To Personalize or Not to Personalize: Modeling Queries with Variation in User Intent

Jaime Teevan
Microsoft Research
Redmond, WA 98053 USA
teevan@microsoft.com

Susan T. Dumais
Microsoft Research
Redmond, WA 98053 USA
sdumais@microsoft.com

Daniel J. Liebling
Microsoft Research
Redmond, WA 98053 USA
danl@microsoft.com

ABSTRACT
In most previous work on personalized search algorithms, the results for all queries are personalized in the same manner. However, as we show in this paper, there is a lot of variation across queries in the benefits that can be achieved through personalization. For some queries, everyone who issues the query is looking for the same thing. For other queries, different people want very different results even though they express their need in the same way. We examine variability in user intent using both explicit relevance judgments and large-scale log analysis of user behavior patterns. While variation in user behavior is correlated with variation in explicit relevance judgments the same query, there are many other factors, such as result entropy, result quality, and task that can also affect the variation in behavior. We characterize queries using a variety of features of the query, the results returned for the query, and people's interaction history with the query. Using these features we build predictive models to identify queries that can benefit from personalization.

Categories and Subject Descriptors
H.3.3 [**Information storage and retrieval**]: Information Search and Retrieval – *query formulation*;

General Terms
Algorithms, Measurement, Performance, Reliability, Experimentation, Human Factors.

Keywords
Potential for personalization, clarity, personalized search.

1. INTRODUCTION
A number of factors are important to consider when ranking Web documents in response to a query. Of primary importance is the topical relevance of each document, or how well each document matches the query, and much research in information retrieval has focused on addressing this problem. However, search on the Web goes beyond ad hoc retrieval tasks based on topical relevance in several ways. People's Web queries are short, varied, and include navigational and resource queries [6, 22]. There are often many more documents that match a Web query than a searcher has time to view, and ranking becomes a problem not only of identifying topically relevant documents, but also of identifying those that are of particular interest to the searcher.

Fidel and Crandall [6] have shown that in addition to topic relevance, variables such as recency, genre, level of detail, and project relevance are important in determining relevance. Algorithms like PageRank [16] and HITS [13] take advantage of aggregate link information to get at some of these non-content features. In addition, Teevan et al. [24] have reported individual variation in what different people personally consider relevant to the same queries. These differences result in a large gap between how well search engines could perform if they personalized results for an individual, and how well they actually do perform by returning a single list designed to satisfy everyone.

Recent work on personalized search systems has focused on developing algorithms that personalize results using a representation of an individual's interests [3, 5, 20, 23]. In these systems, personalization is applied to all queries. However, as found by Dou et al. [5], personalization only improves the results for some queries, and can actually harm other queries. This can happen when unreliable personal information swamps the effects of aggregate group information that is based on considerably more information. Aggregate information can be collected in large quantities for queries an individual has never issued before, and this may be particularly useful when different people's intents for the same query are similar. On the other hand, when there is a lot of information available about what an individual it interested in related to a query, or when a query is very vague, it may make sense to focus primarily on the individual during ranking.

In this paper, we first examine the variability in user intent for a large number of queries using both implicit and explicit measures. We study how well variation in the implicit measures predicts variation in the explicit measures, and look at what other factors can account for variation in the implicit measures. Queries are characterized using a variety of features of the query, the results returned for the query, and the query's interaction history. Using these features we build predictive models to identify the queries that will benefit most from personalization, and explore which features are the most valuable for prediction.

2. RELATED WORK
Two lines of work are relevant to our research: predicting query difficulty and ambiguity, and personalized search. By predicting characteristics of queries or results sets, systems tuned to different types can be developed. For example, if a system knows which queries are hard, it can devote the appropriate resources to improving the results for those queries, or if a system knows which algorithm will work best for which queries, it could improve performance by selecting the right algorithm.

SIGIR'08, July 20–24, 2008, Singapore.

A number of researchers have explored methods for predicting query performance (e.g., [1, 4, 26]). Measures such as clarity [4], Jensen-Shannon divergence [1], and weighted information gain [26] have been developed to predict the performance on a query (as measured by average precision or mean reciprocal rank, for example) using characteristics of the query and/or result sets. Much of the early work attempted to predict query performance for traditional content-based or informational tasks. Zhou and Croft [26] extended this work by developing measures for both content-based and named-page finding tasks using a subset of the Web (the GOV2 TREC collection). We extend this line of work by focusing on the ability to identify the variation in judgments across individuals rather than query performance in the aggregate.

Leskovec et al. [15] used graphical properties of the hyperlinked result set to predict result quality and likelihood of query reformulation. The result quality measures they used (e.g., the ability to discriminate results in ranks 1-20 vs. 40-60, and the top-rated result) were different than the average precision measure used in the context of the TREC collections, so it is difficult to compare these results directly. Song et al. [21] investigated query ambiguity. An initial survey revealed three types of queries: ambiguous queries, broad queries, and clear queries. They then used features of the result set to classify queries as ambiguous or not (which included both broad and clear) using 250 hand-labeled queries. It is unclear to what extent their notion of ambiguity is related to query performance.

Teevan et al. [24] examined the variability in what different individuals found personally relevant to the same query. They evaluated their ideas using explicit relevance judgments from a small number of individuals and queries. Our work is similar to this, but greatly extends it by using explicit judgments for a larger number of queries as well as implicit measures of interest for a very large sample of Web queries. In addition, we develop models to predict variability in relevance across individuals.

Relevance is a complex concept and a review of that work is beyond the scope of this paper (see [18] for a review). One aspect that is of particular interest in our work is to characterize what different individuals find relevant for the same query. Fidel and Crandall [6] have described attributes other than topical relevance such as recency, genre, level of detail, and project relevance that are important in determining the quality of retrieval and filtering systems for individuals. Similar ideas motivated the development of techniques such as PageRank [16] and HITS [13] for ranking Web results, and these methods have been extended to compute different PageRank scores for different groups of users [9]. Our work follows in this tradition by focusing on the variability in what different individuals find relevant to the same query. We refer to this as ***query ambiguity.***

Several researchers have characterized differences in user behavior when interacting with Web search results for the same query (e.g., [5, 14, 25]). Dou et al.'s [5] work is quite relevant since it examines behavioral variability in the context of personalizing search results. They show that click entropy (i.e., the variability in results that people click for a query) is related to how well they can personalize results for a query. We extend this line of work by using both explicit and implicit indicators of relevance, a wide variety of query and result features, and most importantly by developing predictive models of ambiguity.

In addition, several groups have developed systems that personalize search results for individuals (e.g., [3, 20, 23]). These systems differ in many ways including how they model users interests (e.g., ODP categories, history of search queries and results visited, richer desktop history), and the details of the personalization algorithms (e.g., re-ranking using relevance feedback, query modification). Regardless of the details, however, they all apply the same personalization algorithm and parameter settings for every query. Yet, as noted above, personalization does not work equally well on all queries. Our work seeks to identify queries that show the most variability across individuals, and can thus benefit most from personalization methods. Being able to accurately identify these queries should provide useful input to all of these personalized search systems, as well as to new methods for supporting searchers in articulating their information needs.

The work reported in this paper examines the variation in user's search intent by measuring both explicit relevance judgments and large-scale log analysis of user interaction patterns. Although this is related to work on query performance, our main interest is in understanding differences in individual relevance with the goal of improving systems that personalize search. We characterize queries using features of the query, the results returned for the query, and people's interaction history with the query. This allows us to systematically explore the contributions of query history and results information. Using these features we build predictive models to identify queries that can benefit from personalized ranking. We do not attempt to classify queries into a small number of types (e.g., content-based or named-page) but rather try to directly predict the behavior of interest, which enables us to generalize to a wide range of user tasks.

3. METHODS

This section describes the methods we employed to understand query ambiguity. It begins with a description of the two data sets used to explore variation in query intent – one behavior based and the other comprised of explicit relevance judgments. It then presents several measures of query ambiguity and describes classes of features that can be used to predict query ambiguity.

3.1 Data Sets

To understand which queries have the potential to benefit from personalization we looked at a large sample of queries issued to the Live Search engine from October 4, 2007 to October 11, 2007 by more than ten unique individuals. We focused on queries that at least ten people issued to ensure we had sufficient data to understand the variation in behavior across people issuing the same query. Because we are also interested in predicting ambiguity for queries with fewer unique users (including queries that have never been issued before), in this paper we examine features that can be computed from only a single query as well as those that require previous history.

For each query, the results displayed to the users and the results that were clicked were extracted from the logs. In order to remove variability caused by geographic location and language, we only studied queries generated in the English speaking United States ISO locale. In total, we report on data from 2,400,645 query instances, covering 44,002 distinct queries. The queries were issued by 1,532,022 distinct users.

While this is a very large dataset, it can only be used to implicitly determine whether the queries might benefit from personalization (using, for example, variation in the results that were clicked). A more direct way to determine if different individuals consider different results relevant to the same query is to explicitly ask

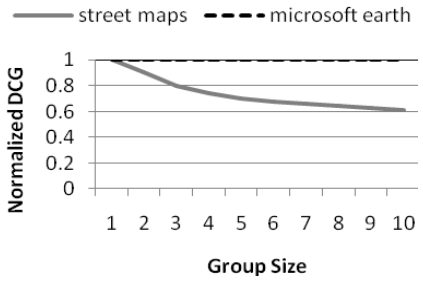

Figure 1. The potential for personalization curve for an unambiguous query like "microsoft earth" is flat, but dips with group size for a more ambiguous query like "street maps".

them. For this reason, we collected explicit relevance judgments from 128 people for 12 of the distinct queries in the logs. Between 4 and 81 individuals judged the top 50 results for each query (presented in random order) as *highly relevant*, *relevant*, or *not relevant*. In total 292 sets of judgments were collected.

3.2 Measures of Query Ambiguity

Using both of these datasets, we measured query ambiguity by calculating the variation in the explicit relevance judgments or the user behavior available from the log data.

3.2.1 Measures for Explicit Data

Several measures of query ambiguity have been explored in previous work. One common way to determine how much explicit relevance judgments differ between judges is to calculate the inter-rater reliability. As a measure of inter-rater reliability, we calculated Fleiss' kappa (κ) [7]. Kappa measures the extent to which the observed probability of agreement (P) exceeds the expected probability of agreement (P_e) if all raters were to make their ratings randomly:

$$\kappa = (P - P_e) / (1 - P_e).$$

Another measure of ambiguity is the *potential for personalization* curve [24]. Figure 1 shows example curves for two different queries ("street maps" and "microsoft earth"). Different group sizes are shown on the *x*-axis, and the *y*-axis represents how well a single result list can satisfy each group member in a group of that size (measured using normalized Discounted Cumulative Gain (nDCG) [10]). For a group of size one, the best list is one that returns the results that the individual considers relevant first. Such a list satisfies the single group member perfectly, and has an nDCG of 1. For larger group sizes, a single ranked list can no longer satisfy all individuals perfectly (unless they have identical ratings), so the average quality for group members drops. In the case of the query "street maps," nDCG for groups of two individuals decreases to 0.904. As the group size grows, so does the gap between the personalized performance for individuals (nDCG of 1) and the best possible performance for the group.

The shape of the potential for personalization curve depends on the query. When everyone has the same relevance judgments for a set of query results, then the same list makes everyone maximally happy, regardless of group size. The curve in such cases is flat at a normalized DCG of 1, as can be seen in Figure 1 for the query "microsoft earth." As different people's notions of relevance of the same results to the same query vary, the gap between what is ideal for the group and what is ideal for an individual grows, as it does for the query "street maps." Queries with big gaps are likely to benefit from personalization.

In this paper we quantify the curve by measuring the gap between the group curve and the ideal (nDCG of 1) at different group sizes (e.g., 5 or 10), and by clustering similar curves (discussed later).

3.2.2 Measures for Implicit Data

Both inter-rater reliability and the potential for personalization curves require explicit relevance judgments to calculate. Because explicit judgments are expensive to gather, we also explore two measures of ambiguity that rely on implicit data instead. These measures use clicks as a proxy for relevance, and capture the variation in the results searchers click on, on the assumption that queries for which there is great variation in clicked results also have great variation in what people consider relevant.

The first implicit measure of query ambiguity that we use in this paper is an implicit potential for personalization curve constructed using clicks as an approximation for relevance, with clicked results treated as results that were judged relevant. The example in Figure 1 was actually constructed from clicks, not explicit judgments. It shows that people clicked on the same results for "microsoft earth," but different results for "street maps."

The second implicit measure is click entropy, explored by Dou et al. [5], which measure the variability in clicked results across individuals. Click entropy is calculated as:

$$\text{Click entropy}(q) = -\sum_{\text{URL } u} p(c_u|q) * \log_2(p(c_u|q)),$$

where $p(c_u|q)$ is the probability that URL u was clicked following query q. A large click entropy means many pages were clicked for the query, while a small click entropy means only a few were.

3.3 Features Used to Predict Ambiguity

We considered a wide range of features that can be used to predict query ambiguity. These features represent different types of information that can be gathered in operational settings.

The simplest features require only the query string and are available for all queries. Other features require additional information about the query such as the result set, and may require an associated index to compute. Still others require history information about the query and/or results set, and thus are only useful for common (or "head") queries. Since many queries are unique [22], features that require query history will not be applicable to all queries for determining whether a query will personalize well. For this reason, we explore how well a wide range of features can be used to predict query ambiguity.

A summary of the features is shown in Table 1, broken down by the amount of information required to calculate the feature (query or result information), and the amount of query history necessary to calculate the feature (no history or some history). The features for each quadrant of the table are calculated for the 44k distinct queries in our data set using all 2.4 million query instances, since even for the same query there can be variation in the results returned, users' interactions with the results, and the time of day when the query is issued.

3.3.1 Query Features

Features that can be calculated using only a single issuance of the query without any additional information are shown in the upper left-hand cell of Table 1. These features include properties of the query string such as the query length, and whether a query uses

advanced syntax, mentions a geographic location, or contains a URL fragment. Although we have not explored such features, other query-based features could include the number of meanings the query has, as determined, for example, by WordNet.

In addition to features that characterize the query string, there are several features that relate to a single query instance, including temporal aspects of the query (e.g., was the query issued during work hours?). Finally, there are features that relate to characteristics of the corpus (but not the content of results), such as the number of results for that query, query suggestions, ads, or definitive results for the query.

3.3.2 Features that Require the Result Set

Other features can be calculated given knowledge of the set of results returned for a query. These features are shown in the lower left hand corner of Table 1. To calculate them we downloaded the title, summary, and URL for the top 20 results of each of the queries studied. Using this information we calculated features such as query clarity. Query clarity, proposed by Cronen-Townsend et al. [4], is a measure of the quality of the results returned for a query that does not require the query to have been seen by a search engine before. It measures the relative entropy between a query language model and the collection language model, and is calculated as follows:

$$\text{Clarity}(q) = -\sum_{\text{Terms } t} p(t|q) * \log_2\left(\frac{p(t|q)}{p(t)}\right),$$

where $p(t|q)$ is the probability of the term occurring given the result set returned for the query, and $p(t)$ is the probability of the term occurring in the index.

We also categorized each result according to what category it fell into in a large, human edited Web directory (the Open Directory Project, www.dmoz.org). Doing so allowed us to compute features related to the number of results that appeared in the Open Directory, the number of distinct categories covered by the result set, and the entropy of the categories covered. We further evaluated features of the results such as the number of distinct domains results were from, and the portion of results from different top level domains.

Table 1. Features used to predict query ambiguity, broken down by whether they require a history of interaction with the query (like click entropy) or the result set (like query clarity).

		History	
		No	Yes
Information	Query	Query length (chars, words) Contains location Contains URL fragment Contains advanced operator(s) Time of day issued Issued during work hours Number of results # of query suggestions offered # of ads (mainline and sidebar) Has a definitive result	Reformulation probability # of times query issued # of users who issued query Avg/σ time of day issued Avg/σ issued during work Avg/σ number of results Avg/σ # of query suggest. Avg/σ # of ads
	Results	Query clarity ODP category entropy # of ODP categories # of distinct ODP categories # of URLs matching ODP Portion of results non-html Portion that are ".com"/".edu" # of distinct domains	Result entropy Avg/σ click position Avg/σ seconds to click Avg/σ clicks per user Click entropy Potential for personalization

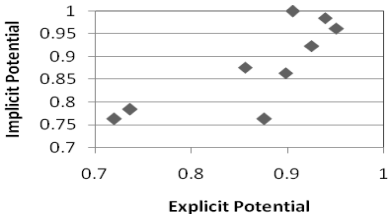

Figure 2. The potential for personalization for groups of size four, for curves computed for queries explicitly (using relevance judgments) and implicitly (using clicks). Implicit curves are strongly correlated with the explicit curves.

3.3.3 Features that Require History

The features shown in the right hand column of Table 1 can only be calculated if the query has been issued before. Features that rely solely on the query being seen before are shown in the upper right hand corner. These include the average and standard deviation of the features that can be calculated for a single query instance, the number of times the query has been issued, and the number of unique users who issue the query.

If there is further information about the history of the results that have previously been returned for the query and people's interactions with them, we can calculate more complex features. Given the history of the results displayed for a query, we can capture how often results change by calculating the result entropy:

$$\text{Result entropy}(q) = -\sum_{\text{URL } u} p(u|q) * \log_2(p(u|q)),$$

where $p(u|q)$ is the number of times the URL u was returned in the top ten results any time the query q was issued.

Other features in the lower right-hand quadrant of Table 1 are the average number of results clicked per user, the average rank of the clicked results, the average amount of time it took to click a result following a query, and the average number of results an individual clicks for the query. Our implicit target features of click entropy and the potential for personalization curve (discussed in Section 3.2) also fall into this quadrant.

4. UNDERSTANDING AMBIGUITY

This section discusses our explorations into understanding the relationships among different measures of query ambiguity. We look at how closely the implicit measures (click entropy and implicit potential for personalization) track the measures based on explicit relevance judgments (inter-rater reliability and explicit potential for personalization curves). We also correlate the query features in Table 1 with the implicit measures. We finish the section by highlighting several influences on the implicit measures beyond query ambiguity.

4.1 Comparing Explicit & Implicit Measures

The main focus of this paper is on understanding query ambiguity using implicit measures which can be obtained for a wide range of users and tasks. To confirm that the click-based implicit measures we used were a reasonable target, we examined their relationship to the explicit measures for the twelve queries for which we had both explicit and implicit data from at least four people.

The implicit measures of query ambiguity appear to correspond well with the explicit measures. As click entropy increases (meaning people click on a greater variety of results), the explicit measures of ambiguity decrease. The correlation between click

Table 2. Several key features and their correlations with implicit measures of query ambiguity, for all queries (All) and for queries with low result entropy (Low RE).

	Click entropy		Potential at 10	
	All	Low RE	All	Low RE
Query length (words)	0.20	0.16	0.17	0.11
Query length (chars)	-0.04	0.03	-0.06	0.00
URL fragment	-0.36	-0.23	-0.33	-0.18
Location mentioned	-0.03	-0.04	-0.02	-0.02
Advanced query	-0.01	-0.02	-0.01	-0.02
# of query suggestions	0.12	0.15	0.11	0.11
# of times issued	0.00	-0.01	-0.02	-0.04
# of distinct users	-0.01	0.00	-0.02	-0.04
Avg. # of results	0.03	-0.02	0.03	-0.01
% issued during work	-0.10	-0.04	-0.11	-0.05
Query clarity	0.02	-0.02	0.03	-0.01
Category entropy	-0.01	0.01	-0.04	-0.01
# of distinct categories	0.01	-0.02	0.03	0.01
# of URLs in ODP	0.09	0.05	0.12	0.07
Top level domain entropy	0.02	0.04	0.03	0.04
# of distinct hosts	0.19	0.17	0.16	0.13
Click entropy	1.00	1.00	0.87	0.86
Potential at 10	0.87	0.86	1.00	1.00
Result entropy (RE)	0.53	-0.04	0.40	-0.05
Avg. clicks per user	0.73	0.69	0.52	0.54
Avg. click position	0.90	0.86	0.86	0.83
Avg. seconds to click	0.03	0.05	0.04	0.06

entropy and the kappa inter-rater reliability is -0.36, and between click entropy and the potential for personalization curve at a group size of four is -0.46. The relationships trend in the right direction, but are not statistically reliable given the small sample size.

The implicit click-based potential for personalization curve is more strongly related to variation in explicit judgments than click entropy. In Figure 2, the value of the implicit potential for personalization curve at a group size of four is plotted against the explicit potential for personalization curve at the same group size. The values are highly correlated (correlation coefficient of 0.77, $p<0.01$). The implicit values are somewhat higher than the explicit values, which is to be expected given there is typically less variation in click behavior than in relevance judgments [24]. The correlation between the implicit potential for personalization curve at four and the kappa inter-rater reliability is 0.75 ($p<0.01$).

4.2 Correlating Features & Implicit Measures

The correlation coefficients between many of the features listed in Table 1 and the two implicit measures of query ambiguity (*Click entropy* and *Potential at 10*) are shown in Table 2. These correlations are based on 44k queries. They are broken down separately for queries with low results entropy (Low RE) and for all queries (All), as we describe in more detail in the next section.

Not surprisingly, the strongest correlations are for features that involve query history. The two implicit measures of query ambiguity, *Click entropy* and *Potential at 10*, are highly correlated. Figure 3 illistrates this relationship. It shows the

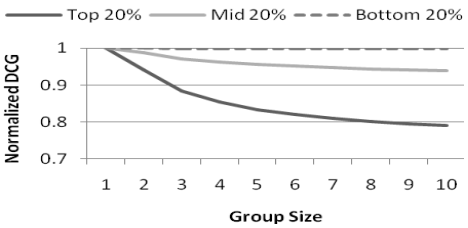

Figure 3. Potential for personalization curve as a function of click entropy. For low click entropy, there is almost no potential for personalization, while for high click entropy there is a lot.

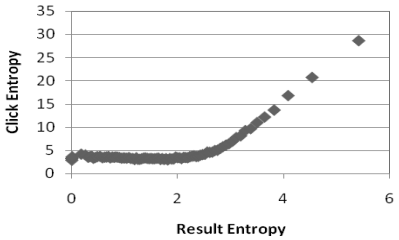

Figure 4. Result entropy and click entropy are correlated, but only for queries with result entropy greater than 2.

click-based potential for personalization curves for queries with high click entropy and low click entropy. There is a much higher potential for personalization for queries with high click entropy than there is for queries with low click entropy. Average click position is also strongly correlated with both measures. Note, however, that the time to click the first result is not correlated with *Click entropy* or *Potential at 10*.

Although smaller, there are also correlations between our implict measures of query ambiguity and some query features (e.g., query length in words and whether the query contains a URL fragment) and result set features (e.g., the number of distinct hosts).

4.3 Influences on the Implicit Measures

We now examine these correlations in more detail, focusing on several factors other than variation in intent that can influence our implicit measures of query ambiguity. We identify several reasons why a query could have high click entropy or a large gap in the potential for personalization curve, and yet not be a good candidate for personalization, including variation in the results presented, variation in the task, and variation in result quality.

4.3.1 Variation in the Results Presented

Queries may have high click entropy because there is a lot of variation in the results displayed for the query. Clearly if different results are displayed to one user compared to what is displayed to another, the users will click on different results even if they would consider the same results relevant. Selberg and Etzioni [19] studied the rate of change of search results, and found the results presented for the same query changed regularly. Further, some queries experience greater result churn than others, and thus will have higher click entropy despite not necessarily being good candidates for personalization.

Figure 4 shows click entropy as a function of result entropy. As can be seen, high result entropy is correlated with click entropy. Queries with result entropy greater than 2 have a 0.55 correlation with click entropy, while queries with result entropy less than 2

Table 3. **Example queries with an average number of clicks per user, broken down by high and low click entropy and high and low result entropy.**

		Click Entropy		
		Low	Mid	High
Clicks/User	Low	www.schoolloop.com usps.gov men's health magazine espn2	fox news network ontario airport wvu larry king	ecw fcc arrow internet explorer update
	Mid	corvette america cleartype petfinder.org pfchang	michigan state football alaska cruise trivia quiz knee injury	toyota camry rachel ray recipes bruce springsteen lyrics stress hormones
	High	(no queries)	restaurant guide famous poems calculate bmi woodrow wilson	first aid hand foot mouth disease cupcake recipes house spiders

have a -0.04 correlation. This trend holds for the potential for personalization for groups of different sizes. In many of our analyses we control for the effects of result entropy by focusing on queries with result entropy lower than two. Because there is a high amount of result churn in the real world, we also include some discussion of the effect of result entropy on our predictions.

4.3.2 Task Variation

Some of the observed variation in click entropy could result from the nature of the user's task (e.g., navigational or informational). While many queries, such as navigational queries like "CNN" or "Google", are followed by on average only one click, others are followed by a number of clicks. For example, a person searching for "cancer" may click on several results while learning about the topic. A query where half the people click on one result and the other half click on another result has the same click entropy as a query where everyone clicks on both results. But the variation between individuals in what is relevant to the query is clearly very different in the two cases, the first having a lot of variation, and the second having none. Thus it is not surprising that click entropy is correlated with the average number of clicks per user. Table 3 shows example queries with high and low click entropy and high and low average clicks per user.

The potential for personalization curves capture task variation somewhat in their shape. If we look at the curves for queries with the same click entropy but a different average number of clicks per user, queries where people click on fewer results have a greater gap at large group sizes than queries where people click on many results. This can be seen graphically in Figure 5.

4.3.3 Result Quality

There is also some indication that variation in clicks can be influenced by the quality of the results. Previous research [8, 11] has shown that people are more likely to click on the first result regardless of its relevance, so we would expect search results lists where the result being sought is not first to contain more variation. Although there is a bias towards clicking the first result, a lower average click position can still indicate lower result quality because the first result is not satisfying the searcher. We find that click position is highly correlated with the measures of ambiguity.

5. PREDICTING AMBIGUITY

We built predictive models to identify queries that will benefit most from personalization. In this section we talk about how the models are built and present our findings.

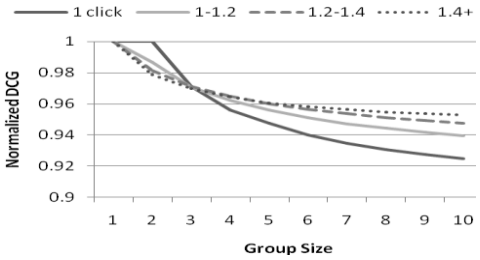

Figure 5. The potential for personalization curve for queries with a different number of average clicks per user. Result entropy and click entropy are held constant.

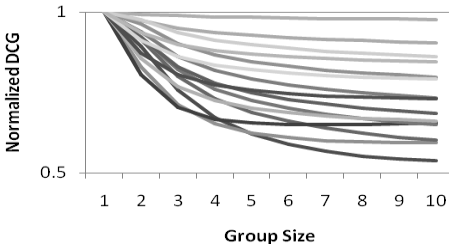

Figure 6. Potential for personalization cluster centroids. Some clusters show a sharp initial drop in what different people consider relevant to the query, while others drop more slowly.

5.1 Building a Model

To model query ambiguity, we learned Bayesian dependency networks that best explain the training data [2]. In the resulting models the conditional distributions at each node take the form of probabilistic decision trees. Parameters for restricting the density of the dependency network were estimated via cross validation on the training set. All results reported use five-fold cross validation.

Bayesian structure search to learn dependency networks is one of several feasible learning procedures. We used the method because it scales nicely to large training sets with large numbers of continuous and discrete features. Also, the dependency models and trees output by the method allowed us to inspect graphical relationships among observations and predictions. A comparison of other machine learning algorithms (e.g., logistic regression, support vector machines, etc.) is an item of future work.

Our targets for learning were a query's click entropy (binned into four equal-sized bins) and its implicit potential for personalization curve, characterized in several ways. One was to look at the gap at different group sizes (we used 5 and 10, again binned into four equal bins). We also tried to capture the nature of the curves by applying a clustering algorithm to group curves that have similar shape and magnitude. Specifically, we use repeated-bisection clustering [12] with a cosine similarity metric and the ratio of intra- to extra- cluster similarity as the objective function. In practice, we find that clusters are fairly stable regardless of the specific clustering or similarity metric. By varying the number of clusters and testing within- and between-cluster similarity, we found that the optimal ratio occurred at around 15 clusters. The clusters centroids are shown in Figure 6.

We predicted these four output variables (click entropy, potential at 5, potential at 10, and potential cluster) using various amounts of information, as described below.

Table 4. The model accuracy using different features and predicting different targets.

Features used			Result entropy	Click entropy		Potential at 5		Potential at 10		Potential cluster	
Query	Results	History		Baseline	Accuracy	Baseline	Accuracy	Baseline	Accuracy	Baseline	Accuracy
Yes	No	No	All	0.254	0.399	0.256	0.385	0.260	0.389	0.498	0.498
Yes	Yes	No	All	0.254	0.399	0.256	0.393	0.260	0.392	0.498	0.498
Yes	No	Yes	All	0.254	0.426	0.256	0.389	0.260	0.391	0.498	0.495
Yes	Yes	Yes	All	0.254	0.813	0.256	0.820	0.260	0.797	0.498	0.611
Yes	No	No	Low	0.258	0.360	0.258	0.360	0.257	0.355	0.342	0.342
Yes	Yes	No	Low	0.258	0.366	0.258	0.357	0.257	0.355	0.342	0.340
Yes	No	Yes	Low	0.258	0.360	0.258	0.390	0.257	0.376	0.342	0.341
Yes	Yes	Yes	Low	0.258	0.788	0.258	0.794	0.257	0.786	0.342	0.495

5.2 Model Accuracy

Table 4 shows how the model performed when trained under a variety of conditions. Each row represents one set of input variables reflecting different amounts of information: Just information that can be gleaned from a single query instance (the upper left-hand corner of Table 1, rows Yes-No-No in Table 4), just information that can be gleaned from the query and the result set (left-hand column of Table 1, rows Yes-Yes-No in Table 4), just information that can be gleaned from every issuance of the query (the top row of Table 1 – rows Yes-No-Yes in Table 4), and all of the information available to us except the prediction target (the entire Table 1 – rows Yes-Yes-Yes in Table 4). We look both at models that control for result entropy (Low) and models that do not (All). The baseline represents the performance when the most likely target class is selected.

In general, the patterns of results and the overall levels of accuracy are similar when predicting *Click entropy*, *Potential at 5* and *Potential at 10*. When predicting the *Potential cluster* the improvements are much smaller and the only notable advantages are obtained when using all variables. Predicting clusters is a challenging task with 15 target classes of varying frequencies that differ from each other in sometimes subtle ways.

Using the query features alone gives a sizeable improvement in prediction accuracy for all target outputs, except the *Potential cluster*. The improvement ranges from 50% to 57% over the baseline when we do not control for result entropy. When result entropy is held constant, the query features provide a smaller improvement for predicting, ranging from 38% to 39% for *Click entropy* and *Potential at 5* and *10*. The accuracies are consistently lower when controlling for result entropy, which is expected given we restrict the range of queries in this case.

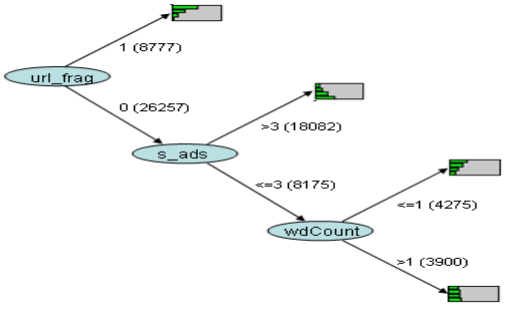

Figure 7. Portion of the learned tree for predicting Click Entropy using only query features.

Figure 7 shows a portion of the learned decision tree for the output variable *Click entropy* predicted using only query features. Nodes correspond to input features and each leaf node shows the probability distribution for the output variable, which is shown as a histogram. Labels on the edges show the splitting criteria of the parent variable, and the numbers in parenthesis show the number of training examples routed over that edge.

The three query features shown in this figure are whether the query string contains a URL fragment, whether the query is commercial (as measured by the number of ads that are shown), and the number of words in the query. The URL fragment and query length variables help distinguish between navigational queries that have low click entropy and informational queries that have higher click entropy. Queries that have a commercial intent have higher click entropy than those that do not. Although the overall level of prediction accuracy is moderate and can be improved using features of the interaction history (as we discuss below), it does suggest that we can identify queries with potential for personalization, to some extent, using only features gleaned from a single query instance.

There are some small improvements in prediction accuracy when query history features (without interaction history) or result set features (without query history) are added. Result set features add at most a 2% improvement. The query history features show somewhat larger gains of 6 to 8%. This suggests query history may be more valuable in predicting query ambiguity than the result set. It is interesting that adding result set features produces almost no advantages for our task since query clarity, weighted information gain, and Jensen-Shannon divergence depend heavily on such features. Our prediction task, however, is to identify queries with large differences across users and not to predict aggregate query difficulty, and different features appear to be relevant for doing so.

Using both query history and result set features together always produces a sizeable jump over other combinations, and high overall accuracies (approximately 80% in most cases). This is not surprising since, as we saw in Table 2, *Click entropy* and the *Potential at 10* were highly correlated. Practically it means that if we have previous evidence about how different users have interacted with the search results for query we should use it to predict whether to personalize or not. In our current models we require a minimum of ten previous users, and an interesting

direction for future research is to examine how few users are needed to still achieve high accuracy.

We are encouraged by these initial results and look forward to developing new features to improve accuracy. The extent to which these learned models can improve personalized search (e.g., by choosing different algorithms or parameter weights for different queries) is an important direction for future research. A complete answer to this question will involve examining performance in one or more personalized search systems, with and without using our query-by-query predictions to guide the choice of personalization algorithms or parameter settings.

6. CONCLUSION AND FUTURE WORK

In this paper we explored using the variation in search result click-through to identify queries that can benefit from personalization. Drawing on explicit relevance judgments and large-scale log analysis of user behavior patterns, we found that several click-based measures (click entropy and potential for personalization curves) reliably indicate when different people will find different results relevant to the same query. We also explored a number of additional factors that influence these implicit measures, such as result churn, task, and result quality.

Because click-based measures of query ambiguity are only useful for queries with a history of interaction, we investigated how well they could be predicted using many additional features of the query, including features of the query string, the result set, and history information about the query. We found that features of the query string alone were able to help us predict variation in clicks. Additional information about the result set or query history did not add much value except when taken in conjunction.

There are many ways the predictive models of query variation presented here could be used. We plan to explore their use within a personalized search framework. We believe we can provide a significant improvement to the search experience by personalizing results for queries that are ambiguous, while relying on the rich aggregate group data used in Web ranking for queries that are not. We would also like to explore using features related specifically to individual searchers to identify candidate queries for personalization. For example, individuals may benefit from personalized results when they want different results for that query than most people do. Finally, models for predicting query ambiguity may also be useful in identifying queries where additional assistance could be provided to searchers in articulating their information needs, or where search results could be diversified to satisfy a wider range of individual goals.

7. REFERENCES

[1] Carmel, D., Yom-Tov, E., Darlow, A., and Pelleg, D. (2006). What makes a query difficult? In Proceedings of SIGIR '06, 390-397.

[2] Chickering, D. M. (2002). The WinMine Toolkit. Technical Report MSR-TR-2002-103, Microsoft, Redmond, WA.

[3] Chirita, P. A.,Nejdl,W., Paiu, R., and Kohlschutter, R,C. (2005). Using ODP metadata to personalize search. In Proceedings of SIGIR 2005, 178–185.

[4] Cronen-Townsend, S., Zhou, Y., and Croft, W. B. (2002). Predicting query performance. In Proceedings of SIGIR '02, 299-306.

[5] Dou, Z., Song, R., and Wen, J.R. (2007). A large-scale evaluation and analysis of personalized search strategies. In Proceedings of WWW '07, 581-590.

[6] Fidel, R. and Crandall, M. (1997). Users' perception of the performance of a filtering system. In Proceedings of SIGIR 1997, 198-205.

[7] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378-382.

[8] Guan, Z. and Cutrell, E. (2007). An eye tracking study of the effects of target rank on Web search. In Proceedings of CHI 2007, 417-420.

[9] Haveliwala, T. (2002). Topic-sensitive PageRank. In Proceedings of WWW '02, 517-526.

[10] Järvelin, K. and Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In Proceedings of SIGIR '00, 41-48.

[11] Joachims, T., Granka, L., Pan, B., Hembrooke, H. and Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In Proceedings of SIGIR 2005, 154-161,

[12] Karypis, G. (2002). Cluto: A clustering toolkit. http://www.cs.umn.edu/~cluto

[13] Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5): 604–632.

[14] Lee, U., Liu, Z. and Cho, J. (2005). Automatic identification of user goals in Web search. Proceedings of WWW '05, 391-400.

[15] Leskovec, J. Dumais, S. and Horvitz, E. (2007). Web projections: Learning from contextual subgraphs of the Web. In Proceedings of WWW '07.

[16] Page, L., Brin, S., Motwani, R. and Winograd, T. (1998). The pagerank citation ranking: Bringing order to the Web. Technical report, Stanford Dig. Lib. Tech. Proj.

[17] Rose, D. E. & Levinson, D. (2004). Understanding user goals in Web search. In Proceedings of WWW '04, 13-19.

[18] Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *JASIST*, 58(3): 1915-1933.

[19] Selberg, E. and Etzinoi, O. (2000). On the instability of Web search engines. In Proceedings of RIAO 2000.

[20] Shen, X., Tan, B., and Zhai, C. X. Implicit user modeling for personalized search. In Proceedings of CIKM '05, 824–831.

[21] Song, R., Luo, Z, Wen, J.-R., Yu, Y., and Hon, H.-W. (2007). Identifying ambiguous queries in Web search. In Proceedings of WWW '07, 1169-1170.

[22] Spink, A. and Jansen, B. (2004). Web Search: Public Searching of the Web. Kluwer Academic Publishers.

[23] Teevan, J., Dumais, S.T., and Horvitz, E. (2005a). Personalizing search via automated analysis of interests and activities. In Proceedings of SIGIR '05, 449-456.

[24] Teevan, J., Dumais, S. T., and Horvitz, E. (2008). Potential for Personalization. Under submission.

[25] White, R. and Drucker, S. (2007). Investigating behavioral variability in Web search. In Proceedings of WWW '07, 21-30.

[26] Zhou, Y. and Croft, W. B. (2007). Query performance prediction in Web search environments. In Proceedings of SIGIR'07, 543-550.