Personalizing Search via Automated Analysis of Interests and Activities

Jaime Teevan
MIT, CSAIL
32 Vassar Street, G472
Cambridge, MA 02138 USA
teevan@csail.mit.edu

Susan T. Dumais
Microsoft Research
One Microsoft Way
Redmond, WA 98052 USA
sdumais@microsoft.com

Eric Horvitz
Microsoft Research
One Microsoft Way
Redmond, WA 98052 USA
horvitz@microsoft.com

ABSTRACT
We formulate and study search algorithms that consider a user's prior interactions with a wide variety of content to personalize that user's current Web search. Rather than relying on the unrealistic assumption that people will precisely specify their intent when searching, we pursue techniques that leverage implicit information about the user's interests. This information is used to re-rank Web search results within a relevance feedback framework. We explore rich models of user interests, built from both search-related information, such as previously issued queries and previously visited Web pages, and other information about the user such as documents and email the user has read and created. Our research suggests that rich representations of the user and the corpus are important for personalization, but that it is possible to approximate these representations and provide efficient client-side algorithms for personalizing search. We show that such personalization algorithms can significantly improve on current Web search.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Relevance feedback.

General Terms
Algorithms, Measurement, Design, Experimentation, Human Factors, Languages.

Keywords
Personalized search, Web search tools, adaptive interfaces.

1. INTRODUCTION
A Web search for "IR" returns a wide range of results, including stock quotes for the Ingersoll-Rand Company, Web pages in Arabic from Iran (with the .ir extension), details on infrared light, and a few about information retrieval. Readers of this paper would likely be uninterested in most of the search results returned for this query. This paper explores how search can be personalized, so that a search for "IR" returns results like the SIGIR homepage for the information-retrieval researcher, stock quotes for Ingersoll-Rand for the financial analyst, and pages about infrared light for the chemist.

In an earlier study [26], we investigated the variance in the informational goals of people using search engines, and the ability of current search tools to address such different goals. Our study showed that people differed significantly in the search results they considered to be relevant for the same query. This was true when users had different information needs (*e.g.*, information retrieval vs. Ingersoll-Rand) as well as when they expressed their underlying query intent in very much the same way (*e.g.*, "key papers in information retrieval" vs. "important papers in information retrieval"). The analysis suggested that, while current Web search tools do a good job of retrieving results to satisfy the range of intents people have for a given query, they do not do a very good job of discerning individuals' search goals.

We found that there is an opportunity to achieve significant improvement by custom-tailoring search results to individuals, and were thus motivated to pursue search algorithms that return personalized results instead of treating all users the same. Pitkow et al. [18] describe two general approaches to personalizing search results for individual users. In one case, the user's query is modified or augmented. For example, the query "IR", when issued by an information-retrieval researcher, might be expanded to "information retrieval". In the other case, the same search request is issued for all users, but the results are re-ranked using information about individuals. In our example, the SIGIR homepages might be pushed to the top of the ranking, while pages from Iran would fall to the end. In this paper we focus on result re-ranking, but also consider the broader space of designs for personalization

Because people are not good at specifying detailed informational goals, we use information about the searcher that we can glean in an automated manner to infer an implicit goal or intent. We explore the use of a very rich user profile, based both on search-related information such as previously issued queries and previously visited Web pages, and on other information such as documents and email the user has read and created. Our research suggests that by treating the implicitly constructed user profile as a form of relevance feedback, we can obtain better performance than explicit relevance feedback and can improve on Web search. Although most successful personalization algorithms rely both on a rich user profile and a rich corpus representation, it is possible to approximate the corpus and the text of the top-ranking documents based on the results returned by the Web search engine, making efficient client-side computation possible.

SIGIR '05, August 15–19, 2005, Salvador, Brazil.

2. RELATED WORK

There have been several prior attempts to personalize Web search. One approach to personalization is to have users describe their general interests. For example, *Google Personal* asks users to build a profile of themselves by selecting categories of interests [7]. This profile can then be used to personalize search results by mapping Web pages to the same categories. Many commercial information filtering systems use this approach, and it has been explored before to personalize Web search results by Gauch *et al.* [6][22]. Personal profiles have also been used in the context of the Web search to create a personalized version of PageRank [10] for setting the query-independent priors on Web pages. Liu *et al.* [14] used a similar technique for mapping user queries to categories based on the user's search history.

Information about the user's intent can also be collected at query time by means of techniques such as relevance feedback or query refinement. Koenemann and Belkin [12] examined several different interface techniques that varied in their transparency for allowing users to specify how their queries should be expanded. Anick [2] and McKeown *et al.* [15] explored alternative methods for generating query refinements. Relevance feedback and query refinement harness a very short-term model of a user's interest, and require that a query first be issued then modified. In practice, especially in Web search, explicit query refinement is rarely used [2]. More generally, people are typically unwilling to spend extra effort on specifying their intentions. A study by Teevan *et al.* [25] suggests that instead of fully specifying their search goals up front, people often browse to their targets via pages identified by less precise but more easily specified queries. This result resonates with the intuitions of Nielsen [17], who cautions against requiring users to perform extra work for personalization. Even when people are motivated to spend additional effort on specifying their search intent, they are not always successful in doing so [2].

A promising approach to personalizing search results is to develop algorithms that infer intentions implicitly rather requiring that the user's intentions be explicitly specified. Kelly and Teevan [11] review research on the use of implicit measures to improve search, highlighting several approaches in the literature that seek to tailor results for individuals. A wide range of implicit user activities have been proposed as sources of information for enhanced Web search, including the user's query history [20][22], browsing history [16][23], Web communities [13][23], and rich client-side interactions [3][4][16].

We focus in this paper on the use of implicit representations of a user's long-term and short-term interests. With this approach to personalization, there is no need for the user to specify or maintain a profile of interests. Unlike the systems described above, we explore very rich client models that include both search-related information such as previously issued queries and previously visited Web pages, and other information about the user such as documents and email the user has read or created. The paradigm allows us to evaluate the contribution of different sources of information over different periods of time to the quality of personalization in different contexts.

We take a relevance-feedback perspective [21] on modeling personalization. Relevance feedback has a solid theoretical foundation and a long history of application to information retrieval. Our approach differs from standard relevance feedback in that it does not require explicit judgments. The methods are distinguished from blind or pseudo-relevance feedback as they operate over a longer time frame than an individual query [19].

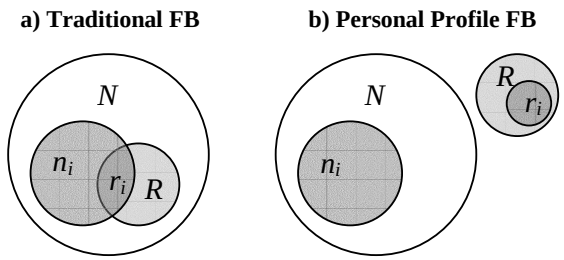

Figure 1. In traditional relevance feedback (a) relevance information (R, r_i) comes from the corpus. In our approach to user profiling (b), profiles are derived from a personal store, so we use $N' = (N+R)$ and $n_i' = (n_i + r_i)$ to represent the corpus instead.

To summarize, in our approach to Web search personalization, we use a wide range of implicit user activities over a long period of time to develop an implicit user profile. This profile is used to re-rank Web search results employing a relevance feedback framework. In our current approach, all profile storage and processing is done on the client machine. However, this can be generalized as we shall discuss.

3. PURSUIT OF PERSONALIZATION

For a Web search engine to incorporate information about a user, a user profile must either be communicated to the server where the Web corpus resides or information about the results must be downloaded to the client machine where a user profile is stored. We have focused on the latter case, on re-ranking the top search results locally, for several reasons. For one, such a methodology ensures privacy; users may be uncomfortable with having personal information broadcast across the Internet to a search engine, or other uncertain destinations. Second, in the re-ranking paradigm, it is feasible to include computationally-intensive procedures because we only work on a relatively small set of documents at any time. Third, re-ranking methods facilitate straightforward evaluation. To explore re-ranking, we need only collect ratings for the top-*k* returned documents, instead of undertaking the infeasible task of collecting evaluations for all documents on the Web. Within the re-ranking framework, we also examined lightweight user models that could be collected on the server side or sent to the server as query expansions.

We explored Web search personalization by modifying BM25 [21], a well known probabilistic weighting scheme. BM25 ranks documents based on their probability of relevance given a query. In use, the method essentially sums over query terms the log odds of the query terms occurring in relevant and non-relevant documents. The algorithm easily incorporates relevance feedback. Relevance feedback can be considered a very simple and short-term user profile, based on documents the user has selected as relevant to the particular query. We incorporate more complex profiles in the same manner that relevance feedback operates on the few documents identified by users as relevant.

We shall focus briefly on additional details of BM25. The method ranks documents by summing over terms of interest the product of the term weight (w_i) and the frequency with which that term appears in the document (tf_i). When no relevance information is available, the term weight for term *i* is:

$$w_i = \log \frac{N}{n_i}$$

where N is the number of documents in the corpus, and n_i is the number of documents in the corpus that contain the term *i*.

When relevance information is available, two additional parameters are used to calculate the weight for each term. R is the number of documents for which relevance feedback has been provided, and r_i is the number of these documents that contain the term. As shown graphically in Figure 1a, the term weight in traditional feedback is modified to:

$$w_i = \log \frac{(r_i + 0.5)(N - n_i - R + r_i + 0.5)}{(n_i - r_i + 0.5)(R - r_i + 0.5)}$$

In our approach, we do not have access to explicit relevance judgments on the documents returned for each query. Instead, we infer relevance based on a local store of information that we consider to be implicitly relevant. (We describe how the user profile is collected in more detail below.) The local user profile contains information that is not in the Web corpus, and this requires an extension to traditional relevance feedback models.

Figure 1b shows graphically how one can conceptualize using information outside of the Web corpus for relevance feedback as pulling the relevant documents outside of the document space. Thus we must extend the notion of corpus for the purpose of BM25 weighting to include the outside documents. We use $N' = (N+R)$ and $n_i' = (n_i + r_i)$ to represent the corpus instead. Substituting these values into the previous equation and simplifying, we get the following equation for the term weights:

$$w_i = \log \frac{(r_i + 0.5)(N - n_i + 0.5)}{(n_i + 0.5)(R - r_i + 0.5)}$$

To personalize search results, we compute the similarity of a query to a document, summing these term weights over all terms in the query (or expanded query).

There are a number of different ways to represent the corpus, the user profile, and different approaches to selecting the query terms that are summed over. We explore several different approaches, summarized in Table 1, in greater detail below. Three important components of our model are: Corpus Representation (how to obtain estimates for N and n_i); User Representation (how to obtain estimates for R and r_i), and Document/Query Representation (what terms are summed over).

3.1 Corpus Representation

Because the domain of our algorithms is Web search, the corpus is the Web. The parameter N represents the number of documents on the Web, and n_i, the number of documents on the Web that contain term *i*. A disadvantage of performing personalization on the client is that the client does not have direct access to details of the Web corpus. As a proxy for a Web index, we used the number of results reported by the Web search engine. To obtain estimates for n_i, we probed the Web by issuing one word queries. To obtain an estimate for N, we used the most frequent word in English, "the", as the query.

The query issued by the user can be used to focus the corpus representation. Corpus statistics can either be gathered from all of the documents on the Web, or from only the subset of documents that are relevant to the query (which we will refer to as a *query focus*). For example, if the query is "IR", a query-focused corpus consists only of documents that contain the term "IR". Thus, N, instead of being equal to the number of documents on the Web, is the number of documents that contain the term "IR". Similarly, n_i represents the number of documents that contain both term *i* AND "IR", instead of just the documents that contain term *i*. When the corpus representation is limited to a query focus, the user representation (which we describe below) is correspondingly query focused.

In practice, it is impractical to require the personalized search system to issue a query for each term it needs statistics for. Consequently, we also looked at approximating the corpus using statistics derived from the result set. We collected the corpus statistics either from the full text of every document in the result set, or from the title and snippet of each result. Using the full text of returned documents requires additional downloads, but using only the title and snippet does not require any additional information and is quite efficient. Collecting the corpus statistics in this way generates a query-skewed view of the results, but the approach serves to discriminate the user from the general population on the topic of the query.

3.2 User Representation

To represent a user we employed a rich index of personal content that captured a user's interests and computational activities. Such a representation could be obtained from a desktop index such as that described in Stuff I've Seen [5] or available in desktop indices such as Copernic, Google Desktop Search, Mac Tiger, Windows Desktop Search, Yahoo! Desktop Search or X1. The system we used indexed all of the information created, copied, or viewed by a user. Indexed content includes Web pages that the user viewed, email messages that were viewed or sent, calendar items, and documents stored on the client machine. All of this information can be used to create a rich but unstructured profile of the user. The most straightforward way to use this index is to treat every document in it as a source of evidence about the user's interests, independent of the query. Thus, R is the number of documents in the index, and r_i is the number of documents in the index that contain term *i*. As in the case of the corpus representation, the user profile can also be query focused, with R representing instead the number of documents in the user's index that match the user's query, and r_i, the subset that also contains term *i*.

We experimented with several techniques for using subsets of a user's index (each which could either be query focused or query independent) to compute R and r_i. For example, we explored the value of considering all document types (*e.g.*, email messages, office documents, and Web pages) versus restricting the document type to only Web pages. The motivation for exploring such a restriction is that the statistical properties of the terms might be significantly different in the user's full index than on the Web because of inherent differences in word frequencies associated with different types of information. As another class of restriction along a different dimension, we considered limiting documents to

the most recent ones. Because a user's interests may change over time, documents created or viewed more recently may give a better indication of a user's current interests than older documents. In the general case, we can consider that the time sensitivity of representations of a user's interests may differ by document type, and draw from a user's index, combinations of different types of documents, each restricted to different time horizons. In our studies, we looked at the value of only considering documents indexed in the last month versus the full index of documents.

Beyond analysis of the user's personal index, we considered two lighter-weight representations of the user's interests. For one, we used the query terms that the user had issued in the past. For the other, we boosted the search results with URLs from domains that the user had visited in past. Results associated with URLs where the last three components of the URL's domain name (*e.g.*, http://www.***sigir.confmaster.net***) matched a previously visited URL were boosted to the top, followed by those where the last two components matched (*e.g.*, http://www.sigir.***confmaster.net***). Both of these methods for representing a user's interests could easily be collected on servers hosting search services.

3.3 Document and Query Representation

The document representation is important in determining both what terms (i) are included and how often they occur (tf_i). Using the full text of documents in the results set is a natural starting place. However, accessing the full text of each document takes considerable time. Thus, we also experimented with using only the title and the snippet of the document returned by the Web search engine. We note that because the Web search engine we used derived its snippets based on the query terms, the snippet is inherently query focused.

In the absence of any information other than the user's query, a document's score is calculated by summing over the query terms, the product of the query term weight (w_i) and the query term occurrence in the document (tf_i). However, when relevance feedback is used, it is very common to use some form of query expansion. A straightforward approach to query expansion that we experimented with is the inclusion of all of the terms occurring in the relevant documents. This is a kind of blind or *pseudo-relevance* feedback in which the top-k documents are considered relevant [19].

Thus, for the query "cancer", if a document contained the following words,

> The American Cancer Society is dedicated to eliminating cancer as a major health problem by preventing cancer, saving lives, and diminishing suffering through...

each word would affect the document score. To maintain a degree of emphasis on the query, we also tried selecting from the documents the subset of terms that were relevant to the query. This was done in a simple manner, by including the words that occurred near the query term. For example, the following underlined terms would be selected from the previous snippet:

> The American Cancer Society is dedicated to eliminating cancer as a major health problem by preventing cancer, saving lives, and diminishing suffering through...

To summarize, we explore several different techniques for representing the corpus, the user, and the documents. These include *Corpus Representation*: Counts derived from the Web or from the returned set for estimating N and n_i; *User Representation*: Counts from the full index, temporal subsets or type subsets for estimating R and r_i; and *Document/Query Representation*: Words obtained from the full text or snippets of documents, and words at different distances from the query terms.

4. EVALUATION FRAMEWORK

To examine the usefulness of our personalized search system, we created an evaluation collection by having 15 participants evaluate the top 50 Web search results for approximately 10 self-selected queries each. Web search results were collected from MSN Search. For each search result, the participant was asked to determine whether they personally found the result *highly relevant*, *relevant*, or *not relevant* to the query. So as not to bias the participants, the results were presented in a random order.

The queries evaluated were selected in two different manners, at the participants' discretion. In one approach, users were asked to choose a query to mimic a search they had performed earlier that day, based on a diary of Web searches they had been asked to keep. We believe that these queries closely mirrored the searches that the participants conducted in the real world. In another approach, users were asked to select a query from a list formulated to be of general interest (*e.g.*, "cancer", "Bush", "Web search"). For the pre-selected queries, users were asked to describe their intent and to rate the relevance of documents relative to their intent. By allowing the participants to decide whether or not they wanted to evaluate a particular query, we sought to provide them with a query and associated results that would have some meaning for them.

Ignoring queries with no results, or where no results were marked relevant, we collected a total of 131 queries. Of those, 53 were pre-selected queries and 78 were self-generated queries. Each participant also provided us with an index of the information on their personal computer, so we could compute their personalized term weights. Their indices ranged approximately in size from 10,000 to 100,000 items. All participants were employees of Microsoft. Their job functions included software engineers, researchers, program managers, and administrators. All were computer literate and familiar with Web search.

To measure the ranking quality, we use the Discounted Cumulative Gain (DCG) [9]. DCG is a measure that gives more weight to highly ranked documents and allows us to incorporate different relevance levels (*highly relevant*, *relevant*, and *not relevant*) by giving them different gain values.

$$\mathrm{DCG}(i) = \begin{cases} \mathrm{G}(1), & \text{if } i = 1 \\ \mathrm{DCG}(i-1) + \mathrm{G}(i)/\log(i), & \text{otherwise.} \end{cases}$$

For the results we present here, we used $\mathrm{G}(i) = 1$ for relevant results, and $\mathrm{G}(i)=2$ for highly relevant results, reflecting their relative importance. Because queries associated with higher numbers of relevant documents will have a higher DCG, the DCG was normalized to a value between 0 (the worst possible DCG given the ratings) and 1 (the best possible DCG given the ratings) to facilitate averaging over queries. We also explored different gain functions (from $\mathrm{G}(i)=2$ to $\mathrm{G}(i)=100$, for highly relevant results), and an alternative overall performance measure (percent of *relevant* or *highly relevant* documents in the top ten results). In almost all cases, results are the same as those for the normalized DCG measure with a gain of 2 which we report here.

Table 1. Summary of differences between personalization variables. Significant differences ($p<0.01$) are marked with <, weakly significant differences ($p<0.05$) with '≤', and non-significant differences are marked as equal.

Corpus Representation (N, n_i)
Full text of documents in result set < Web < Snippet text in result set
Query focused = Based on all documents
User Representation (R, r_i)
No user model = Query history ≤ Indexed Web documents < Recently indexed < Full index
Query focused = Based on all documents
Document Representation (terms i summed over)
Snippet text < Full document text
Words near query terms < All words in document

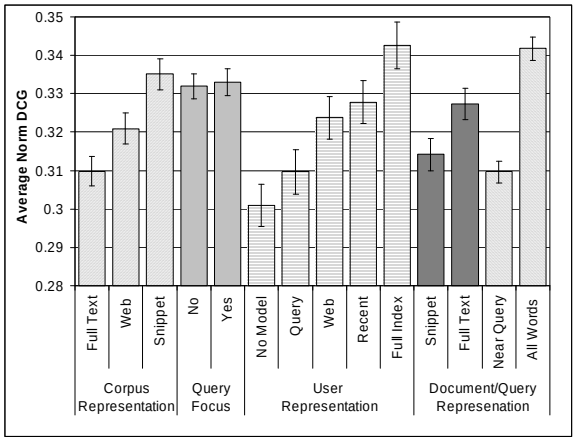

Figure 2. Average normalized DCG for different variables, shown with error bars representing the standard error about the mean. Richer representations tend to perform better.

5. RESULTS

We experimented with 67 different combinations of how the corpus, users, and documents could be represented, as discussed above, and used these combinations to re-rank Web search results. We also explored several different baselines.

We first present the results of ranking the top fifty documents for a query based purely on their textual features, ignoring the ranking returned by the Web search engine. We compare the best parameter settings for personalizing the search results with several baselines. We then report on augmentations of the personalized content-based rankings that incorporate the Web ranking. Web rankings take into account many factors, including textual content, anchor text, and query-independent importance factors such as PageRank.

5.1 Alternative Representations

The many different combinations of corpus, user, and document representations we explored resulted in a complex experimental design. For ease of presentation, we first summarize one-way effects in which we hold all but one variable constant, and explore the effects of varying that variable (e.g., User Representation – No model, Queries, Web pages, Recent index, Full index). This approach does not examine interaction effects, so at the end of this section we also summarize findings from the best combination of variables.

Results of the one-way analyses are shown in Figure 2. The scores reported in Figure 2 are the normalized DCG for the 131 queries in our test set, averaged across levels of the other variables. Analyses for statistical significance were performed using two-tailed paired *t*-tests The key effects are summarized in Table 1, along with their statistical significance levels.

The one-way sensitivity analysis showed that a rich representation of both the user and the corpus was important. The more data used to represent the user (*User Representation* in Figure 2), the better. Performance with the user's entire desktop index was best (*Full Index*), followed by representations based on subsets of the index (*Recent*ly indexed content, and *Web* pages only). Using only the user's query history (*Query*) or no user-specific representation (*No Model*) did not perform as well. Similarly, the richer the document and query representation (*Document/Query Representation*), the better the performance. We found that the best personalization occurred when using the full document text to represent the document (*Full Tex*t), rather than its snippet (*Snippet*), and performance was better when using all words in the document (*All Words*) than when using only the words immediately surrounding the query terms (*Near Query*).

The only place that more information did not improve the ranking was in the corpus representation (*Corpus Representation*). Representing the corpus based only on the title and snippets of the documents in the result set (*Snippet*) performed the best. An explanation for this finding is that when using only documents related to the query to represent the corpus, the term weights represent how different the user is from the average person who submits the query. We were interested to find that using the result set statistics performs both better (when the title and snippet are used) and worse (when the *Full Text* is used) than the Web (*Web*). One potential advantage to using the title and snippets to represent the corpus instead of using the full text is that the snippets are relatively uniform in length relative to the full text of the documents. Thus, one or two long documents do not dominate the corpus statistics. Another possible advantage of using the snippets is that they extract query-relevant portions of the document, which is important for documents that cover more than one topic.

Because of the query expansion performed during the result re-ranking, the user's query does not necessarily play a significant role in re-ranking. However, emphasizing the query during re-ranking does not appear to be necessary. Using all terms for query expansion was significantly better than using only the terms immediately surrounding the user's query (*Document/Query Representation*, *All Words* vs. *Near Query*). Using query focused corpus and user representations (*Query Focus*, *Yes*) showed no significant difference from a non-query focused representations (*Query Focus*, *No*). It could be that a query focus provides some benefit, but that the tradeoff between having a query focus and using more information in the ranking is relatively balanced. Alternatively, the lack of importance of the query could be

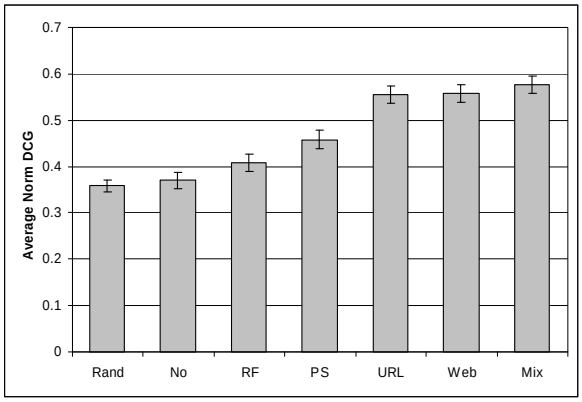

Figure 3. Personalized search (*PS*) compared with a random ranking (*Rand*), no user model (*No*), relevance feedback (*RF*), URL boost (*URL*), the Web (*Web*), and personalized search combined with the Web (*Mix*).

because all of the documents being ranked are more or less relevant to the query, making the primary job of the personalization to match the user.

These results indicate that, although the corpus can be approximated, a rich document representation and a rich user representation are both important. In practice, a system must choose between performing personalization on the client, where the rich user representation resides, or on the server side, where rich document representations reside. Thus, we also looked at the interaction between parameters. We found that it was more important to have a rich user profile than to have a rich document representation.

Because the parameters interact, the relationship is somewhat more complex than suggested by the results of the one-way analyses reported in Figure 2. The best combination of parameters we found was:

> **Corpus Representation:** Approximated by the result set title and snippets, which is inherently query focused.
>
> **User Representation:** Built from the user's entire personal index, query focused.
>
> **Document and Query Representation:** Documents represented by the title and snippet returned by the search engine, with query expansion based on words that occur near the query term.

This parameter combination received a normalized DCG of 0.46, and was the best combination selected for each query using leave-one-out cross-validation.

The corpus and user representations for the best parameter combination are consistent with what we found in the one-way sensitivity analyses. However, the document representation differs. The best combination calls for documents to be represented by their titles and snippets, rather than their full text. This makes sense given that the corpus representation is based on the documents titles and snippets as well. Corpus statistics are not available for terms that appear in the full text but not the snippet.

In addition to performing well, this combination is easy to implement entirely on the client's machine, requiring only the download of the search engine results. Thus, we investigated it further in comparison with several non-personalized baseline conditions.

5.2 Baseline Comparisons

To assess how well personalization performed, we also compared the results with several key baselines. These comparisons can be seen in Figure 3. The scores reported in Figure 3 are the normalized DCG for the 131 queries in our test set, and statistical analyses were performed using two-tailed paired *t*-tests with 130 degrees of freedom. The results of the best personalized search algorithm are shown in the bar labeled *PS*. The baseline conditions include: random ordering of the top-50 results (*Rand*), no user model (*No*), and an idealized version of relevance feedback (*RF*). For the cases of no user model and relevance feedback, we used a BM25 ranking based on the same content as the personalized search algorithms, with the best corpus and document/query representation selected for each. Only the user representation (R, r_i) differed. In the no user model case, R and r_i were equal to zero, and for the relevance feedback case, they were based on the documents in the evaluation test set that the user marked as *highly relevant* or *relevant*.

Not surprisingly, personalized search re-ranking (*PS*) significantly outperformed a random ordering of search results (*Rand*, $p<0.01$), and search with no user model (*No*, $p<0.01$). We were somewhat surprised to find that Web search personalization also performed somewhat better than ideal relevance feedback (*RF*, $p<0.05$). While this result may seem counterintuitive, it is important to note that in relevance feedback, the relevant documents are used to expand the terms considered and to modify the term weights. This does not guarantee that the documents used for augmentation will be at the top of the re-ranked list. In addition, the rich user profile used in *PS* may contain useful discriminating terms that are not present in the relevant documents in the top-50 results.

Figure 3 also shows a comparison of the best personalized content-based rankings with the Web ranking (*Web*). Personalized search performed significantly worse ($p<0.01$) than the Web rank, which had a normalized DCG of 0.56. This is probably because Web search engines use information about the documents they index in addition to the text properties, most notably linkage information, and this has been shown to improve results for many search tasks, including those of finding homepages finding and identifying resources [8]. For this reason, we looked at incorporating into our personalization algorithm the ranking information returned by the search engine.

5.3 Combining Rankings

To understand whether there was potential value in combining Web rankings and personalized results, we examined how similar these two rankings were to each other and to the user's ideal ranking. To construct the user's ideal ranking, the documents the user considered *highly relevant* were ranked first, *relevant* next, and *not relevant* last. Because such a ranking does not yield a completely ordered list, we computed the Kendall-Tau distance for partially ordered lists [1] to measure the similarity of rankings. The Kendall-Tau distance counts the number of pair-wise disagreements between two lists, and normalizes by the maximum possible disagreements. When the Kendall-Tau distance is 0, the two lists are exactly the same, and when it is 1, the lists are in reverse order. Two random lists have, on average, a distance of 0.5. The personalized ranking was significantly closer to the ideal

than it was to the Web ranking (τ = 0.45 vs. τ = 0.49, p<0.01). Similarly, the Web ranking was significantly closer to the ideal than to the personalized ranking (τ = 0.45 vs. τ = 0.49, p<0.05). The finding that the Web ranking and the personalized ranking were both closer to the ideal than to each other suggests that what is good about each list is different.

In an attempt to take advantage of the best of both lists, we merged the Web ranking with the personalized text-based ranking. For the personalized results, we had BM25 match scores. For the Web results, we only had access to rank information. For this reason, we considered only rank information in the merge. To merge the two lists, we weight each position in accordance with the probability that a result at that position from the Web is relevant. This probability was computed based on all queries in our test set except the one being evaluated. Because the probability curve typically drops quickly after the first couple of results, merging results in this way has the effect of keeping the first couple of results similar to the Web, while more heavily personalizing the results further down the list. The combination of the Web ranking and the personalized ranking (*Mix* in Figure 3) yielded an average normalized DCG of 0.58, a small but significant (p<0.05) improvement over the Web's average normalized DCG of 0.56. In contrast, boosting previously visited URLs by merging them with the Web results (*URL*) yielded no significant change to the Web's average normalized DCG.

Note that the personalization algorithm used in this analysis was not selected because it produced the best results when merged with the Web ranking, but because it performed well on its own and is feasible to implement. Greater improvement might be obtained by selecting the parameter setting that produces the best results when merged with the Web ranking. We also believe that we could further improve the advantages of personalization by combining server and client profiles in richer ways. For example, in the same way that we personalize content-based matching, we could personalize link-based computations as proposed by Jeh and Widom [10]. Richer application programming interfaces (APIs) to Web indices could also provide richer corpus or document statistics, further improving our ability to personalize both content-based and query-independent factors in ranking.

6. CONCLUSION AND FUTURE WORK

We have investigated the feasibility of personalizing Web search by using an automatically constructed user profile as relevance feedback in our ranking algorithm. Our research suggests that the most successful text-based personalization algorithms perform significantly better than explicit relevance feedback where the user has fully specified the relevant documents, and that combining this algorithm with the Web ranking yields a small but statistically significant improvement over the default Web ranking. Although many successful algorithms rely both on a rich user profile and a rich corpus representation, it is possible to approximate the corpus, making efficient client-side computation feasible.

There are a number of interesting directions to investigate. We now discuss several avenues for research and look at real-world challenges with deploying a personalized search system. We also consider interface design and interaction issues that arise in personalizing search.

6.1 Further Exploration

The parameters of the personalization procedure explored in this paper represent only a small subset of the space of parameterizations. As an example of an extension, the user profile could incorporate a more complex notion of time and current interest by being more influenced by documents seen recently than documents seen a long time ago. Within the BM25 framework, we could explore tuning the algorithm's parameters, using more complex terms (*e.g.*, phrases), and incorporating length normalization. We could also examine alternative frameworks for incorporating differential term weighting to personalize search.

In our experiments, no one parameter setting consistently returned better results than the original Web ranking, but *there was always some parameter setting that led to improvements*. This result highlights the opportunity for using machine learning to select the best parameters based on the query at hand. This selection could be done based on the individual (*e.g.*, the user's interests change often, so recently seen documents might be weighted more heavily in the construction of the user profile), the query (*e.g.*, the query term is very common in the user's personal index, suggesting that a great deal of personalization is needed), and properties of the result set (*e.g.*, the documents in the result set have widely varying scores, suggesting that personalization would be useful).

Differences across queries are particularly interesting to explore. There was some indication that personalization was more effective for shorter queries and more ambiguous queries (measured by the number of Web documents matching the query). For example, the queries *discount hotel London*, *activities Seattle* and *trebuchet* improved with personalization but the queries *habanero chiles*, *Snoqualmie ridge QFC* and *Pandora ranking 60 level-1 level-2* did not. However the effects were quite variable and were not statistically reliable given the relatively small numbers of queries used in our experiment.

In another direction, we could introduce additional classes of text- and non-text based content and activities in the construction of interest profiles. These profiles could incorporate information about activities such as current or recent software application usage, the pattern of topics of content at the current or recent focus of attention, and the history of locations visited by the user as sensed via GPS and other location-tracking methodologies.

6.2 Making Personalization Practical

There are a number of practical issues to consider in deploying a personalized Web search engine, especially if the user and corpus models exist on different machines. Making key information available, via programmatic interfaces for personalization, may mitigate a number of the issues. For example, snippets could be designed for personalization, and richer information about corpus statistics and score could be provided for merging.

For efficiency reasons, it would also be good to determine the ideal number of documents returned from a Web search engine so as to provide an appropriate substrate for a re-ranking procedure. This might be dependant on the user and the query, and query expansion might sometimes be necessary to retrieve the appropriate documents for re-ranking. For example, a user from San Francisco searching for "weather" may retrieve at least one Web page on Bay Area weather in the top results, and thus will find what they want at the top of a re-ranked list following

personalization. On the other hand, a person from Centralia, PA (population 21), probably will not have the same success unless the number of pages considered for re-ranking is increased.

6.3 On the User Interface

There are also a number of design and research challenges in the realm of human-computer interaction when personalizing search. We have explored several designs for displaying the initial and personalized lists of results. A good interface should help users understand how the personalization occurs, for example, by highlighting those terms that most influence the re-ranking. Personalization interfaces should also provide the user with control over key personalization parameters. Exposing all of the parameters will likely be of little use, but an interface that, for example, initially returns the Web search engine ranking and allows the user to personalize the results with a slider, would allow the user to control in a smooth manner the importance of personalization in the displayed list. There are also challenges and opportunities in handling the potential volatility of personalized (and non-personalized) rankings for the same query issued over time [24]. Personalized results for the same query can change as the user's profile changes. Tools for caching previous results for the same queries, or for allowing a user to control the profile used may be valuable for allowing users to retrieve documents they have successfully accessed in the past.

Our research shows that it is possible to provide effective and efficient personalized Web search using a rich and automatically derived user profile. We are currently extending the work by examining new algorithmic and user interface approaches outlined above to further improve our ability to personalize search.

7. REFERENCES

[1] Adler, L. M. (1957). A modification of Kendall's tau for the case of arbitrary ties in both rankings. *Journal of the American Statistical Society, 52:* 33-35.

[2] Anick, P. (2004). Using terminological feedback for Web search refinement: a log-based study. In *Proceedings of WWW '04*, 89-95.

[3] Bharat, K. (2000). SearchPad: Explicit capture of search context to support Web search. In *Proceedings of WWW '00*, 493-501.

[4] Budzik, J. and Hammond, K. (1999). Watson: Anticipating and contextualizing information needs. In *Proceedings of ASISIT '99*, 727-740.

[5] Dumais, S. T., Cutrell, E., Cadiz, J. J., Jancke, G., Sarin, R. and Robbins, D. (2003). Stuff I've Seen: A system for personal information retrieval and re-use. In *Proceedings of SIGIR '03*, 72-79.

[6] Gauch, S., Chafee, J. and Pretschner, A. (2004). Ontology-based personalized search and browsing. *Web Intelligence and Agent Systems, 1(3-4):* 219-234.

[7] Google Personal. http://labs.google.com/personalized

[8] Hawking, D. and Craswell, N. (2001). Overview of the TREC-2001 Web Track. In *Proceedings of TREC '01*, 61-68.

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

[10] Jeh, G. and Widom, J. (2003). Scaling personalized Web search. In *Proceedings of WWW '03*, 271-279.

[11] Kelly, D. and Teevan, J. (2003). Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum, 37(2):* 18-28.

[12] Koenmann, J. and Belkin, N. (1996). A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Proceedings of CHI '96*, 205-212.

[13] Kritikopoulos, A. and Sideri, M. (2003). The Compass Filter: Search engine result personalization using Web communities. In *Proceedings of ITWP*.

[14] Liu, F., Yu, C. and Meng, W. (2002). Personalized Web search by mapping user queries to categories. In *Proceedings of CIKM '02*, 558-565.

[15] McKeown, K. R., Elhadad, N. and Hatzivassiloglou, V. (2003). Leveraging a common representation for personalized search and summarization in a medical digital library. In *Proceedings of ICDL '03*, 159-170.

[16] Morita, M. and Shinoda, Y. (1994). Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of SIGIR '94*, 272-281.

[17] Nielsen, J. Personalization is overrated. In Jakob Nielsen's Alertbox for October 4, 1998. http://www.useit.com/alertbox/981004.html.

[18] Pitkow, J., Schutze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E. and Breuel, T. (2002). Personalized search. *Communications of the ACM, 45(9):* 50-55.

[19] Ruthven, I. and Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review, 18(2):* 95-145.

[20] Shen, X. and Zhai, C. X. (2003). Exploiting query history for document ranking in interactive information retrieval. In *Proceedings of SIGIR '03 (Poster)*, 377-378.

[21] Sparck Jones, K., Walker, S. and Robertson, S. A. (1998). Probabilistic model of information retrieval: Development and status. Technical Report TR-446, Cambridge University Computer Laboratory.

[22] Speretta, M. and Gauch, S. (2004). Personalizing search based on user search history. Submitted to *CIKM '04*. http://www.ittc.ku.edu/keyconcept/

[23] Sugiyama, K., Hatano, K. and Yoshikawa, M. (2004). Adaptive Web search based on user profile constructed without any effort from user. In *Proceedings of WWW '04*, 675-684.

[24] Teevan, J. (2005). The Re:Search Engine: Helping people return to information on the Web. To appear in *Proceedings of SIGIR '05 (Doctoral Consortium)*.

[25] Teevan, J., Alvarado, C., Ackerman, M. S. and Karger, D. R. (2004). The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Proceedings of CHI '04*, 415-422.

[26] Teevan, J., Dumais, S. T. and Horvitz, E. (2005). Beyond the commons: Investigating the value of personalizing Web search. In *Proceedings of the Workshop on New Technologies for Personalized Information Access (PIA)*.