Potential for Personalization

Jaime Teevan, Susan T. Dumais, and Eric Horvitz, Microsoft Research [1]

Abstract

Current Web search tools do a good job of retrieving documents that satisfy the wide range of intentions that people associate with a query – but do not do a very good job of discerning different individuals' unique search goals. In this paper we explore the variation in what different people consider relevant to the same query by mining three data sources: 1) *explicit* relevance judgments, 2) clicks on search results (a *behavior-based implicit* measure of relevance), and 3) the similarity of desktop content to search results (a *content-based implicit* measure of relevance). We find that people's explicit judgments for the same queries differ greatly. As a result, there is a large gap between how well search engines could perform if they were to tailor results to the individual, and how well they currently perform by returning results designed to satisfy everyone. We call this gap the *potential for personalization.* The two implicit indicators we studied provide complimentary value for approximating this variation in result relevance between people. We discuss several uses of our findings, including a personalized search system that takes advantage of the implicit measures to improve users' search experiences by ranking personally relevant results more highly and improving click-through rates.

ACM Classification: H3.3 [Information storage and retrieval]: Information search and retrieval – *Query formulation;* H5 [Information interfaces and presentation (e.g., HCI)]

General terms: Experimentation, Human Factors

Keywords: Personalized search, Web search, individual differences, user modeling

[1] One Microsoft Way, Redmond, WA, USA 98052, {teevan, sdumais, horvitz}@microsoft.com

1. Introduction

The interface to most Web search engines is simple. A user enters a few words into a search box, and receives a long list of results in return. Despite the simplicity of the interaction, people use Web search engines for many complex tasks; people conduct research, plan trips, entertain themselves, purchase items, and find new jobs using Web search engines. The challenge for a search engine is to translate people's simple, short queries into lists of documents that satisfy their different information needs.

It is unlikely that a typical two- or three-word query can unambiguously describe a user's informational goal. For example, a Web search for "CHI" returns a wide range of results, including stock quotes for the Calamos Convertible Opportunities & Income Fund, Web pages about Chicago, details on balancing your natural energy (or ch'i) through acupuncture, and a few results about computer-human interaction. Although each of these results is relevant to some query for "CHI", readers of this paper would likely be uninterested in most of what is returned.

There have been attempts to help people articulate their information needs more clearly, either by supporting interactions with search results after an initial query (e.g., via query suggestions [Anick 2003], interactive query expansion [Koenemann and Belkin 1996, Ruthven 2003], or filtering [Dumais et al. 2003]), or through the elicitation of richer queries and background knowledge [e.g., Kelly and Fu 2007]. Although search interfaces can be improved to encourage less ambiguous queries in laboratory settings, there will always be many cases where people are unable to clearly articulate their needs because they lack the knowledge or vocabulary to do so, or where search engines cannot take good advantage of the additional information. In addition, there is evidence that it is easier cognitively to generate a short query and to use navigation to get to the desired information [Teevan et al. 2004]. The research presented in this paper compliments previous efforts to help people articulate their search

target by characterizing the variability in what different searchers find relevant to the same query and suggesting ways that this variation can be successfully captured to help people find what they are looking for.

To better understand the variation in what people using the same query are searching for, we examine explicit relevance judgments and implicit indicators of user interest. We develop analytical techniques to summarize the amount of variation across individuals (*potential for personalization* curves), and compare different data mining techniques for generating these measures. Through analysis of the explicit relevance judgments made by different individuals for the same queries, we find that significant variation exists not only when people have very different underlying information needs (e.g., "computer-human-interaction" vs. "ch'i"), but also when they appear to have very similar needs (e.g., "key papers in human-computer interaction" vs. "important papers in human-computer interaction"). While explicit relevance judgments are impractical for a search engine to collect, search engines do typically have a large amount of implicit data generated through users' interactions with their service. By exploring several of these sources of implicit data, we find it is possible to use them to approximate the variation in explicit relevance judgments and improve the user experience. Implicit measures that are behavior-based (e.g., related to the similarity of a result to previously visited URLs) appear to hold potential for capturing relevance, while measures that are content-based (e.g., related to the similarity of a result to other electronic content the individual has viewed) appear to hold the potential for capturing variation across individuals.

Our findings can be used in many ways to improve the search experience. We briefly describe one case study showing how the implicit measures we studied in this paper can be used to personalize search results. Personalization can make it possible for a search for "CHI" to return results like the TOCHI homepage for the computer-human interaction researcher, stock quotes for the Calamos fund for

the financial analyst, and pages about ch'i for the feng shui practitioner. By using both behavior- and content-based implicit measures of relevance, our search system captures the observed notions of relevance and variation between individuals necessary to significantly improve the search experience. We show the personalized system helps people to find what they are looking for even when searching using the easy, underspecified queries that come naturally to them.

We begin the paper with a discussion of related work. We then present the three data sets we have mined to learn more about what people consider explicitly and implicitly relevant to a query. The first data set consists of different individual's explicit relevance judgments for 699 queries, the second of the search engine click logs for 2,400,645 queries, which provides an implicit behavior-based relevance judgment, and the third of term frequencies of content stored on individual's desktops for 822 queries, which provides an implicit, content-based relevance judgment. Using these three data sets, we look at how a result's Web search rank correlates with its judged relevance, and find that many relevant results are ranked low. Because these low ranked relevant results are judged differently by different individuals, we then show that ranking results for an individual instead for a group of people has great potential to improve the quality of the search experience. Finally, we discuss ways our findings can be used and present a system that uses the different implicit measures of relevance explored to personalize search results.

2. Related Work

Most research into the development and study of search engines has focused on providing the same results to everyone who issues the same query, without consideration of the searcher's individual context. For example, the most common way to assess the relevance of a result to a query is to ask an expert judge to explicitly rate its relevance. The intent is that two different judges should assign the same rating to the same result. Judges are given a detailed description of an information need with a

query (it might be their own need or one expressed by another searcher), and asked to provide judgments about the topical relevance of the document, rather the quality of the document or the relevance to an individual searcher.

There have been efforts to measure inter-rater agreement in relevance judgments given the same detailed description of an information need [e.g., Harter 1996; Voorhees 1997]. In our work we also explore the agreement between different judges for the same queries, but focus, in contrast, on judgments made given the judge's individual search intent. Our judges are not asked to judge whether a result is a reasonable one for a given need, but rather whether it is what they personally would want for that query. In an earlier poster [Teevan et al. 2007a] we reported some preliminary results from our explorations in this area. In this paper, we extend the earlier work by analyzing a richer data set of explicit judgments as well as two additional sources of implicit measures of relevance obtained through content analysis and click behavior, and show an example of how our findings can be used to personalize search results. In the information science community, there has been a good deal of work on evaluating the relevance of results to a query. Understanding "relevance" is a complex problem [Mizzaro 1997; Saracevic 1976; Saracevic 2006; Schamber 1994], and in our work we address only the portion of that pertains to individual assessments of relevance.

In addition to explicit ratings of relevance, several groups have examined using implicit measures to model user interest and relevance. Kelly and Teevan [Kelly and Teevan 2003] provide an overview of this work. For example, Claypool et al. [Claypool et al. 2001] examined the relationship between explicit judgments of quality and implicit indicators of interest such as time and scrolling for general Web browsing activities. Morita and Shinoda [Morita and Shinoda 1994] measured the relationship between reading time and explicit relevance judgments. Joachims et al. [Joachims et al. 2005] looked at the relationship between clicks and explicit judgments for a search task. Agchtein et al. and Fox et al.

[Agchtein et al. 2006; Fox et al. 2005] developed more complex learned models to predict relevance judgments or preferences using a variety of implicit measures including clicks, dwell time and query reformulations. A recent paper by Wu et al. [Wu et al. 2008] investigated the variability in results that users click on for the same query. They measured the agreement in clicks independent of position, using a small number of queries (135) from a close-knit community. Our behavior-based analyses extends this line of work by developing additional measures of agreement using potential for personalization curves and examining a much larger number of queries (more than 44,000) for a more heterogeneous set of users. New research in the information foraging literature is beginning to examine social foraging including an analysis of different user's previous history of Web browsing [Chi and Pirolli 2006], although this work has not been used to personalize search as far as we know. Our work extends these efforts by focusing on differences in implicit measures across individuals, and by using content as well as interaction data to model users' information goals.

As a result of our analyses, we provide an example of how what we learn about people's variation in intent can be applied to improve personalized search. There are a few publicly available systems for personalizing Web results (e.g., http://www.google.com/psearch), but the technical details and evaluations of these commercial systems are proprietary. Pitkow et al. [Pitkow et al. 2002] describe two general approaches to personalizing Web search results, one involving modifying the user's query and the other re-ranking search results. The approach we take is the latter. Research systems that personalize search results model their users in different ways. Some rely on users explicitly specifying their interests [Ma et al. 2007], or on demographic/cognitive characteristics [Frias-Martinez et al. 2007], but user supplied information can be hard to collect and keep up to date. Others have built implicit user models based on content the user has read or their history of interaction with Web pages (e.g.,[Chirita et al. 2006; Dou et al. 2007; Shen et al. 2005; Sugiyama et al. 2004; Teevan et al. 2005]).

In this paper we briefly summarize an operational personalized search system that we developed that uses implicit measures [Teevan et al. 2005], and highlight how the insights derived from our analyses can help us to understand how the system performs. The evaluation of personalized search systems typically consists of measuring aggregate user performance for a small number of queries and users, and showing that personalized results lead to better search accuracy or user preferences. The results we presented earlier of our personalized search system [Teevan et al. 2005] are expanded in this paper to include the analysis of many more queries (699 queries v. 131 queries) as well as a more in-depth discussion of the effects of varying the weight of different implicit factors when ranking. We also provide evidence that personalized search algorithms positively impact user behavior with a longitudinal study of the system's use.

To summarize, the research reported in this paper builds on earlier research on relevance and search personalization, but examines in detail differences in search intents for searchers issuing the same query and formalizes this notion using a measure that summarizes the *potential for personalization*. We compare relevance measure obtained using three different sources of evidence (explicit ratings, click behavior, and content), and describe a prototype search system that uses these ideas.

3. Methods and Data Sets

There are numerous ways to assess whether a document is relevant to a query. In this paper we explore three different ways, and each is described in greater detail in this section. For one, we use explicit relevance judgments, which are the most commonly used measure for assessing the relevance of a document to a user's query in the information retrieval literature. We also use two implicit measures of relevance, one content-based (how similar is the content of a result to content a person has seen before), and the other behavior-based (how likely is a person to have visited the document before,

for how long, etc.). These two broad classes of implicit measures are the most commonly used in operational systems for personalization.

The data we collected for each of the three measures are summarized in Table 1. We collected explicit relevance judgments for 699 queries, content-based implicit relevance judgments for 822 queries, and behavior-based implicit relevance judgments for 2,400,654 queries. Because we are concerned with the variation in judgment across individuals for the same query, the table also provides information about how many of the queries in the dataset are unique, and how many have relevance judgments from multiple different users. In this section we describe in detail how each set of judgments was collected, and in particular how we were able to obtain many sets of judgments for the same queries. In subsequent sections we use these measures to understand how well Web search engines currently perform, how well they could ideally perform if they tailored results to each individual, and how well a proposed personalized search system does perform.

Table 1. This paper explores three types of relevance measures: explicit relevance judgments, implicit content-based measures, and implicit behavior-based measures. This table lists the number of people from whom each measure was gathered, the total number of queries gathered for each, the number of unique queries, and the number of queries with judgments from more than one individual. Also listed is how each measure is quantified, which is labeled *gain*.

Relevance Measure		# Users	# Queries	# Unique	>5 Users	Gain
Explicit judgments		125	699	119	17 [Table 2]	2 if highly relevant 1 if relevant 0 if irrelevant
Implicit	Content	59	822	24	24	Cosine similarity
	Behavior	1,532,022	2,400,645	44,002	44,002	1 if clicked 0 if not clicked

3.1 Explicit Judgments

The most straightforward way to determine whether an individual considers a result relevant to a query is to explicitly ask that individual. The TREC benchmark collections used in the evaluation of information retrieval systems, for example, are developed using such explicit judgments [Voorhees and Harman 2005]. In TREC, expert judges are asked to rate the relevance of results to a query based on a detailed description of an information need. The following is an example of a TREC topic:

<num> Number: 403

<title> osteoporosis

<desc> Description:

Find information on the effects of the dietary intakes of potassium, magnesium and fruits and vegetables as determinants of bone mineral density in elderly men and women thus preventing osteoporosis (bone decay).

<narr> Narrative:

A relevant document may include one or more of the dietary intakes in the prevention of osteoporosis. Any discussion of the disturbance of nutrition and mineral metabolism that results in a decrease in bone mass is also relevant. The purpose of the topic is to unambiguously describe an information goal. While discrepancy in what judges consider relevant have been noted even with such detailed topics, the goal is to maximize inter-judge agreement in interpreting the query intent.

The TREC topic scenario is unrealistic for Web search, where it is well-known that people issue very short queries to describe their information needs [Spink and Jansen 2004]. It is unlikely that the same short Web query, when issued by two different people, has the same unambiguous information goal behind it. In addition, TREC studies focus on whether documents are topically relevant to the query and not whether an individual user would be satisfied with the document. In this paper we are interested in understanding what people consider individually relevant to typical Web queries (i.e., what documents would satisfy their information need even when it is expressed with a short ambiguous query). For this reason, rather than have people evaluate results for fully defined information goals, we asked our

judges to indicate which results they *personally would consider relevant* to an information need specified with a more typical Web query.

Participants in our study were asked to evaluate how personally relevant the top 40+ Web search results for six to ten queries were to them. Web search results were collected from Live Search, and presented to participants in the same format as Web results are normally shown, with a title, snippet and URL. The actual result page could be seen by clicking on the title or URL, but was only viewed when the participant felt doing so was necessary to make a judgment. For each search result, each participant was asked to determine whether they personally found the result to be *highly relevant*, *relevant*, or *not relevant* to the query. So as not to bias the participants, the results were presented in a random order. Relevance judgments made by assessing a list of randomly ordered search results do not necessarily reflect exactly what would satisfy that person's information need if they were to issue the associated query, particularly as the query may appear as part of a session and analysis of the results may include browsing behavior to pages beyond those contained in the result set. But the judgments do serve as a reasonable proxy for relevance, and the approach is consistent with previous approaches for measuring search result quality.

The queries evaluated were selected in two different manners, at the participants' discretion. In one approach (*self-generated queries*), participants were asked to choose a query that mimicked a search they had performed earlier that day, based on a diary of Web searches they had been asked to keep. In another approach (*pre-generated queries*), participants were asked to select a query from a list of queries. For a subset of the self- and pre-generated queries, participants were asked to write a more detailed description of the informational goal or intent they had in mind when they issued the query, or were interested in for the pre-generated queries.

Using these two types of queries allowed us to balance that value that could be obtained by studying naturally occurring self-generated queries with the need to collect multiple judgments for the same query. Asking people to select results from a pre-generated list enabled us to explore the consistency with which different individuals evaluated the same results to the same query. Such data would have been very difficult to collect using only self-generated queries, since it would have required us to wait until different participants coincidentally issued the same query on their own. Thus the pre-generated queries provide a way to obtain overlap in queries across people. Participants were encouraged to select only pre-generated queries that were of direct interest to them. The intent was that this selection process would somewhat mitigate the artificial nature of using pre-generated queries. We collected self-generated queries so that we could directly compare overall patterns of judgments for self-generated and pre-generated queries and explore any potential discrepancies. Russell and Grimes [Russell and Grimes 2007] have shown that searchers behave somewhat differently for assigned and self-generated search tasks (e.g., spending more time on tasks but generating fewer queries). As we describe in more detail below in our discussion of Figure 1, there are no differences in the overall distribution of explicit relevance judgments for the two types of queries used in our studies. We suspect that our more focused explicit relevance task minimized such differences compared with Russell and Grimes' analyses of search sessions.

Participants were all employees of Microsoft. All were computer literate and familiar with Web search. They came from variety of backgrounds, including administration, consulting, legal, product support, research, sales, and software development. The pre-generated queries were chosen to be of general interest to Microsoft employees, and reasonably common based on an examination of general query logs and logs of queries issued from within Microsoft. Judgments were done on the participants own machine at their own pace over a few days so timing estimates are quite rough, but the average time per participant was one to two hours.

Ignoring queries where fewer than 40 results were evaluated, we collected explicit relevance judgments for 699 queries associated with 119 unique queries. Of the queries, 601 were pre-generated and 98 were self-generated. The number of people who evaluated the results for the same pre-generated query ranged from three to 77. There were 17 unique queries that had judgments from more than 5 people. Two different lists of pre-generated queries were used, shown in Table 2; the different lists are associated with previous work [Teevan et al. 2007a, List I] and a new round of data collection [List II]. Fifty three of the pre-generated queries evaluated were associated with List I, and 548 with LII.

Table 2. In addition to generating their own queries, participants were encouraged to select queries to evaluate from one of two lists of pre-generated queries. The two lists of pre-generated queries are listed here, along with the number of participants who explicitly evaluated the results for each query.

List I

Query	Users
bush	3
cancer	6
gates	6
longhorn	8
microsoft	9
seattle	5
traffic	4
uw	6
Web search	4

List II

Query	Users
black & white photography	12
bread recipes	27
business intelligence	9
c# delegates	22
cat on computer	25
live meeting	18
microsoft new technologies	28
Photosynth	25
powershell tutorial	15
redmond restaurant	38
slr digital camera	24
toilet train dog	30

Explicit judgments are nice because they allow us to examine the consistency in relevance assessments across judges in a controlled setting. They do, however, also have some drawbacks. For

one, it is cumbersome for people to give explicit judgments and thus challenging to gather sufficient data to generalize across a broad variety of people, tasks, and queries. Explicit judgments are also typically captured outside of an end-to-end search session in which several queries may be issued and combined with navigation. For this reason, we supplement the explicit judgments with implicit measures of personal relevance and analyze the correspondence between implicit and explicit measures. Implicit measures are easier to collect, thus allowing us to analyze many more queries from a wider variety of searchers, but are also more difficult to interpret. We explore two implicit measures which are commonly used for personalization, content-based and behavior-based measures.

3.2 Behavior-Based Implicit Measures

Behavior-based measures of relevance use people's behavior, such as their prior interactions with search result lists, to infer what is relevant. Click-through is a common behavior-based proxy for relevance [Kelly and Teevan 2003]. We collected click-through data to use for this purpose by analyzing the fully anonymized logs of queries issued to Live Search. For each query instance, the logs contained a unique user ID, time stamp, and list of clicked results. Because in this paper we are interested in studying queries for which we have relevance judgments from multiple judges, only those queries issued by more than ten unique individuals were considered. In order to remove variability caused by geographic and linguistic variation in search behavior, we filtered for queries generated in the English speaking United States ISO locale. Our analyses are based on a sample of queries from October 4, 2007 to October 11, 2007, and represent 2,400,645 queries instances "evaluated" by more than 1,532,022 million unique users. Of the total queries, 44,002 are unique.

While the manually collected relevance data only represents a few hundred queries, because we were able to collect the behavior-based dataset implicitly, it includes information about millions of

queries issued by millions of users. Using this dataset we are able to study many different users' interactions with the same self-generated queries in a way that would be infeasible with explicit data.

3.3 Content-Based Implicit Measures

Content-based implicit measures of relevance use a textual representation of people's interests to infer which results are relevant to their current need. There are many ways of representing people's interests, including explicit user profiles, implicit profiles based on previous query history, and richer implicit profiles based on the full content of documents. In this paper we use a very rich interest profile based on the frequencies of terms in previously viewed documents. Such a representation can be obtained from a desktop index such as that described in Stuff I've Seen [Dumais et al. 2003] or available in desktop indices such as Google Desktop Search, Mac OS X Spotlight, Windows Desktop Search, X1 or Yahoo! Desktop Search. The system we used to collect relevance judgments based on content-based profiles indexes all of the information created, copied, or viewed by an individual. Indexed content includes Web pages that the individual has viewed, email messages that were viewed or sent, calendar items, and documents stored on the client machine.

The data we analyze in this paper is derived from a dataset collected by Radlinski and Dumais [Radlinski and Dumais 2006] for other purposes. This dataset allows us to measure how closely the top 40 search results for 24 unique queries matched 59 participants' user profiles. The 24 queries are shown in Table 3. To collect content-based implicit relevance judgments for these queries, the participants, all Microsoft employees, ran a simple software application on their computer. This application used standard information retrieval measures to calculate the similarity of the participants' content-based profile to each search result in a preset list, and reported the results back. Note that participants did not have to actually issue the queries to provide the relevance assessments for them.

Instead the measure of relevance was based entirely on the pre-existing content on their desktop computer. In total this dataset provided us with relevance judgments for 822 query instances.

Table 3. The 24 queries used to gather content-based information, and the number of users for each query.

Query	Users
animal control	59
aurora	52
bandwidth test	46
bank of america	45
bespelled	44
best buy	41
big brother	40
canada	39
circuit city	37
eiffel tower	36
first national bank	35
hawaii	34
hoffman	34
mercy hospital	29
painter	27
qvc	27
science direct	27
seattle map	27
t-shirt	26
union station	26
walmart	23
weather	23
world map	23
yahoo	23

Results that were more similar to a participant's user profile can be considered implicitly more relevant to the individual, and results that are further from the profile less relevant. To calculate the similarity of a result to a profile, the user profile, **u**, and each search result (or document), **d**, were

represented as vectors of terms. Terms were assigned scores using the BM25 weighting scheme [Sparck Jones et al. 1998]. The BM25 weight is intended to represent how "important" a term is in a document by taking into account its frequency in the document (a term that occurs more often in a document is more important) and its frequency in the corpus (a term that occurs more often in the corpus is less important). We then computed the cosine between the document vector and the user profile vector. For efficiency purposes, each search result was represented using the snippet returned by the search engine for the associated query (i.e., the page's title and a query focused summary of the content). So d_i is based on the number of times term *i* occurred in the title and snippet of the page.

4. Analysis of Rank and Rating

We begin our analysis of the three sets of relevance judgments by looking at how relevant the results ranked at different positions by the Web search engine were according to each measure, explicit and implicit.

4.1 Rank and Explicit Rating

Figure 1 shows the average relevance, or gain, for each result with an explicit relevance judgment as a function of rank. The gain used for each measure is summarized in Table 1. To compute the gain for results with explicit judgments, the rating *irrelevant* was given a gain of 0, *relevant* a gain of 1, and *highly relevant* a gain of 2. Values were averaged across all queries and all users. Separate curves are shown for the pre-generated (dashed line) and self-generated (solid line) queries from our earlier work and for the new pre-generated queries (dotted line). Clearly there is some relationship between rank and relevance. All curves show higher than average relevance for results ranked at the top of the result list. But it is important to note that there are still many relevant results at ranks 11-40, well beyond

what users typically see. A similar pattern is found using just binary ratings (gain of 0 for irrelevant and 1 for relevant or very relevant), although the mean is shifted down.

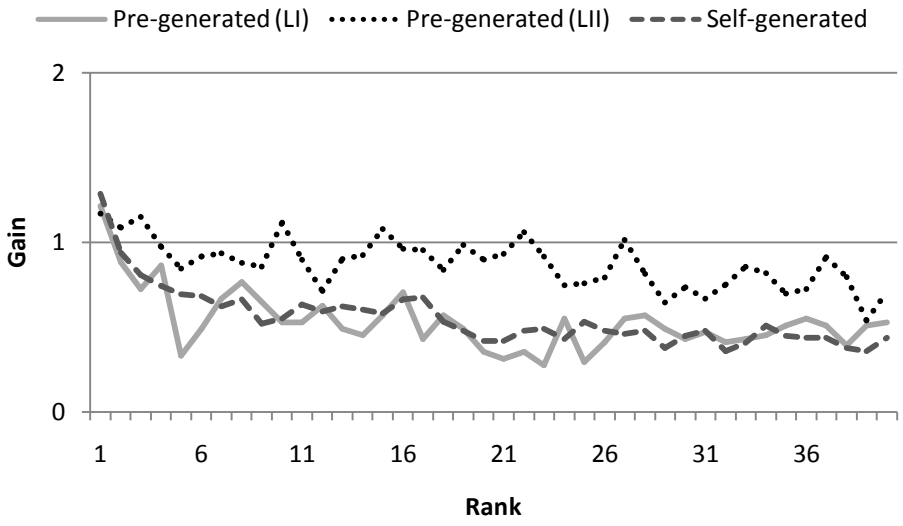


Figure 1. The average ratings for Web search engine results as a function of rank. Separate curves are shown for the pre-generated (solid line for queries from List I and dashed line for queries from List II) and self-generated (solid line) queries. While there is some relationship between rank and relevance, many relevant results are not ranked in the top ten.

Two other aspects of these results are worth noting. First, there are no differences between the overall ratings for pre-generated and self-generated queries in our study; the pre-generated (LI) and self-generated judgments were obtained from the same set of subjects and the lines in Figure 1 overlap consistently. This suggests that although there are sometimes differences in user behavior between assigned and self-generated search tasks, this does not appear to be the case for explicit relevance judgments. Second, there is a difference between the two sets of pre-generated queries, with judgments for the LII set being somewhat higher than the judgments for the LI set. It is not clear what

accounts for this difference. Microsoft employees served as judges in both cases. It could be that the LII queries are more specific or that the Live Search engine has improved in the two years between the LI and LII experiments.

The general pattern of results seen in Figure 1 is not unique to our sample of users or queries. A reanalysis of data from the TREC Web track [Hawking 2000] shows a similar pattern. In the TREC-9 Web track, the top 40 results from 50 Web queries were rated by trained judges using a similar three-valued scale, *highly relevant*, *relevant*, and *not relevant*. Results for three of the top-performing search systems are shown in Figure 2.

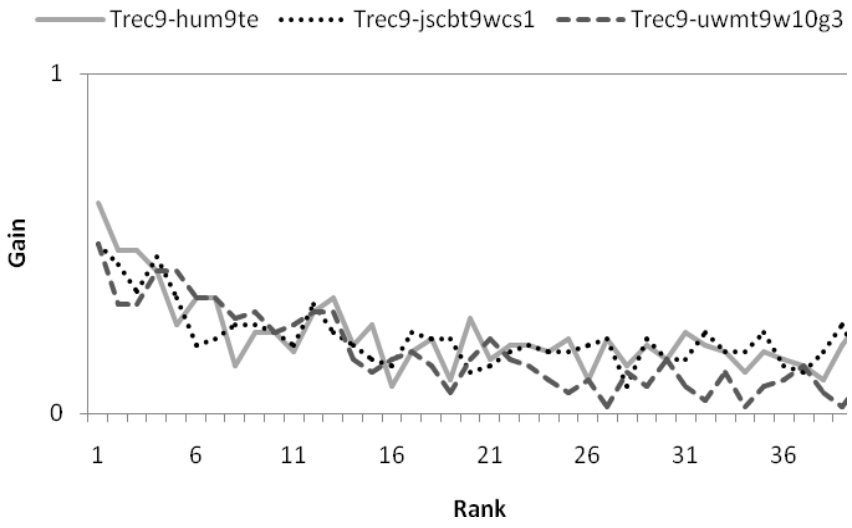


Figure 2. Average ratings for TREC Web track results as a function of rank. As in Figure 1, many relevant results are ranked below the top ten.

The above analysis shows that rank and rating are not perfectly correlated, and that there are many relevant documents in lower ranks. These findings suggest that if everyone rated the same low-ranked

documents as highly relevant, effort should be invested in improving the search engine's algorithm to rank those results more highly, thus making everyone happier. However, despite the many commonalities among our participants (e.g., all were employees of the same company, lived in the same area, and had similar computer literacy), deeper analyses of the data demonstrates a great deal of variation in their rating of the results for the same query.

Instead of rating the same documents as relevant, participants appeared to use the same query to mean very different things. This was evidenced by the variation in the explicit intents our participants wrote for 131 of the queries with explicit judgments (intent was collected in the initial set of queries, but not for the expanded collection). For example, the explicit intents we observed for the query "cancer" ranged from "information about cancer treatments" to "information about the astronomical/ astrological sign of cancer". Ranges were evident both for the pre-generated queries, where the user had to come up with an intent based on the query, and for the self-generated queries, where the query was generated to describe a pre-existing intent. Although we did not observe any duplicate self-generated queries, many self-generated queries, like "rice" (described as "information about rice university"), and "rancho seco date" (described as "date rancho seco power plant was opened") were clearly ambiguous.

Even when our participants expressed similar intents for the same query, they still rated the query results very differently. This highlights the difficulty of articulating information needs and suggests that the participants did not describe their intent to the level of detail required to distinguish their different goals. For example, for the query "Microsoft", three participants expressed these similar intents:

- "information about Microsoft, the company"
- "Things related to the Microsoft corporation"
- "Information on Microsoft Corp"

Despite the similarity of the stated intent, only one page (http://www.microsoft.com) was given the same rating by all three individuals. Twenty-six of the 40 results were rated *relevant* or *highly relevant* by one of these three people, and for only six of those 26 did more than one rating agree.

The observed disparity in rating likely arises because of ambiguity; the detailed intents people wrote were not very descriptive. Searches for a simple query term were often elaborated as "information on *query term*" (e.g., "UW" became "information about UW", leaving open whether they meant the University of Washington or the University of Wisconsin, or something else entirely). It appears our participants had difficulty stating their intent, not only for the pre-generated queries, where we expected they might have some difficulty creating an intent (mitigated by the fact that they only rated pre-generated queries by choice), but also for the self-generated queries.

Although explicit intents generally did not fully explain the query term, they did sometimes provide some additional information. For example, "trailblazer" was expanded to "Information about the Chevrolet TrailBlazer," clarifying the participant was interested in the car, as opposed to, for example, the basketball team. Further study is necessary to determine why people did not always include this additional information in their original query. It does suggest that there is some opportunity to develop interfaces that encouraged people to provide more information about their target when searching (e.g., [Kelly and Fu 2007]). However, even if people were to provide more information in their queries, they would probably still not be able to express exactly what wanted. For example, the Trailblazer example above did not clarify exactly what kind of information (e.g., pricing or safety ratings) was sought. This suggests that alternative methods should be explored to help searchers iteratively refine their needs during the course of a session using query suggestions or navigation support, or to enable search systems to better infer destinations from queries.

4.2 Rank and Implicit Rating

To complement and extend the results from explicit judgments of relevance, we explore whether content-based and behavior-based implicit measures of personal relevance behave similarly to the explicit manual judgments of personal relevance. We saw in Figures 1 and 2 that a significant number of results that people explicitly judge to be relevant exist at low ranks. The question explored in this section is whether content- or behavior-based proxies for relevance have the potential to identify these low-ranking relevant results.

Figure 3 shows the same type of graph presented in Figures 1 and 2 for behavior- and content-based data and for our explicit judgments. As summarized in Table 1, the gain used throughout this paper to represent the relevance of the behavior-based measure is 1 when a result was clicked, and 0 when a result was not clicked. Thus the average behavior-based gain for a result at a given rank is the probability of a result being clicked at that rank. The gain used to represent the relevance of the content-based measure is the cosine similarity between the result and a vector representing the user's interest. For comparison purposes, the curves are normalized so that the area under each sums to one.

The dotted line, representing the measure of relevance based on click-through data, is much higher for results ranked first or second than is to be expected given the explicitly collected relevance judgments (solid line). On the other hand, results that appear later in the ranking (e.g., results 9 and 10) receive significantly fewer clicks than they warrant based on the explicit relevance judgments. A person's behavioral interactions with a search result list appear, as expected given previous research [Guan and Cutrell 2007; Joachims et al. 2005], to be strongly influenced by presentation.

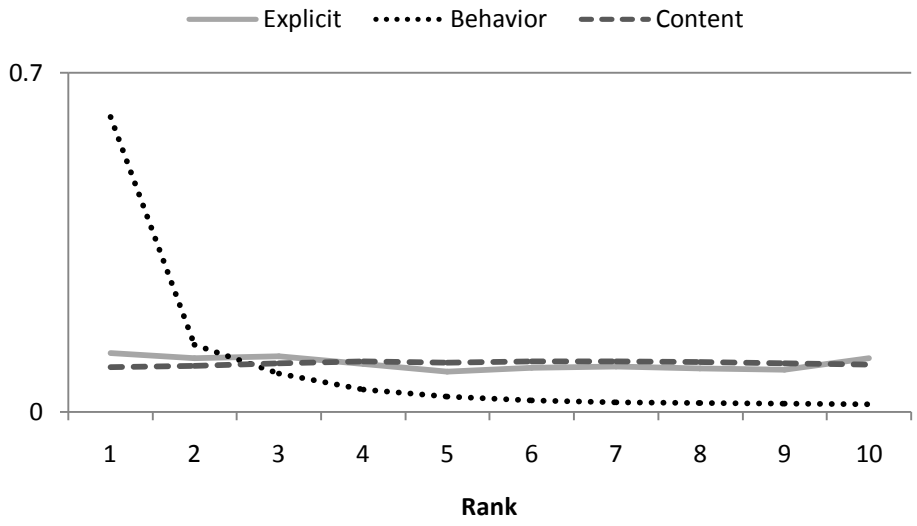


Figure 3. Average ratings for Web search engine results as a function of rank. Explicit relevance judgments (solid line) are compared with content-based (dashed line) and behavior-based (dotted line) implicit judgments. While the behavior-based judgments depend highly on rank, the content-based judgments do not.

In contrast, the content based curve (dashed line) is very flat, flatter even than the curve constructed from explicit relevance judgments. This may be because search results are all relevant topically (they all match the query words), while many other variables are used in judging personal relevance. For example, Fidel and Crandall [Fidel and Crandall 1997] showed that variables such as recency, genre, level of detail, and project relevance were important in determining relevance for an individual in addition to topic relevance.

5. Potential for Personalization

We have seen that Web search engines do not currently do a perfect job ranking results, and have provided some evidence that they are unlikely to be able to do so because of the variation in what

different people consider relevant to the same query. In this section we quantify the differences in result relevance between individuals. This analysis enables us to better understand the potential benefit to be gained from personalization. We first describe how this "potential for personalization" is quantified, and then describe findings using the explicit judgments and implicit measures.

5.1 Calculating the Potential for Personalization

To summarize the quality of a ranked list of results, we use *Discounted Cumulative* Gain (DCG), a measure commonly used for this purpose in information retrieval research [Järvelin and Kekäläinen 2000]. DCG summarizes the quality of a result set by counting the number of relevant results in the set, and further refines this simple measure with two important ideas: 1) that higher ranks should contribute more to the score, and 2) that very relevant items should contribute more to the score. DCG incorporates the idea that highly-ranked documents are worth more than lower-ranked documents by weighting the value of a document's occurrence in the list inversely proportional to its rank (i) thus providing a "discount" for lower ranks. The discount factor used in Equation 2 is $1/\log(i)$. DCG also incorporates the notion of multiple relevance levels by, for example, giving *highly relevant* documents a different "gain" value than *relevant* documents. As shown in Table 2, the gains ($G(i)$) are 2, 1, and 0 for highly relevant, relevant an non-relevant documents.

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

For each query, scores are summed ("cumulated") for all ranks giving us a single summary measure for the quality of a set of results. Because queries that have more relevant documents will have a higher DCG, the DCG is normalized to a value between 0 (the worst possible DCG given the ratings) and 1 (the best possible DCG given the ratings) when averaging across queries.

As an example, Table 4 shows the Web search results for the query "slr digital camera" and the gain associated with each result for two different users. In this example, User A rated four results as relevant to his information need, and User B rated one as very relevant and two as relevant to her information need. Using these scores we compute a DCG measure for each column. The normalized DCG is 0.52 for User A, and 0.23 for User B. On average, as shown in the column labeled *A+B*, the normalized DCG for the Web ranking is 0.38 for these two people.

Table 4. A ranked list of results for the query "slr digital camera" and the gain for two users based on their explicit judgments. The collective gain for both users (A+B) represents the quality of a result for the two of them. On average, the normalized DCG for the Web ranking for Users A and B is 0.38.

Web Result	Gain A	Gain B	A+B
usa.canon.com/consumer/controller?act=ProductCatIndexAct&fcategoryid=111	1	0	1
cameras.about.com/od/professionals/tp/slr.htm	1	1	2
cameras.about.com/od/camerareviews/ig/Digital-SLR-Camera-Gallery/index.htm	0	1	1
amazon.com/Canon-Digital-Rebel-XT-f3-5-5-6/dp/B0007QKN22	0	0	0
amazon.com/Canon-40D-10-1MP-Digital-Camera/dp/B000V5P90K	0	0	0
en.wikipedia.org/wiki/Digital_single-lens_reflex_camera	1	0	1
en.wikipedia.org/wiki/DSLR	1	2	3
olympusamerica.com/e1/default.asp	0	0	0
olympusamerica.com/e1/sys_body_spec.asp	0	0	0
astore.amazon.com/photograph-london-20	0	0	0
	User A	**User B**	**Avg**
Normalized DCG	0.52	0.23	0.38

If we take DCG value to be a summary measure of the quality of a ranked list of results, the best possible ranking for a query is the ranking with the highest DCG. DCG can be maximized by listing the results with the highest gain first. For example, for queries with explicit judgments where only one

participant evaluated the results, DCG can be maximized by ranking *highly relevant* documents first, *relevant* documents next, and *irrelevant* documents last. The best ranking of the results for "slr digital camera" for Users A and B individually can be seen in the two left columns of Table 5. Because these lists are the best possible for the individual, the normalized DCG for these rankings is 1 in each case. Note that this is an ideal case in which we have explicit judgments of how relevant each result is for each user. The best a search engine could do for this user is to match these judgments by returning the results in this order. There may be other ways of collecting judgments, e.g., after judges have looked thoroughly at each page or after they have completed their task. But, for this discussion we treat the explicit judgments of personal relevance that we have obtained as the gold standard.

Table 5. The best ranking of the results for "slr digital camera" for User A and for User B. The rightmost section shows the best possible ranking if the same list must be returned to User A and User B. The normalized DCG for the best ranking when only one person is taken into account is 1. When more than one person must be accounted for, the normalized DCG drops.

Best Ranking for User A		**Best Ranking for User B**		**Best Ranking for Group (A + B)**			
	Gain		Gain		Gain		
Web Result	A	Web Result	B	Web Result	A	B	A+B
usa.canon.com/consu…	1	..wikipedia.org/DSLR	2	..wikipedia.org/DSLR	1	2	3
..about.com/professio…	1	..about.com/professio…	1	..about.com/professio…	1	1	2
..wikipedia.org/Digital…	1	..about.com/..reviews…	1	usa.canon.com/consu…	1	0	1
..wikipedia.org/DSLR	1	usa.canon.com/consu…	0	..about.com/..reviews…	0	1	1
..about.com/..reviews…	0	amazon.com/..-Rebel-…	0	..wikipedia.org/Digital…	1	0	1
amazon.com/..-Rebel-…	0	amazon.com/Canon-4…	0	amazon.com/..-Rebel-…	0	0	0
amazon.com/Canon-4…	0	..wikipedia.org/Digital…	0	amazon.com/Canon-4…	0	0	0
olympusamerica.com/…	0	olympusamerica.com/…	0	olympusamerica.com/…	0	0	0
olympusamerica..body…	0	olympusamerica..body…	0	olympusamerica..body…	0	0	0
astore.amazon.com/p…	0	astore.amazon.com/p…	0	astore.amazon.com/p…	0	0	0
	A		**B**		**A**	**B**	**Avg**
Normalized DCG	1.00	**Normalized DCG**	1.00	**Normalized DCG**	0.97	0.96	0.97

When there are more than one set of ratings for a result list, the ranking that maximized DCG ranks the results that have the highest collective gain across raters first. For queries with explicit judgments, this means that results that all raters thought were *highly relevant* are ranked first, followed by those that most people thought were *highly relevant* but a few people thought were just *relevant*, followed by results most people thought were *relevant*, etc. The collective gain for Users A and B for the query "slr digital camera" is shown in the rightmost column of Table 4, and the results ranked according to that gain can be seen in the rightmost column of Table 5. Because the list must satisfy more than one person, it is no longer the best list for either individual. Instead, as shown in Table 5, the normalized DCG for this best group ranking is 0.97 for User A and 0.96 for User B, for an average of 0.97. This group normalized DCG (0.97) is lower than the normalized DCG for the best individual rankings (1.00 for both A and B).

For those queries where we have measures of relevance from multiple people, it is possible to find the best possible result ranking for each individual, as well as the best possible ranking for different sized groups of individuals. As additional people are added to the group, the gap between user satisfaction with the individualized rankings and the group ranking grows. The gap between the optimal rating for an individual and the optimal rating for the group is what we call the *potential for personalization*.

5.2 Potential for Personalization Using Explicit Measures

We explored the potential for personalization using the explicit measures of relevance we collected. The graph depicted in Figure 4 shows the average normalized DCG for the best individual (dotted line), group (solid line), or current Web rankings (dashed line) as a function of the number of individuals in the group. These data were derived from the 17 pre-generated queries for which more than five individuals made explicit relevance evaluations of the results (see Table 2).

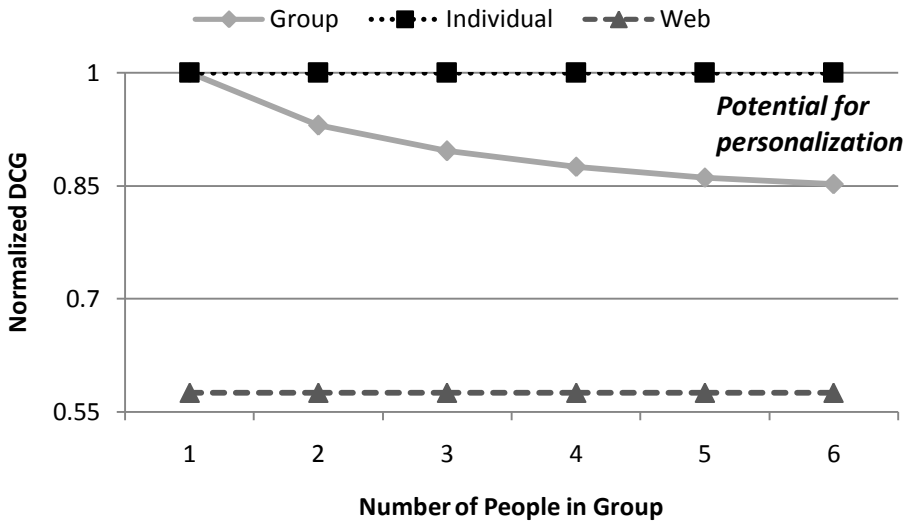


Figure 4. With perfect personalization, the average normalized DCG for an individual is 1. As more people's interests are taken into account to generate a ranking, the average normalized DCG for each individual drops for the ideal group ranking. The gap represents the potential value to be gained by personalizing the search results. There is also a gap between the current normalized DCG for the Web results and the best group ranking, which represents the potential improvement to be gained merely by improving results without consideration of the individual.

Using the explicit data we could only explore small groups. Search engines that do not tailor the search experience to individual users must try to find the best possible result ranking for the much larger "group" of people consisting of all possible searchers. In our data analysis, it is impossible for us to explore the preferences of everyone, as we cannot feasibly collect relevance judgments (even implicitly) from everyone. Instead we use the potential for personalization curves to make a good guess about what the potential would be among large groups by looking at how it increases as group size increases.

On average, when considering groups of six people, the best group ranking based on explicit data yielded a 46% improvement in DCG over the current Web ranking (0.85 vs. 0.58), while the best

individual ranking led to a 70% improvement (1.00 vs. 0.58). From the shape of the curves it appears likely that the best group ranking for a larger sample of users would result in even lower DCG values and be closer to the Web ranking which aims to satisfy a large number of searchers interests for the same query.

These analyses of people's explicit relevance judgments underscore the promise of providing users with better search result quality by personalizing results. Improving core search algorithms is difficult, with research leading typically to very small improvements. We have learned that rather than improving the overall ranking for a particular query, we can obtain significant boosts by working to improve results to match the intentions behind it – and that these intentions can be different for different individuals.

5.3 Potential for Personalization Using the Implicit Measures

For both of the implicit measures studied (content and behavior), we constructed, for groups of different sizes, the best group ranking that we could using the measure. We then measured the quality of each group ranking using the implicit gains to assess the normalized DCG. This allowed us to create potential for personalization curves for both implicit measures that are similar to the explicit one displayed in Figure 4. The distance of these implicit potential for personalization curves from what is ideal for the individual (a normalized DCG of 1) gives us an idea of how much room there is to improve search results using personalization, versus improvement to the general results.

Figure 5 shows the same potential for personalization curve computed for the explicit relevance judgments in Figure 4 (solid line) for the behavior-based (dotted line) and content-based (dashed line) relevance proxies. The curves have a similar shape for all three measures of an individual user's intent. The potential for personalization based on behavior-based data is smaller than the actual variation observed in explicit relevance judgments. This is most likely due to the fact that despite variation in

intent, people's click behavior is strongly influenced by where the results appear in the ranked list [Guan and Cutrell 2007; Joachim et al. 2005].

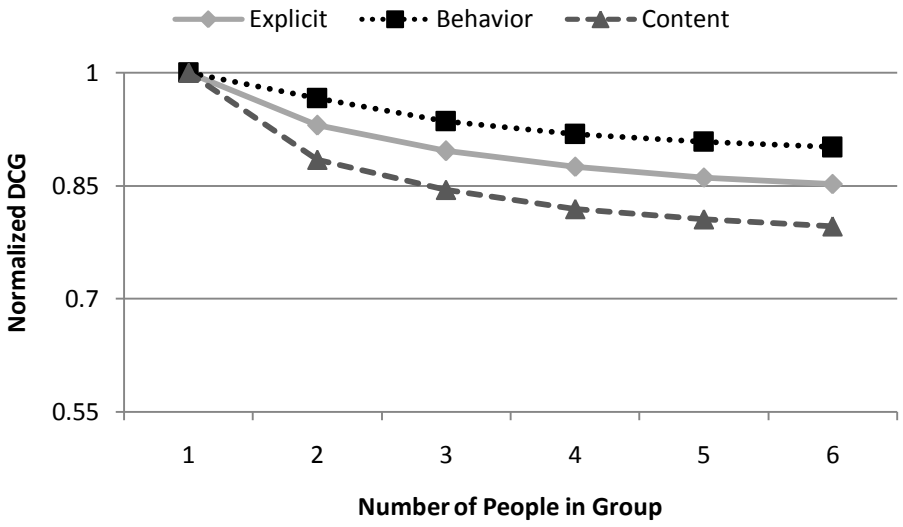


Figure 5. The potential for personalization curves according to the three different measures of relevance. Explicit relevance judgments for the 17 unique queries that more than 5 people evaluated are compared with 24 queries for which there are at least six content-based implicit judgments and the 44,002 behavior-based queries for which there are behavior-based implicit judgments.

In contrast, the content-based curve displays greater variation than the curve built from the explicit judgments. This suggests there is more variation in the content that has been previously viewed by an individual than there is variation in relevance judgments. It may be possible to leverage this variation to present the most personally relevant results to an individual.

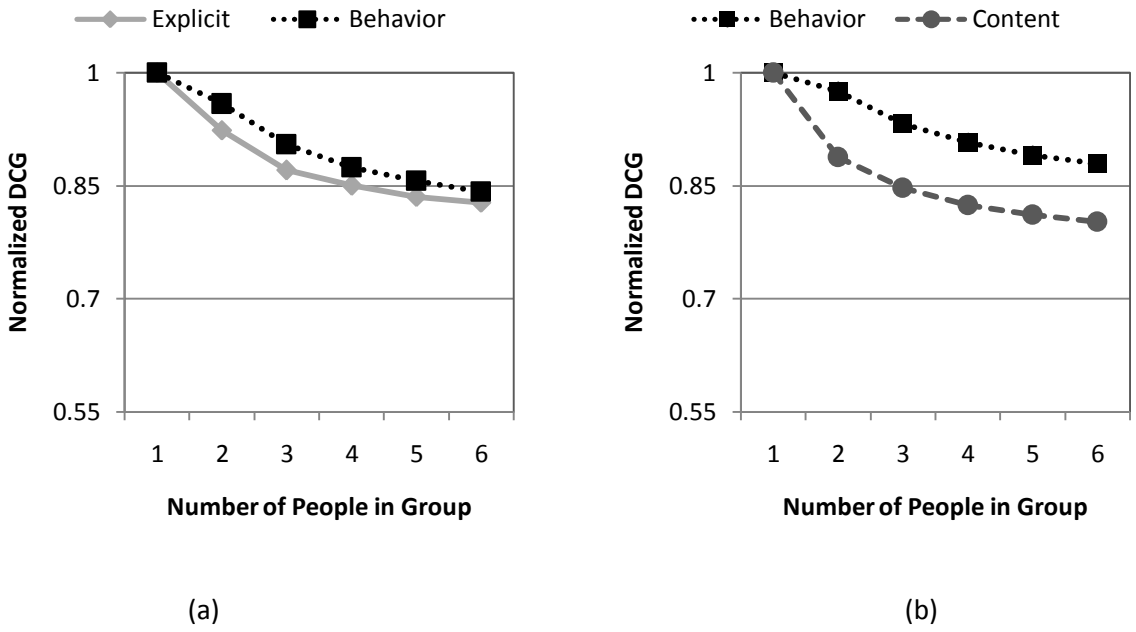


Figure 6. Behavior-based explicit potential for personalization curves for (a) the three overlapping queries where more than 5 people evaluated and (b) for the 14 overlapping content-based queries. The exact values of the curves are different from what was seen in Figure 5 because individual queries vary, but the general patterns remain.

Some of the variation we observe across measures may arise from the fact that the set of queries used for each measure varies somewhat, and it is possible that different queries have greater potential for personalization than others. However, a handful queries overlap across measures that we can examine to get an idea about whether the same pattern exists when there is no additional variation due to different queries. For three queries ("microsoft", "cancer", and "gates") we have both explicit and behavior-based implicit relevance judgments. As can be seen in Figure 6(a), the same pattern observed in Figure 5 holds, where the behavior-based measure suggests less potential for personalization than the explicit based measure. For 14 queries ("animal control", "aurora", "bandwidth test", "bank of america", "bespelled", "best buy", "canada", "circuit city", "hoffman", "union station", "walmart", "weather", "world map", and "yahoo") we have both behavior-based and content-based implicit

relevance judgments for at least six people. Figure 6(b) shows that the same relationship between the two implicit measures observed in Figure 5 also hold when compared across identical sets of queries.

Because we were able to collect content- and behavior-based data for the same queries from larger groups of people than we were with explicit data, it is possible to extend the curves to examine the potential for personalization when we consider larger numbers of individuals. Figure 7 shows the potential for personalization for groups of up to 23. Both the content-based curve and the behavior-based curves appear to continue to flatten out as more people's interests are considered. Where the curves asymptote represents the best possible results a non-personalized Web search engine can hope to achieve based on those implicit measures.

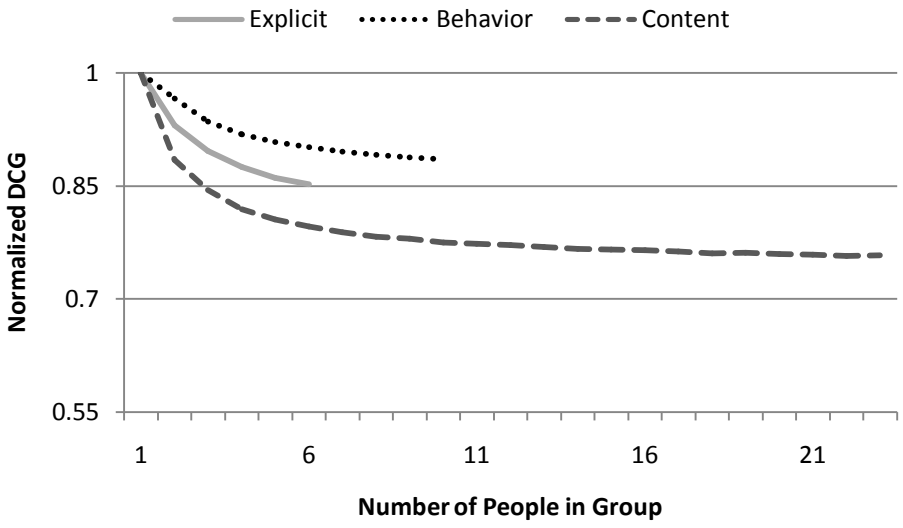


Figure 7. The potential for personalization curves for various relevance measures extended past groups of six. Again, explicit relevance judgments are compared with content-based and behavior-based implicit judgments.

Taken along with our analysis of rank and rating, Figures 5, 6, and 7 suggest that the information contained in behavior-based measures is different from what is contained in content-based measures. Related literature [Duo et al. 2007] suggests behavior-based measures are more accurate at identifying relevant results than content-based measures, but click measures also have several drawbacks. For one, as seen in Figure 3, they are unlikely to identify relevant results ranked low in the result list since they do not provide very much information about those results. Additionally, they do not appear to do a good job of identifying variation between individuals. Because there is less variation in the metric than there is in explicit judgments, it would be impossible for the metric to capture all of the variation in judgments. In contrast, content-based measures, while perhaps less accurate in indicating relevance, seem more likely to successfully achieve these goals because they can identify low ranked results and they do contain a sufficient amount of variation.

In this section, we have used different explicit and implicit measures to characterize the benefit that could be obtained if we could personalize optimally for each individual. In practice, of course, there will be many challenges in modeling users' interests and in developing an end-to-end search system that uses these models to present results differently for different individuals. In the next section, we describe several such systems and look closely one example which we have developed to personalize Web search results.

6. Personalizing Search Using Content and Behavior

The data and analysis methods described above show how we can characterize the variability in what different users find relevant to the same query using potential for personalization curves. Further, we have shown how two complimentary implicit measures (mined from user behavior and content) are related to explicit relevance judgments. There are many ways in which such analyses and understanding can be applied to improve the user's search experience. For example, search results could be re-ranked

using information about a user's interests and activities. Or, we could analyze the potential for personalization for different queries, and provide addition support for helping users articulate their needs or understand the results for queries that have a large variation in what different users find relevant (e.g., by soliciting greater elaboration about a user's intent, providing query suggestions, or grouping search results for queries). Below we describe in more detail a case study of how we used the ideas developed in this paper to personalize search results. This is not intended to be a detailed discussion of the system or its evaluation, but rather to illustrate how the ideas and methods might be used in practice.

As a specific example of a way to use the analyses described above, we built a system to personalize the search results an individual receives using implicit behavior- and content-based indicators of what that individual might consider relevant. The algorithm used in the personalized search system is described in greater detail elsewhere [Teevan et al. 2005]. In this paper we highlight how insights derived from the analyses presented in this paper help us to understand how the system functions and we show how the system performs over the new, larger data set of explicit judgments we collected and from deployment. We first show how implicit content-based measures of relevance can be used to personalize the results, and then show how implicit behavior-based measures of relevance can be used. We finish with a discussion of how well the different measures perform.

6.1 Content-Based Personalization

We implemented content-based Web search personalization by modifying BM25 [Sparck Jones et al. 1998], a well-known text-based probabilistic weighting scheme for information retrieval. As described in Section 3.3, BM25 assigns weights to individual terms in a document based on their frequency of occurrence in the document and the corpus, and uses these weights to estimate the probability that the document is relevant to a query. When relevance information is available, term weights can be

modified by giving additional weight to terms that discriminate relevant documents from irrelevant documents, a technique known as *relevance feedback*. Relevance information is typically obtained by explicitly asking users which documents are relevant; this kind of feedback can be considered a very simple and short-term content-based user profile, based on documents the user has selected as relevant to the particular query. For our content-based personalized search system, instead of relying on explicit relevance judgments, we incorporate implicit long-term content-based feedback using the content-based profile described in Section 3.3. This implicit feedback is used to modify term weights in the same manner as relevance feedback operates by computing the log odds of a term appearing in relevant and non-relevant documents (see [Teevan et al. 2005] for details).

In previous work we explored many different approaches to implementing the specifics of this content-based personalization algorithm. In this paper, we look only at the most successful approach. Roughly speaking, this approach re-scores Web search results by giving more weight to terms that occur relatively more frequently in the user's model than in the Web search results. As we did earlier when exploring content-based implicit measures of relevance, we use only the title and snippet of the document returned by the Web search engine to represent the document. Global statistics for each term are approximated using the title and snippets in the result set. This allows the personalization to be computed client-side without access to an index of the entire Web. 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.

6.2 Behavior-Based Personalization

In addition to content-based analysis of the user's personal index, we also considered a behavior-based representation of the user's interests. We used a simple behavior-based ranking method that boosts previously viewed results to the top of the result list, followed by results that were from domains

the individual tends to frequent. Results associated with URLs where the last three components of the URL's domain name (*e.g.*, http://***tochi.acm.org***) matched a previously visited URL were boosted more than those results where the last two components matched (*e.g.*, http://tochi ***acm.org***).

For each Web search result we computed the behavior-based similarity score based on a user's previous history of interaction using the simple domain matching algorithm described above.

6.3 Personalization Performance

In this section we look at how well content- and behavior-based personalization performs in practice. We do so by examining the quality of the rankings produced by each of the two implicit measures alone compared with simple text-based baselines, and when compared with the Web. We also present the results of a longitudinal user study we conducted to explore how personalization affects people's real world interactions with search results.

6.3.1 Quality of Ranking Produced by Personalized Search

In earlier work [Teevan et al. 2005] we analyzed the performance of the content- and behavior-based personalization algorithms on the 131 queries listed in Table 2, LI. Here we present the performance based on the complete set of 699 queries for which we now have explicit relevance judgments. Note that in this analysis we use all of the queries for which we gathered explicit judgments, and not just those 21 unique queries for which multiple people evaluated the results. In all cases when explicit relevance judgments were gathered, we also collected the necessary content- and behavior-based features to perform personalization so that we could explore developing personalization algorithms to match the judgments.

Statistical analyses are performed using two-tailed paired *t*-tests with 698 degrees of freedom.

Comparing Personalization to Text-Based Baselines

We begin by comparing the content- and behavior-based personalization methods with a simple text baseline in which no personalization is done. To calculate the baseline we ranked each result using the BM25 scoring formula with no user profile. Results for this ranking, which uses no user model (*No*), are summarized using the normalized DCG measure and presented in the left-most bar in Figure 8. Performance for the two personalized search algorithms are shown in the bars labeled *Cont* (for "content") and *Beh* (for "behavior"). Not surprisingly, search with no user model (*No*) was significantly outperformed by both content-based (*Cont*; $t(698)=9.32$, $p<0.01$) and behavior-based (*Beh*; $t(698)=15.71$, $p<0.01$) personalization.

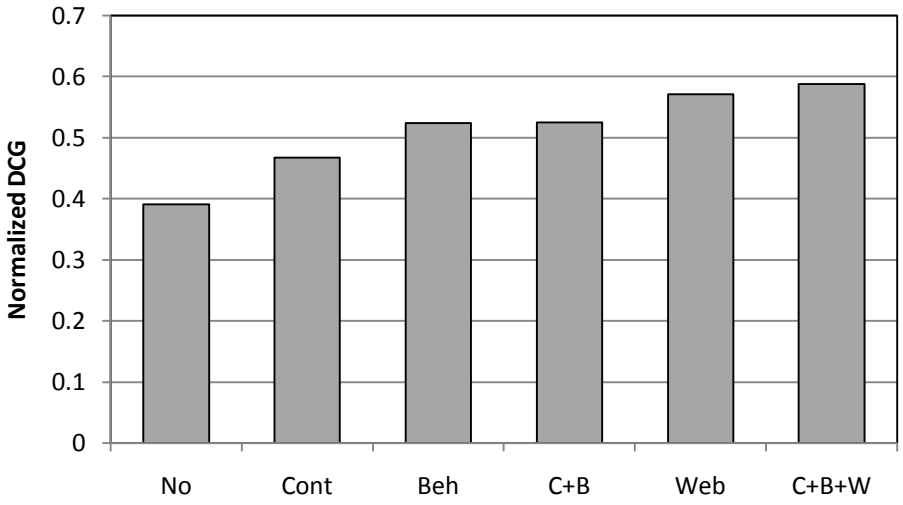


Figure 8. Content-based personalized search (*Cont*) and behavior-based personalization (*Beh*) compared with no user model (*No*), the Web (*Web*), and personalization combined with the Web (*C+B+W*).

Figure 8 also shows a comparison of the personalized content- and behavior-based rankings with the Web ranking (*Web*). While Web search engines do not rank documents in a personalized manner, they

do take advantage of a large amount of information about the documents in their indices beyond the document's textual content, such as linkage information and behavior-based. Using these rich and varied data sources has been shown to improve results for many search tasks [Hawking and Craswell 2001]. Thus it is not surprising that although content- and behavior-based personalization both performed better than a purely text-based algorithm, they performed worse than the Web ranking. The Web rank had a normalized DCG of 0.57, compared with 0.47 for content-based personalization ($t(698)=11.22$, $p<0.01$) and 0.53 for behavior-based personalization ($t(698)=4.99$, $p<0.01$).

Merging Information from Different Sources

Given the Web ranking uses rich data to capture the common elements of what people consider relevant better than text alone can, and we have seen that the two implicit measures capture different aspects valuable for personalization, we suspected it would be possible to combine the three information sources to produce a ranking that captured the best aspects of each.

To take advantage of both of behavior- and content-based implicit features, we linearly combined the personalized scores derived from each measure and ranked results by the combined score. The quality of the personalized list for the best linear combination (found via leave-one-out cross validation) is labeled *C+B* in Figure 8. Combining the two implicit features yielded a small but insignificant improvement over Behavior (*Beh*) alone, and a significant improvement over content (*Cont*; $t(698)=9.97$, $p>0.01$). The solid line in Figure 9 shows how weighting the behavior- and content-based features differently when the scores are combined affected the quality of the resulting personalized ranking. Performance smoothly varied from models in which the behavior-based feature is not weighted at all (merge weighting is 0) through it being the only feature (merge weighting is 1). The curve peaks at a merge weighting between 0.7 and 0.9.

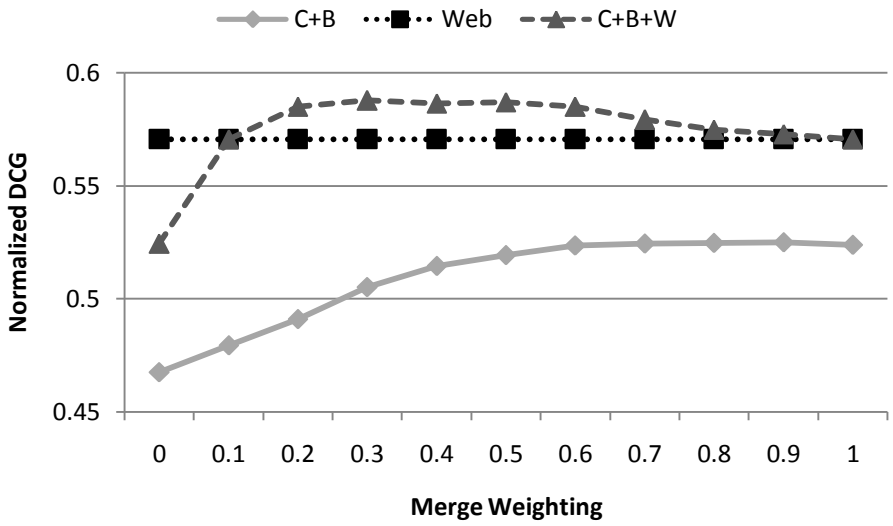


Figure 9. Normalized DCG for the combined personalized search algorithm as a function of how different sources of information used in ranking are weighted. The solid line shows how content-based features and behavior-based features trade off (a weighting of 0 means only content-based features are considered, and a weighting of 1 means only behavior based features are considered). When the best combination of implicit features is then biased by the Web ranking, the performance improves significantly over the Web ranking. A weighting of 0 for the dashed line means the Web ranking is not accounted for at all in conjunction with the best combination of implicit indicators, while a weighting of 1 means only the Web ranking is accounted for.

Note that at this peak the normalized DCG is 0.52, which is still significantly below the flat dotted line representing the Web ranking ($t(698)=4.92$, $p<0.01$). It is still necessary to capture the information present in the Web ranking that is not present in either content- or behavior-based features. To do this we linearly combined a score based on the Web rank with the best personalized score (created from both content- and behavior-based features) to produce a meta-score by which to rank results. To compute a score for a document from the Web rank we used the inverse of the log of the rank. This is consistent with the weighting put on different rankings by DCG and with the general relevance curves seen earlier in Section 0. Scoring Web results in this way has the effect in practice of keeping the first

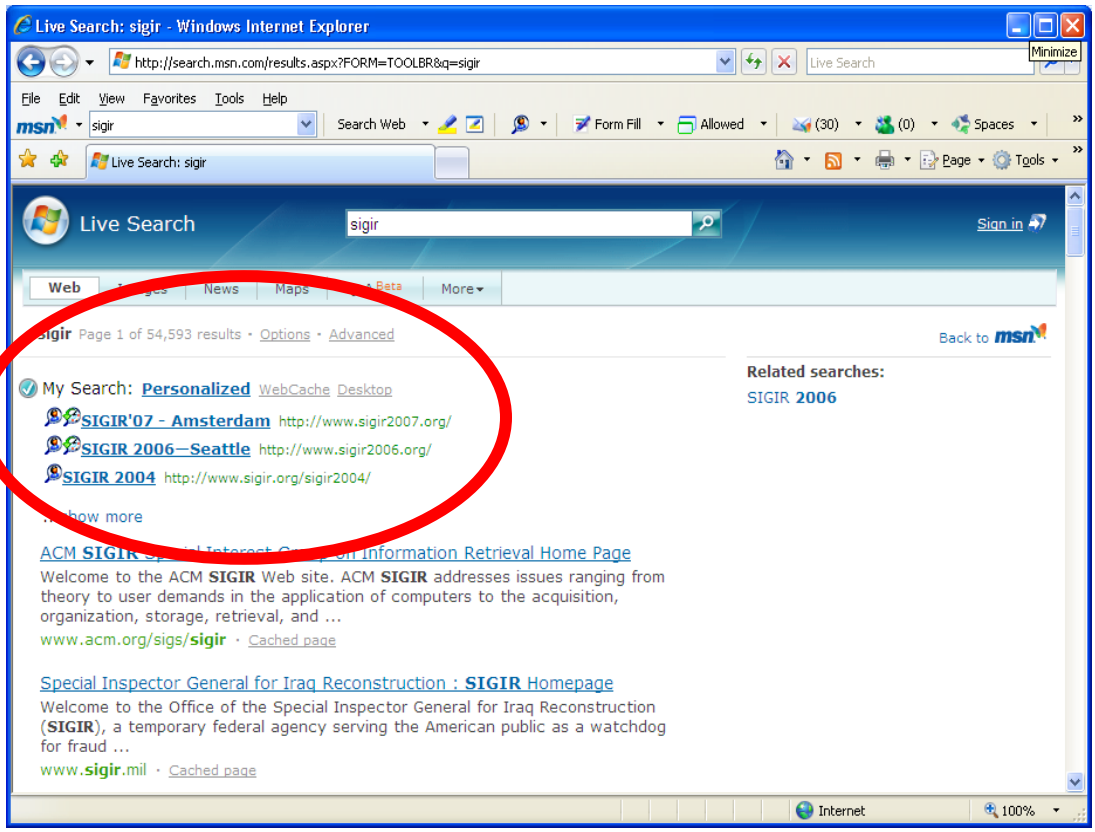
few results similar to the Web, while more heavily personalizing the results further down the list. The combination of the Web ranking and the personalized ranking (*B+C+W* in Figure 8) yielded an average normalized DCG of 0.59, a significant ($t(698)=3.12$, $p<0.01$) improvement over the Web's average normalized DCG of 0.57. As can be seen in Figure 9, the advantages were observed for a wide range of mixing parameter values.

6.3.2 User Study of Personalized Search

Although we saw an improvement in the quality of personalized results using our relatively small test set of explicit judgments, we wanted to know more about how personalization using content- and behavior-based features affects the user's search experience in realistic search settings. To explore this we deployed a Web browser plug-in that provides personalized search capabilities. For each query, the top 100 Web search results are retrieved (title, snippet and URL) and re-ranked using the algorithms described above. The prototype was used by 170 Microsoft employees for five weeks, and data was collected about their use patterns, including the queries issued and the results clicked. We observed 10,312 queries and 12,295 URL clicks during this observation period.

In the prototype, we display the title and URLs of the three best-matching personalized search results in a separate personalized region above the regular Web results, as shown in Figure 10. Icons are used to indicate that the results have been personalized and (when appropriate) previously visited. In the example in Figure 10, both personalized and Web results are shown for the query *SIGIR*. The personalized results (based on the content model and behavior of one of the authors) are those related to the conference for the ACM Special Interest Group on Information Retrieval (SIGIR), rather than results about the Special Inspector for Iraq Reconstruction, which are usually returned by Live Search.

Figure 10 Screenshot of the personalized search prototype. Personalized results are shown above the general Web search results in a region labeled My Search. The title and URLs of the three best-matching personalize results are shown. Icons are used to indicate whether the results are personalized and previously revisited.



It is difficult to directly compare participants' interactions with the standard Web results and the personalized results because the personalized results were presented in a separate personalized region, did not contain any text beyond the title and URL to describe them, and sometimes took longer to display than the Web results. For this reason, to obtain an appropriate baseline we displayed the top three Web search results in the personalized region for 5% of the queries. There were additional queries (31.1% of all recorded) where no personalization was possible because the queries did not match the user profile sufficiently , and in these instances the top three Web results were again shown

in the personalized region. The interaction of our participants with the Web results displayed in the personalized region provides a baseline for evaluating the quality of the personalized results. In practice, one would probably not show the personalized region when the results were the same as Web results or when there was insufficient evidence for personalization, but for experimental purposes we chose to show results in this region for all queries.

Table 6. Results from a 5 week user study of personalized search. For the 36.1% of the queries where the results displayed in the personalized area were the same as the Web results, users clicked on in the area 4.3% of the time. However, for the queries where the results in the personalized area were personalized, participants clicked in the area on average 5.5% of the time. The value of the personalized results went up as the number of documents in the user profile matched by the query increased.

		Personalized result clicks	% of total queries issued
Web results		4.3%	36.1%
Personalized		5.5%	63.9%
Items matched	1-5	4.2%	22.4%
	6-10	5.2%	8.5%
	11-50	6.0%	17.2%
	51-100	5.6%	5.5%
	100+	7.5%	10.3%

Table 6 contains a summary of how often participants clicked on a result in the personalized area under differing conditions. For the queries where the results shown in the personalized area were the same as the Web results, users clicked on results in the area 4.3% of the time. However, for the queries where the results in the personalized area were personalized, participants clicked on results in the area

on average 5.5% of the time. It appears personalization positively affected their search behavior to produce greater click-through. We particularly observed more clicks in the personalized area when the query topic was well represented in the user's profile. Table 6 also shows that as the number of items in the user profile that matched the query increased, clicks in the personalized area also increased. It appears that the more information a participant has in their profile related to a query, the more likely personalization is to be beneficial. It also appears that personalization was particularly useful when what the searcher was looking for was not already ranked highly by the Web search engine. participants clicked in the personalized area 26% of the time when the result they found for the query was below-the-fold in the Web results.

In this section we have described a system which takes a first step in personalizing Web search results by taking advantage of the potential for personalization using content- and behavior-based measures. We have found that the best system combines the current Web ranking (which is based on many variables) with implicit content-based and behavior-based, and that this system positively impacts user behavior.

7. Conclusion

In this paper we have explored the variation in what different people consider relevant to the same query both empirically and theoretically. We mined three complimentary sources of evidence – *explicit* relevance judgments, a *behavior-based implicit* measure of relevance, and a *content-based implicit* measure of relevance – to measure variability in judgments and behavior for the same query. We found large variations in all three measures among different individuals. As a result, there is a large gap between how well search engines could perform if they were to tailor results to individuals, and how well they currently perform by returning a single ranked list of results designed to satisfy everyone. We called this gap the *potential for personalization,* and showed how it can be quantified using measures

such as discounted cumulative gain. We discussed several uses of such analyses including a personalized search system that uses these insights to improve the search experience by ranking personally relevant results more highly.

Looking forward, we plan to examine additional implicit measures of relevance, such as the link-based measures proposed by Jeh and Widom [Jeh and Widom 2003], and richer behavioral models, including dwell-time and interactions patterns (e.g., Fox et al. 2005; Joachims et al. 2005). Richer application programming interfaces (APIs) to Web search engines could also provide richer information about the Web ranking, further improving our ability to incorporate implicit user-specific features with more global features contained in Web search engine result rankings.

We are also exploring differences in the *potential for personalization* across different queries, users and implicit measures. We noticed variation across queries in the potential for personalizing the results to the query. For some queries there was a large gap between the individual and group curves, and for other queries the difference was much smaller, and we have begun to examine this in greater detail [Teevan et al. 2008]. We are also looking in greater depth at the variation between individuals. It may be that some people always want things that are different from the norm, while others want the same. If there are common approaches that sets of people appear to take, or if people's interests fall into groups, perhaps we may be able to back off smoothly in our personalization algorithms, personalizing first to the individual, then to the group, and then accounting for general interests among all people. Such an approach could capture both the full potential for personalization and the full value to be gained from using rich collective group information. Preliminary work in this area has been published by Morris and Teevan [Morris and Teevan 2008; Morris et al 2008].

Finally, we will continue to develop and evaluate new presentation techniques for displaying personalized search results, and to study how personalization impacts users' search experiences.

Acknowledgments

The authors would like to thank Merrie Morris and Steve Bush for their assistance in collecting the large amount of additional explicit relevance judgments reported here, and Dan Liebling for his assistance in collecting the implicit behavior-based information.

References

[Agchtein et al. 2006] Agichtein, E., Brill, E., Dumais, S., and Ragno, R. (2006). Learning user interaction models for predicting Web search preferences. In Proceedings of SIGIR '06, 3-10.

[Anick, P. 2003] Anick (2003). Using terminological feedback for web search refinement: A log based study. In In Proceedings of SIGIR '03, 88-95.

[Claypool et al. 2001] Claypool, M., Brown, D., Le, P., and Waseda, M. (2001). Inferring user interest. IEEE Internet Computing, Nov/Dec, 32-39.

[Chi and Pirolli 2006] Chi, E. and Pirolli, P. (2006). Social information foraging and collaborative search. 2006 Human Computer Interaction Consortium. Available at http://www2.parc.com/istl/projects/uir/publications/author/Pirolli_ab.html

[Chirita et al. 2006] Chirita, P., Firan, C., and Nejdl, W. (2006). Summarizing local context to personalize global Web search. In Proceedings of SIGIR '06, 287-296.

[Dou et al. 2007] 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.

[Dumais et al. 2003] 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.

[Eastman and Jansen 2003] Eastman, C. M. and Jansen, B. J. (2003). Coverage, relevance and ranking: The impact of query operators on Web search engine results. ACM Transaction of Information System, 21(4), 383-411.

[Fidel and Crandall 1997] Fidel, R. and Crandall, M. (1997). Users' perception of the performance of a filtering system. In Proceedings of SIGIR '97, 198-205.

[Frias-Martinez et al. 2007] Frias-Martinez, E., Chen, S. Y., and Liu, X. (2007). Automatic cognitive style identification of digital library users for personalization. Journal of the American Society for Information Science and Technology, 58(2), 237-251.

[Fox et al. 2005] Fox, S., Karnawat, K., Mydland, M., Dumais, S. T., and White, T. (2005). Evaluating implicit measures to improve Web search. ACM Transaction of Information System, 23(2), 147-168.

[Guan and Cutrell 2007] Guan, Z. and Cutrell, E. (2007). An eye-tracking study of the effect of target rank on Web search. In Proceedings of CHI '07, 417-420.

[Harter 1996] Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. Journal of the American Society for Information Science, 47(1), 37-49.

[Hawking 2000] Hawking, D. (2000). Overview of the TREC-9 Web Track. In Proceedings of TREC '00, 87-102.

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

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

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

[Joachims et al. 2005] Joachims, T., Granka, L., Pang, B., Hembrooke, H., and Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In Proceedings of SIGIR '05, 154-161.

[Kelly and Fu 2007] Kelly, D. and Fu, X. (2007). Eliciting better information need descriptions from users of information systems. Information Processing & Management, 43(1), 30-46.

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

[Koenmann and Belkin 1996] 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.

[Ma et al. 2007] Ma, Z., Pant, G., and Sheng, O. (2007). Interest-based personalized search. ACM Transactions on Information Systems, 25(5), Article 5.

[Mizzaro 1997] Mizzaro, S. (1997). Relevance: The whole history. Journal of the American Society of Information Science and Technology, 48(9), 810-832.

[Morita and Shinoda 1994] 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.

[Morris and Teevan 2008] Morris, M. R. and Teevan, J. (2008). Understanding groups' properties as a means of improving collaborative search systems. In Proceedings of the JCDL '08 Workshop on Collaborative Information Retrieval.

[Morris et al. 2008] Morris, M. R., Teevan, J., and Bush, S. (2008). Enhancing collaborative Web search with personalization: Groupization, smart splitting, and group hit-highlighting. In Proceedings of CSCW '08.

[Pitkow et al. 2002] 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.

[Radlinski and Dumais 2006] Radlinski, F. and Dumais. S. (2006). Improving personalized Web search using result diversification, In Proceedings of SIGIR '06, 691-692.

[Russell and Grimes 2007] Russell, D. and Grimes, C. (2007). Assigned and self-chosen tasks are not the same in Web search. In Proceedings of HICSS '07.

[Ruthven, I. 2003] Ruthven (2003). Re-examining the potential effectiveness of interactive query expansion. In In Proceedings of SIGIR '03, 213-220.

[Saracevic 1976] Saracevic, T. (1976). Relevance: A review of the literature and a framework for thinking on the notion in information science. Advances in Librarianship, 6, 81-139.

[Saracevic 2006] Saracevic, T. (2006). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II. Advances in Librarianship, 30, 3-71.

[Schamber 1994] Schamber, L. (1994). Relevance and information behavior. Annual Review of Information Science and Technology, 29, 3-48.

[Shen et al., 2005] Shen, X., Tan, B. and Zhai, C. X. (2005). Implicit user modeling for personalized search. In Proceedings of CIKM ’05, 824-831.

[Sparck Jones et al. 1998] 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.

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

[Sugiyama et al. 2004] 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.

[Teevan et al. 2004] 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.

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

[Teevan et al. 2007a] Teevan, J., Dumais, S. T., and Horvitz, E. (2007). Characterizing the value of personalizing search. In Proceedings of SIGIR ’07, 757-756.

[Teevan et al. 2007b] Teevan, J., Adar, E., Jones, R., and Potts. M. (2007). Information re-retrieval: Repeat queries in Yahoo’s logs. In Proceedings of SIGIR ’07, 151-158.

[Teevan et al. 2008] Teevan, J., Dumais, S. T., and Liebling, D. J. (2008). Personalize or not to personalize: Modeling queries with variation in user intent. In Proceedings of SIGIR ’08, 163-170.

[Voorhees 1998] Voorhees, E. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In Proceedings of SIGIR'98, 315–323.

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

[Wu et al. 2008] Wu, M., Turpin, A., and Zobel, J. (2008). An investigation on a community's web search variability. In Proceedings of the Australian Computer Society Confernece (ACSC 2008).