behaviors). Another observation is that the *Prior* and the *Re-Classify* methods still outperform all other methods. As is the case with the expertise and engine preference cohorts, there is not a clear winner when comparing *Prior* and *Re-Classify*. These methods may be substitutable and further analysis is needed of aspects such as computational cost and flexibility before deciding which methods engines should apply.

Note that the performance of the method on the users not belonging to any cohort will be the same as the *GlobalOnly* baseline where no personalization is employed. For example, the engine preference cohort cover approximately 64% of the users (see Table 1). Personalized models will be used for these users while the *GlobalOnly* baseline will be used for the rest.

5.2.3 Other Experiments

We performed further analysis on the cohort results to try to understand how the type and the size of the cohort may be related to the performance gain due to personalization. Figure 4 shows the accuracy gain of the best performing method using both global and personal data (*Re-Classify* or *Prior*) over the *GlobalOnly* baseline for cohorts of different sizes. We divided the cohorts into two different bins. The cohorts with size more than the average size across all cohorts are in the "large cohorts" bin, while the rest are in the "small cohorts bin". The figure shows that the gain achieved on larger cohorts is more than the gain achieved on smaller cohorts for both methods. This suggests that the proposed methods benefit from larger cohorts because they are more likely to learn user behavior in this cohort as the data increases. The increase of cohort size should not come at the expense of its focus though. To verify this hypothesis, we created a random cohort (users assigned to the cohort at random) and varied the cohort size between *1000* and *10000* users. The performance of all the personalization methods (*PersonalOnly*, *Weighted*, *Re-Classify* and *Prior*) were either identical (not statistically significant difference) or worse than the baselines (*GlobalOnly* and *All*).

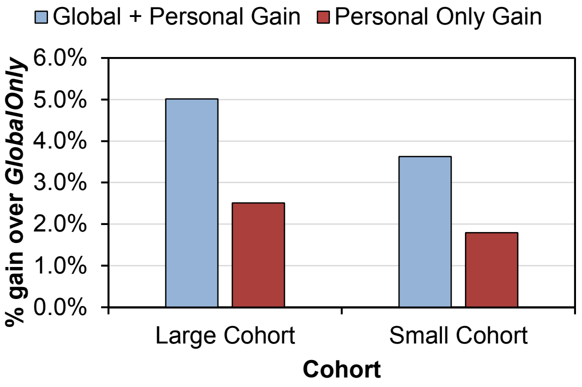

Figure 4. Percentage gain in accuracy from personalization over *GlobalOnly* for different cohort *sizes*.

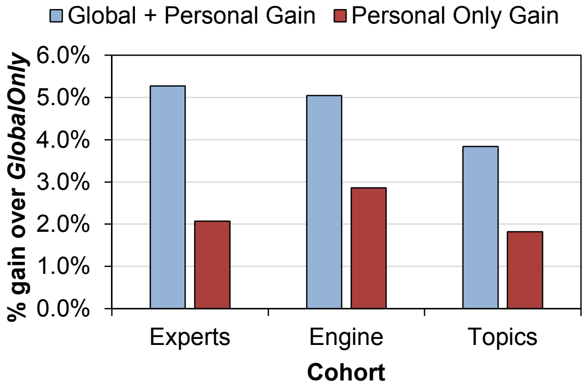

Figure 5. Percentage gain in accuracy from personalization over *GlobalOnly* for different cohort *types*.

Similarly, we also looked at the gain from the personalization methods across different types of cohorts. The results are shown in Figure 5. The figure suggests that the performance gain over the expertise and search engine preference cohorts is more than the gain over the topical interests' cohort. This may suggest that the difference between the global behavior and cohort behavior is clearer in the expertise and engine preference cohorts than in the topical interests' cohort. However, we have to consider that the average size of the topical interests' cohorts is smaller than that of the other cohorts (meaning we could not be sure it was the nature or the size that led to the difference). To better understand this relation, we down-sampled the users in the engine preference cohorts to have the same average size as the topical interest cohorts and we noticed that the gain drops to 4.1% which is still slightly larger than the average gain over topical interests cohorts (3.8%).

6. DISCUSSION AND IMPLICATIONS

We have shown in this study that tailored models of search dissatisfaction can be learned from behavioral signals and used to predict dissatisfaction at a level exceeding that of global methods. We experimented with dissatisfaction models tailored to individuals and to cohorts of searchers who were similar along different topical and behavioral dimensions. We showed although there was a small increase in performance from the individual models, cohort-based models predicted dissatisfaction most accurately.

Satisfaction is a personal emotion, and as we show in Section 3, it manifests in individual behaviors in different ways. Therefore, it was surprising to see in our study that models learned from cohorts of similar users outperformed the personalized satisfaction models tailored to individuals. There are a couple of possible reasons for this. First, personalized models do not have much training data on which to learn clear signals of dissatisfaction for a particular user. Although we did limit the users used for personalized model construction to those with a good amount of data, these models still struggled to perform as well as the cohort-based approaches. Second, given the way in which we defined dissatisfaction in our evaluation (as DSAT switches, a rare event), there were typically only a few DSAT switches each user, limiting the number of labels. Other cohorts could also be created. For example, we could automatically cluster users based on their search behaviors when satisfied directly or incorporate task information to make the models both tailored to the searcher *and* also to the current search scenario.

The strong performance of our methods opens up a range of possibilities to develop personal and group measures of search dissatisfaction. Since our models perform dissatisfaction prediction one action at a time, the model could be applied in real time to estimate search dissatisfaction and make decisions about whether to help the searcher directly (e.g., via new results or enhanced interface). Personalized models could also be applied retrospectively to help better understand the rationales behind logged behaviors.

7. CONCLUSIONS

Satisfaction is a personal emotion. Previous models of search satisfaction have ignored individual differences in behaviors associated with user satisfaction. In this paper we have demonstrated the large extent of individual differences around satisfaction and dissatisfaction. Given these differences we developed machine-learned satisfaction models tailored to individual searchers and searcher cohorts. Although personal models of dissatisfaction may be ideal, the lack of training data for individuals (even if we look over multiple weeks of data) makes the development of strictly personal dissatisfaction models challenging. Our findings show that tailoring models of dissatisfaction to similar users outperforms global models and represents a promising first step toward the development of more tailored satisfaction prediction. Future work involves exploring the development of richer cohort models (including the integration of multiple topical interests/cohort categories for individual users), creating personal models that address data sparseness via more data from more sources (e.g., browsing behavior, social media), and applying these models in both ad-hoc and post-hoc settings.

REFERENCES

[1] M. Ageev, Q. Guo, D. Lagun, and E. Agichtein. (2011). Find it if you can: a game for modeling different types of web search success using interaction data. *Proc. SIGIR*, 345–354.

[2] E. Agichtein, E. Brill, and S.T. Dumais. (2006). Improving web search ranking by incorporating user behavior information. *Proc. SIGIR*, 19–26.

[3] E. Agichtein, E. Brill, S.T. Dumais, and R. Ragno. (2006). Learning user interaction models for predicting web search result preferences. *Proc. SIGIR*, 3-10

[4] A. Aula, R.M. Khan, and Z. Guan. (2010). How does search behavior change as search becomes more difficult? *Proc. SIGCHI*, 35–44.

[5] C. Chelba and A. Acero. (2004). Adaptation of maximum entropy capitalizer: little data can help a lot. *Proc. EMNLP*.

[6] C.W. Cleverdon. (1960). ASLIB Cranfield research project on the comparative efficiency of indexing systems. *ASLIB Proceedings*, XII, 421–431.

[7] M. Cole, J. Gwizdka, C. Liu, R., Bierig, N.J. Belkin, and X. Zhang. (2011). Task and user effects on reading patterns in information search. *Inter. with Comp.*, 23: 346–362.

[8] H. Daumé III (2007). Frustratingly easy domain adaptation. *Proc. ACL*, 256–263.

[9] A. Diriye, R.W. White, G. Buscher, and S.T. Dumais. (2012). Leaving so soon? Understanding and predicting web search abandonment. *Proc. CIKM*, 1025–1034.

[10] D. Downey, S. Dumais, D. Liebling, and E. Horvitz (2008). Understanding the relationship between searchers' queries and information goals. *Proc. CIKM,* 449–458.

[11] B.S. Everitt. (1992). *The Analysis of Contingency Tables*. Chapman and Hall, London.

[12] H.A. Feild, J. Allan, and R. Jones. (2010). Predicting searcher frustration. *Proc. SIGIR*, 34–41.

[13] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. (2005). Evaluating implicit measures to improve web search. *ACM TOIS*, 23.

[14] Q. Guo, R.W. White, Y. Zhang, B. Anderson, and S. Dumais. (2011). Why searchers switch: understanding and predicting engine switching rationales. *Proc. SIGIR*, 335–344.

[15] A. Hassan, R. Jones, and K.L. Klinkner. (2010). Beyond DCG: user behavior as a predictor of a successful search. *Proc. WSDM*, 221–230.

[16] A. Hassan, Y. Song, and L.-W. He. (2011). A task level metric for measuring web search satisfaction and its application on improving relevance estimation. *Proc. CIKM*, 125–134.

[17] A. Hassan. (2012). A semi-supervised approach to modeling web search satisfaction. *Proc. SIGIR*, 275–284.

[18] C. Hölscher and G. Strube. (2000). Web search behavior of internet experts and newbies. *Proc. WWW*, 337–346

[19] S.B. Huffman and M. Hochster. (2007). How well does result relevance predict session satisfaction? *Proc. SIGIR*, 567–574

[20] K. Järvelin and J. Kekalainen. (2002). Cumulated gain-based evaluation of IR techniques. *ACM TOIS*, 20(4): 422–446.

[21] T. Joachims. (2002). Optimizing search engines using clickthrough data. *Proc. SIGKDD*, 132–142.

[22] R. Jones and K. Klinkner. (2008). Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. *Proc. CIKM*, 699–708.

[23] D. Kelly and N. Belkin. (2004). Display time as implicit feedback: understanding task effects. *Proc. SIGIR*, 377–384.

[24] D. Kelly and J. Teevan. (2003). Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2).

[25] R. Kohavi, R. Longbotham, D. Sommerfield, and R.M. Henne. (2009). Controlled experiments on the web: survey and practical guide. *Data Min. Know. Disc.,* 18(1): 140–181.

[26] B. Piwowarski, G. Dupret, and R. Jones. (2009). Mining user web search activity with layered bayesian networks or how to capture a click in its context. *Proc. WSDM*, 162–171.

[27] F. Radlinski, M. Kurup, and T. Joachims. (2008). How does clickthrough data reflect retrieval quality? *CIKM*, 43–52.

[28] P. Serdyukov, G. Dupret, and N. Craswell. (2013). WSCD2013: Workshop on web search click data 2013. *Proc. WSDM*, 787-788.

[29] X. Shen, S. Dumais, and E. Horvitz (2005). Analysis of topic dynamics in web search. *Proc. WWW*, 1102–1103.

[30] E. Voorhees and D. Harman. (2005). *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press.

[31] I. Weber and C. Castillo (2010). The demographics of Web search. *Proc. SIGIR*, 523–530.

[32] R.W. White and S. Dumais. (2009). Characterizing and predicting search engine switching behavior. *Proc. CIKM*, 87-96

[33] R.W. White and J. Huang. (2010). Assessing the scenic route: measuring the value of search trails in web logs. *Proc. SIGIR*, 587–594.

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

Table 2. Accuracy and F1 measure for the individual personalization experiment. The accuracy of the best performing technique is bolded (as are all techniques whose performance is not statistically significantly different at the 95% level). Techniques whose performance is statistically significantly different at the 95% level from the "GlobalOnly" baseline are marked with an *

Individual Personalization						
	GlobalOnly	**PersonalOnly***	**All**	**Weighted***	**Re-Classify***	**Prior***
Accuracy	71.43%	64.29%	71.43%	66.67%	**73.81%**	69.05%
F1	72.98%	64.19%	72.68%	72.67%	**76.55%**	71.10%

Table 3. Accuracy and F1 measure for the Expertise Cohorts

Experts Cohort						
	GlobalOnly	**PersonalOnly***	**All**	**Weighted***	**Re-Classify***	**Prior***
Accuracy	76.62%	78.69%	76.96%	80.47%	**81.42%**	**81.89%**
F1	79.72%	80.55%	79.86%	82.26%	**82.69%**	**83.04%**

Table 4. Accuracy and F1 measure for the Search Engines Cohorts

Engine "A" Cohort						
	GlobalOnly	**PersonalOnly***	**All**	**Weighted***	**Re-Classify***	**Prior***
Accuracy	75.48%	77.68%	76.78%	78.77%	78.89%	**80.25%**
F1	78.63%	78.43%	79.47%	80.36%	80.78%	**81.31%**
Engine "B" Cohort						
	GlobalOnly	**PersonalOnly***	**All**	**Weighted***	**Re-Classify***	**Prior***
Accuracy	75.35%	78.98%	76.93%	78.60%	**78.96%**	**80.47%**
F1	77.49%	79.11%	78.59%	79.45%	**80.83%**	**80.55%**
Engine "C" Cohort						
	GlobalOnly	**PersonalOnly***	**All**	**Weighted***	**Re-Classify***	**Prior***
Accuracy	75.51%	78.25%	75.51%	**79.16%**	**79.61%**	**80.75%**
F1	78.31%	79.08%	78.22%	**80.64%**	**80.86%**	**81.81%**

Table 5. Accuracy and F1 measure for the Topical Interests Cohorts

Topic "Arts" Cohort						
	GlobalOnly	**PersonalOnly**	**All**	**Weighted***	**Re-Classify***	**Prior***
Accuracy	78.61%	78.33%	79.17%	81.11%	**81.16%**	**82.50%**
F1	84.06%	84.02%	84.47%	85.41%	**85.41%**	**86.40%**
Topic "Business" Cohort						
	GlobalOnly	**PersonalOnly**	**All**	**Weighted***	**Re-Classify***	**Prior***
Accuracy	74.34%	76.26%	74.40%	77.52%	77.30%	**79.00%**
F1	78.05%	80.73%	79.60%	81.26%	81.25%	**82.29%**
Topic "Computers" Cohort						
	GlobalOnly	**PersonalOnly***	**All**	**Weighted***	**Re-Classify***	**Prior***
Accuracy	76.74%	79.00%	77.71%	80.29%	78.87%	**79.00%**
F1	81.20%	83.95%	83.13%	84.83%	83.77%	**84.03%**
Topic " Society" Cohort						
	GlobalOnly	**PersonalOnly***	**All**	**Weighted***	**Re-Classify***	**Prior***
Accuracy	75.00%	78.65%	77.01%	79.20%	**82.58%**	80.29%
F1	78.15%	82.14%	80.73%	82.30%	**83.18%**	83.28%
Topic "Shopping" Cohort						
	GlobalOnly	**PersonalOnly**	**All**	**Weighted***	**Re-Classify***	**Prior***
Accuracy	74.45%	75.99%	74.01%	77.97%	**83.06%**	79.53%
F1	78.99%	80.71%	79.01%	81.55%	**80.23%**	81.46%