Demonstration of Improved Search Result Relevancy Using Real-Time Implicit Relevance Feedback

David Hardtke
Surf Canyon Incorporated
274 14th St.
Oakland, CA 94612
hardtke@surfcanyon.com

Mike Wertheim
Surf Canyon Incorporated
274 14th St.
Oakland, CA 94612
mikew@surfcanyon.com

Mark Cramer
Surf Canyon Incorporated
274 14th St.
Oakland, CA 94612
mcramer@surfcanyon.com

ABSTRACT

Surf Canyon has developed real-time implicit personalization technology for web search and implemented the technology in a browser extension that can dynamically modify search engine results pages (Google, Yahoo!, and Live Search). A combination of explicit (queries, reformulations) and implicit (clickthroughs, skips, page reads, etc.) user signals are used to construct a model of instantaneous user intent. This user intent model is combined with the initial search result rankings in order to present recommended search results to the user as well as to reorder subsequent search engine results pages after the initial page. This paper will use data from the first three months of Surf Canyon usage to show that a user intent model built from implicit user signals can dramatically improve the relevancy of search results.

Keywords

Implicit Relevance Feedback, Personalization, Adaptive Search System

1. INTRODUCTION

It has long since been demonstrated that *explicit* relevance feedback can improve both precision and recall in information retrieval[1]. An initial query is used to retrieve a set of documents. The user is then asked to manually rate a subset of the documents as relevant or not relevant. The terms appearing in the relevant document are then added to the initial query to produce a new query. Additionally, non-relevant documents can be used to remove or de-emphasize terms for the reformulated query. This process can be repeated iteratively, but it was found that after a few iterations very few new relevant documents are found [2].

Explicit relevance feedback as described above requires active user participation. An alternative method that does not require specific user participation is *pseudo* relevance feedback. In this scheme, the top N documents from the initial query are assumed to be relevant. The important terms in these documents are then used to expand the original query.

Implicit Relevance Feedback aims to improve the precision and recall of information retrieval by utilizing user actions to infer the relevance or non-relevance of documents. Many different user behavior signals can contribute to a probabilistic evaluation of document relevance. Explicit document relevance determinations are more accurate, but implicit relevance determinations are more easily obtained as they require no additional user effort.

SIGIR '09, July 19-23, 2009, Boston, USA.

2. IMPLICIT SIGNALS AND USER INFORMATION NEED

With the large, open nature of the World Wide Web it is very difficult to evaluate the quality of search engine algorithms using explicit human evaluators. Hence, there have been numerous investigations into using implicit user signals for evaluation and optimization of search engine quality. Several studies have investigated the extent to which a clickthrough on a specific search engine result can be interpreted as a user indication of document relevancy (for a review see [3]). The primary issue involving clickthrough data is that users are most likely to click on higher ranked documents because they tend to read the SERP (search engine results page) from top to bottom. Additionally, users trust that a search engine places the most relevant documents at the highest positions on the SERP.

Joachims *et al* used eye tracking studies combined with manual relevance judgements to investigate the accuracy of clickthrough data for implicit relevance feedback [4]. They conclude that clickthrough data can be used to accurately determine relative document relevancies. If, for instance, a user clicks on a search result after skipping other search results, subsequent evaluation by human judges show that in $\sim$80% of cases the clicked document is more relevant to the query than the documents that were skipped.

In addition to clickthroughs, other user behaviors can be related to document relevancy. Fox *et al.* used a browser add-in to track user behavior for a volunteer sample of office workers[5]. In addition to tracking their search and web usage, the browser add-in would prompt the user for specific relevance evaluations for pages they had visited. Using the observed user behavior and subsequent relevance evaluations, they were able to correlate implicit user signals with explicit user evaluations and determine what user signals are most likely to indicate document relevance. For pages clicked by the user, the user indicated that they were either satisfied or partially satisfied with the document nearly 70% of the time. In the study, two other variables were found to be most important for predicting user satisfaction with a result page visit. The first was the duration of time that the user spent away from the SERP before returning – if the user was away from the SERP for a short period of time they tended to be dissatisfied with the document. The other important variable for predicting user satisfaction was the "Exit type" – users that closed the browser on a result page tended to be satisfied with that result page. The important outcome of this and other studies is that implicit user behavior can be used instead of explicit user feedback to determine the user's information need.

3. IMPLICIT REAL-TIME PERSONALIZATION

As discussed in the previous section, it has been shown that implicit user behavior can often infer satisfaction with visited results pages. The goal of the Surf Canyon technology is to use implicit user behavior to predict which *unseen* documents in a collection are most relevant to the user and to recommend these documents to the user.

Shen, Tan, and Zhai[1] have investigated context-sensitive adaptive information retrieval systems [6]. They use both clickthrough information and query history information to update the retrieval and ranking algorithm. A TREC collection was used since manual relevancy judgements are available. They built an adaptive search interface to this collection, and had 3 volunteers conduct searches on 30 relatively difficult TREC topics. The users could query, re-query, examine document summaries, and examine documents. To quantify the retrieval algorithms, they used Mean Average Precision (MAP) or Precision at 20 documents. As these were difficult TREC topics, users submitted multiple queries for each topic. They found that including query history produced a marginal improvement in MAP, while use of clickthrough information produced dramatic increases (up to nearly 100%) in MAP.

Shen *et al.* also built an experimental adaptive search interface called UCAIR (User-Centered Adaptive Information Retrieval) [7]. Their client-side search agent has the capability of automatic query reformulation and active reranking of unseen search results based on a context driven user model. They evaluated their system by asking 6 graduate students to work on TREC topic distillation tasks. At the end of each topic, the volunteers were asked to manually evaluate the relevance of 30 top ranked search results displayed by the system. The top results shown are mixed between Google rankings and UCAIR rankings (some results overlap), and the evaluators could not distinguish the two. UCAIR rankings show a 20% increase in precision for the top 20 results.

The Surf Canyon browser extension represents the first attempt to integrate implicit relevance feedback directly into the major commercial search engines. Hence, we are able to evaluate this technology *outside* of controlled studies. From a research perspective, this is the first study to investigate this technology in the context of normal searches by normal users. The drawback is that we have no chance to collect *a posteori* relevancy judgements from the searchers or to conduct surveys to evaluate the user experience. We can, however, quickly collect large amounts of user data in order to evaluate the technology.

[1]Shen, Tan, and Zhai are co-authors on one Surf Canyon patent application but were not actively involved in the work presented here

4. TECHNOLOGICAL DETAILS

Surf Canyon's technology can be used as both a traditional web search engine and as a browser extension that dynamically modifies the search results page from commercial search engines (currently Google, Yahoo!, and Live Search). The underlying algorithms in the two cases are mostly identical. As the data presented was gathered using the browser extension, we will describe that here.

Surf Canyon's browser extension was publicly launched on February 19, 2008. From that point forward visitors to the Surf Canyon website[2] were invited to download a small piece of free software that is installed in their browser. The software works with both Internet Explorer and Firefox. Although the implementation differs for the two browsers, the functionality is identical.

Internet Explorer leads in all current studies of web browser market share with March 2008 market share estimated between 60% and 90%. Among users of the Surf Canyon browser extension, however, about 75% use Firefox. Among users who merely visit the extension download page, the breakdown by browser type is nearly 50/50. Part of the skew towards Firefox in both website visitors and users of the product can be attributed to the fact that marketing of the product has been mainly via technology blogs. Readers of technology blogs are more likely to use operating systems for which Internet Explorer is not available (e.g. Mac, Linux). Additionally, we speculate that Firefox may be more prevalent among readers of technology blogs. The difference between the fraction of visitors to the site using Firefox (∼50%) and the fraction of people who install and use the product using Firefox (∼75%) is likely due to the more widespread acceptance towards browser extensions in the Firefox community. The Firefox browser was specifically designed to have minimal core functionality augmented by browser addons submitted by the developer community. The technologies used to implement Internet Explorer browser extensions are also often used to distribute malware so there may be a higher level of distrust among IE users.

Once the browser extension is installed, the user never needs to visit the company web site again to use the product. The user enters a Google, Yahoo!, or Live Search web search query just as they would for any search (using either the search bar built into the browser or by navigating to the URL of the search engine). After the initial query, the search engine results page is returned exactly as it would be were Surf Canyon not installed (for most users who have not specified otherwise, the default number of search results is 10). Two minor modifications are made to the SERP. Small bull's eyes are placed next to the title hyperlink for each search result (see Figure 1). Also, the numbered links to subsequent search engine results pages at the bottom of the SERP are replaced by a single "More Results" link.

The client side browser extension is used to communicate with the central Surf Canyon servers and to dynamically update the search engine results page. The personalization algorithms currently reside on the Surf Canyon servers. This client-server architecture is used primarily to facilitate optimization of the algorithm and to support active research studies. Since web search patterns vary widely by user, the best way to evaluate personalized search algorithms is to vary the algorithms on the same set of users while main-

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

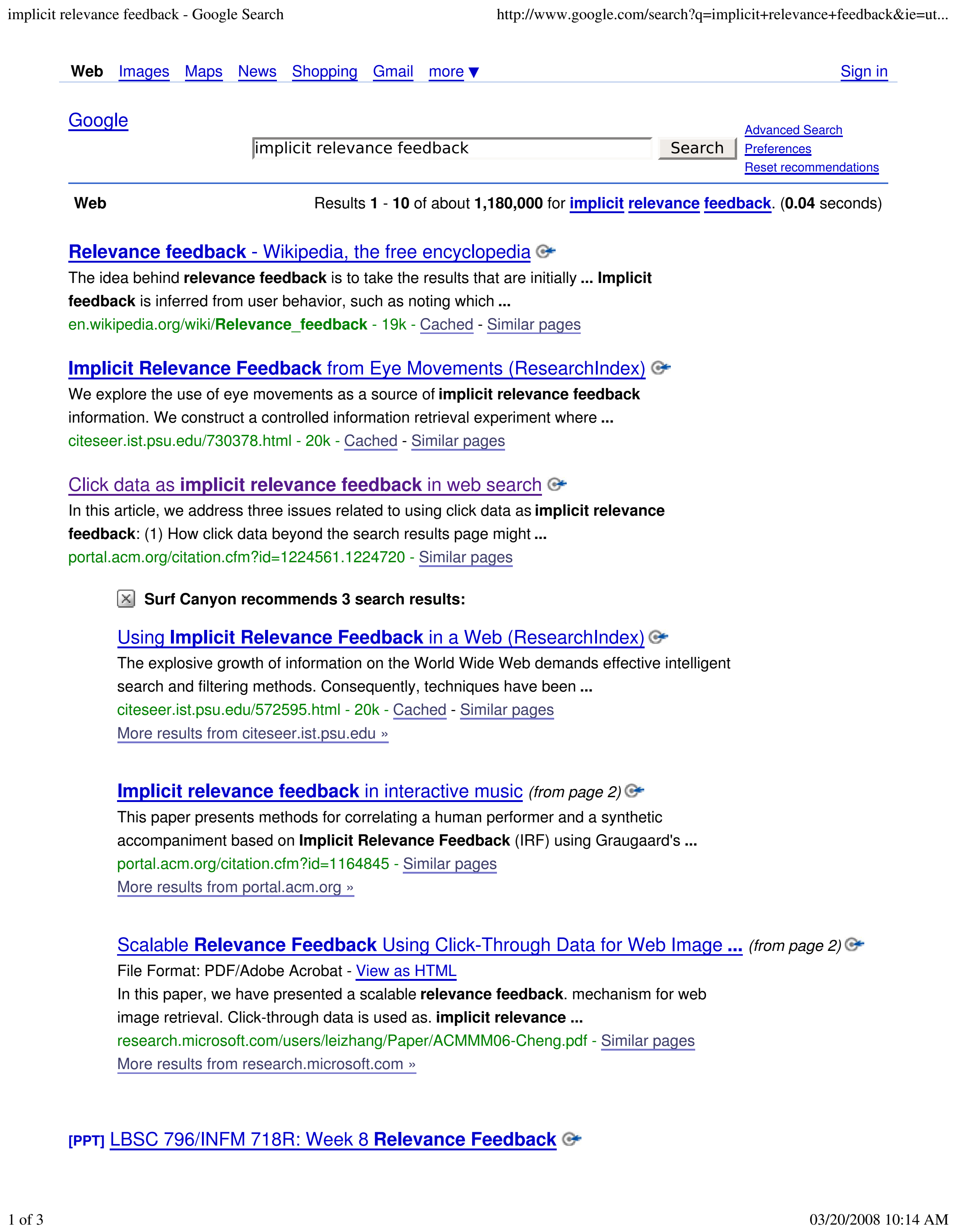

Figure 1: A screenshot of the Google search result page with Surf Canyon installed. The third link was selected by the user, leading to three recommended search results.

taining an identical user interface. With the client-server architecture, the implicit relevance feedback algorithms can be modified without alerting the user to any changes. Nothing fundamental prevents the technology from becoming exclusively client side.

In addition to the ten results displayed by the search engine to the user, a larger set of results (typically 200) for the same query is gathered by the server. With few exceptions, the top 10 links in the larger result set are identical to the results displayed by the search engine. While the user reads the search result page, the back-end servers parse the larger result set and prepare to respond to user actions. Each user action on the search result page is sent to the back-end server (note that we are only using the user's actions on the SERP for personalization and do not follow the user after they leave the SERP). For certain actions (select a link, select a Surf Canyon bull's eye, ask for more results) the back end server sends recommended search results to the browser. The Surf Canyon real-time implicit personalization algorithm incorporates both the initial rank of the result and personalized instantaneous relevancies. The implicit feedback signals used to calculate the real-time search result ranks are cumulative across all recent related queries by that user. The algorithm does not, however, utilize any long-term user profiling or collaborative filtering. The precise details of the Surf Canyon algorithm are proprietary and are not important for the evaluation of the technology presented below. If an undisplayed result from the larger set of results is deemed by Surf Canyon's algorithm to be more relevant than other results displayed below the last selected link, it is shown as an indented recommendation below the last selected link.

The resulting page is shown in Figure 1. Here, the user entered a query for "implicit relevance feedback" on Google[3]. Google returned 10 organic search results (only three of which are displayed in Figure 1) of the 1,180,000 documents in their web index that satisfy the query. The user then selected the third organic search result, a paper from an ACM conference entitled "Click data as implicit relevance feedback in web search". Based on the implicit user signals (which include interactions with this SERP, recent similar queries, and interactions with those results pages) the Surf Canyon algorithm recommends three search results. These links were initially given a higher initial rank (> 10) by the Google algorithm in response to the query "implicit relevance feedback". The real-time personalization algorithm has determined, however, that the three recommended links are more pertinent to this user's information need at this particular time than the results displayed by Google with initial ranks 4-10.

Recommendations are also generated when a user clicks on the small bull's eyes next to the link title. We assume that a selection of a bull's eye indicates that the linked document is similar to but not precisely what the user is looking for. For the analysis below, up to three recommendations are generated for each link selection or bull's eye selection. Unless the user specifically removes recommended search results by clicking on the bull's eye or by clicking the close box, they remain displayed on the page. Recommendations can nest up to three levels deep – if the user clicks on the first recommended result then up to three recommendations are generated immediately below this search result.

At the bottom of the 10 organic search results, there is a link to get "More Results". If the user requests the next page of results, all results shown on the second and subsequent pages are determined using Surf Canyon's instantaneous relevancy algorithm. Unlike the default search engine behavior, subsequent pages of results are added to the existing page. After selecting "More Results" links 1-20 are displayed in the browser, with link 11 focused at the top of the window (the user needs to scroll up to see links 1-10).

[3]http://www.google.com

5. ANALYSIS OF USER BEHAVIOR

Most previous studies of Interactive Information Retrieval systems have used post-search user surveys to evaluate the efficacy of the systems. These studies also tended to recruit test subjects and use closed collections and/or specific research topics. The data presented here was collected from an anonymous (but not necessarily representative) set of web surfers during the course of their interactions with the three leading search engines (Google, Yahoo, and Live Search). The majority of searches were conducted using Google. Where possible, we have analyzed the user data independently for each of the search engines and have not found any cases where the conclusions drawn from this study would differ depending on the user's choice of search engine. The total number of unique search queries analyzed was ∼700,000.

Since the users in this study were acquired primarily from technology web blogs, their search behavior can be expected to be significantly different than the average web surfer. Thus, we cannot evaluate the real-time personalization technology by comparing to previous studies of web user behavior. Also, since we have changed the appearance of the SERP and also dynamically modify the SERP, any metrics calculated from our data cannot be directly compared to historical data due to the different user interface.

Surf Canyon only shows recommendations after a bull's eye or search result is selected. It is therefore interesting to investigate how many actions a user makes for a given query as this tells us how frequently implicit personalization within the same query can be of benefit. Jansen and Spink [8] found from a meta-analysis of search engine log studies that user interaction with the search engine results pages is decreasing. In 1997, 71% of searchers viewed beyond the first page of search results. In 2002 only 27% of searchers looked past the first page of search results. There is a paucity of data on the number of web pages visited per search. Jansen and Spink [9] reported the mean number of web pages visited per query to be 2.5 for AllTheWeb searches in 2001, but they exclude queries where no pages were visited in this estimate. Analysis of the AOL query logs from 2006 [10] gives a mean number of web pages viewed per unique query of 0.97. For the current data sample, the mean number of search results visited is 0.56. The comparatively low number of search results that were selected in the current study has multiple partial explanations. The search results page now contains multiple additional links (news, videos) that are not counted in this study. Additionally, the information that the user is looking for is often on the SERP (e.g. a search for a restaurant often produces the map, phone number, and address). Search engines have replaced bookmarks and direct URL typing for re-visiting web sites. For such navigational searches the user will have either one or zero

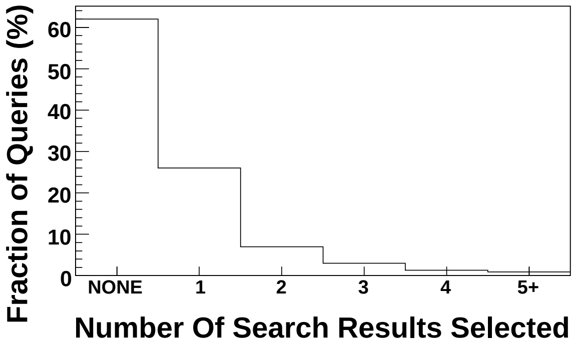

Figure 2: Distribution of total number of selections per query.

clicks depending on whether the specific web page is listed on the SERP. Additionally, it may be that the current sample of users is biased towards searchers who are less likely to click on links.

Figure 2 shows the distribution of the total number of selections per query. 62% of all queries lead to the selection of zero search results. Since Surf Canyon does nothing until after the first selection, this number is intrinsic to the current users interacting with these particular search engines. A recent study by Downey, Dumais and Horvitz also showed that after a query the user's next action is to re-query or end the search session about half the time [11]. In our study, only 12% of queries lead to more than one user selection. A goal of implicit real-time personalization would be to decrease direct query reformulation and to increase the number of informational queries that lead to multiple selections. The current data sample is insufficient to study whether this goal has been achieved.

In order to evaluate the implicit personalization technology developed by Surf Canyon we chose to compare the actions of the same set of users with and without the implicit personalization technology enabled. Our baseline control sample was created by randomly replacing recommended search results with random search results selected from among the results with initial ranks 11-200. These "Random Recommendations" were only shown for 5% of the cases where recommendations were generated. The position (1, 2, or 3) in the recommendation list was also random. These random recommendations were not necessarily poor, as they do come from the list of results generated by the search engine in response to the query.

Figure 3 shows the click frequency for Surf Canyon recommendations as a function of the position of the recommendation relative to the last selected search result. Position 1 is immediately below the last selected search result. Also shown are the click frequencies for "Random Recommendations" placed at the same positions. In both cases, the frequency is relative to the total number of recommendations shown at that position. The increase in click rate (∼60%) is constant within statistical uncertainties for all recommended link positions. Note that the recommendations are generated each time a user selects a link and are considered to be shown even if the user does not return to the SERP. The low absolute click rates (3% or less) are due to the fact that users do not often click on more than one search result as discussed above. The important point, however, is that the Surf Canyon implicit relevance feedback technology increases the click frequency by ∼80% compared to the links presented without any real-time user-intent modelling. The relative increase in clickthrough rate is constant (within statistical errors) for all display positions even though the absolute clickthrough rates rapidly drop as funciton of display position.

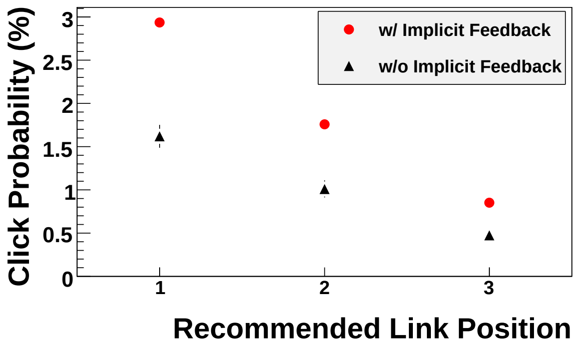

Figure 3: Probability (%) that a recommended search result will be clicked as a function of display position relative to the last selected search result. The red circles are for recommendations selected using Surf Canyon's instantaneous relevancy algorithm, while the black triangles are for the random control sample that does not incorporate relevance feedback.

Figure 4 shows the per query distribution of initial search result ranks for all selected search links in the current data sample. The top 10 links are selected most frequently. Search results beyond 10 are all displayed using Surf Canyon's algorithm (either through a bull's eye selection, a link selection, or when the user selects more results). For the results displayed by Surf Canyon (initial ranks > 10), the selection frequency follows a power-law distribution with $P(IR) = 38\% * IR^{-1.8}$, where IR is the initial rank.

As Surf Canyon's algorithm favors links with higher initial rank, the click frequency distribution does not fully reflect the relevancy of the links as a function of initial rank. Figure 5 shows the probability that a shown recommendation is clicked as a function of the initial rank. This is only for recommendations shown in the first position below the last selected link. After using Surf Canyon's instantaneous relevancy algorithm, this probability shows at most a weak dependence on the initial rank of the search result. The dotted link shows the result of a linear regression to the data, $P(IR) = 3.2 - (0.0025 \pm 0.00101) * IR$. When sufficient data is available we will repeat the same analysis for "Random Recommendations" as that will give us a user-interface independent estimate of the relative relevance for deep links in the search result set before the application of the implicit feedback algorithms.

For the second and subsequent results pages, the browser extension has complete control over all displayed search results. For a short period of time we produced search results pages that mixed Surf Canyon's top ranked results with results having the top initial ranks from the search

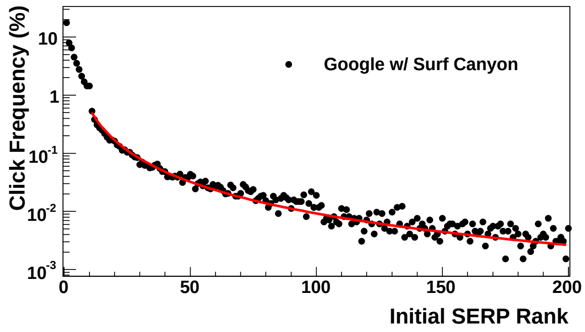

Figure 4: Frequency per non-repeated search query for link selection as a function of initial search result rank.

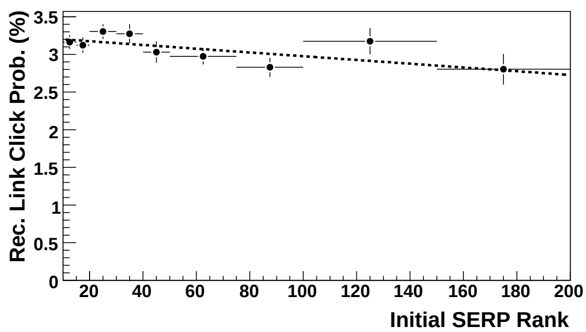

Figure 5: Probability that a *displayed* recommended link is selected as a function of the initial search result rank. This data only include links from the first position immediately below the last selected search result.

engine. This procedure was proposed by Joachims as a way to use clickthrough data to determine relative user preference between two search engine retrieval algorithms [12]. Each time a user requests "More Results", two lists are generated. The first list (SC) contains the remaining search results as ranked by the Surf Canyon's instantaneous relevancy algorithm. The second list (IR) contains the same set of results ranked by their initial display rank from the search engine. The list of results shown to the user is such that the top k_{SC} and k_{IR} results are displayed from each list, with $|k_{SC} - k_{IR}| < 1$. Whenever $k_{SC} = k_{IR}$ the next search result is taken from one of the lists chosen at random. Thus, the topmost search result on the second page will reflect Surf Canyon's ranking half the time and the initial search result order half the time. By mixing the search results this way, the user will see, on average, an equal number of search results from each ranking algorithm in each position on the page. The users have no way of determining which algorithm produced each search result. If the users select more search results from one ranking algorithm compared to the other ranking algorithm it demonstrates an absolute user preference for the retrieval function that led to more selections.

Figure 6 shows the ratio of link clicks for the two retrieval functions. IR is the retrieval function based on the result rank returned from the search engine. SC is the retrieval function incorporating Surf Canyon's implicit relevance feedback technology. The ratio is plotted as a function of the number of links selected previously for that query. Previously selected links are generally considered to be positive content feedback. If, on the other had, no links were selected then the algorithm bases its decision exclusively on negative feedback indications (skipped links) and on the user intent model that may have been developed for similar recent related queries.

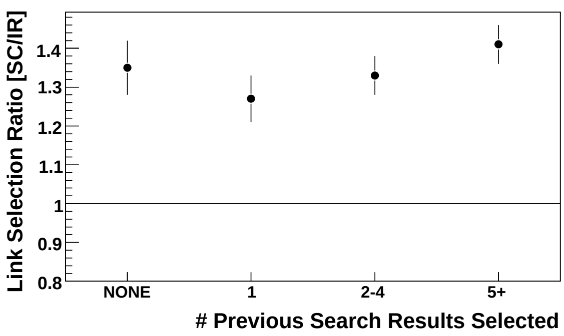

Figure 6: Ratio of click frequency for second and subsequent search results page links ordered by Surf Canyon's Implicit Relevance Feedback algorithm (SC) compared to links ordered by the initial search engine result rank (IR).

We observe that, independent of the number of previous user link selections in the same query, the number of clicks on links from the relevance feedback algorithm is higher than links displayed because of their higher initial rank. This demonstrates an absolute user preference for the ranking algorithm that utilizes implicit relevance feedback. Remarkably, the significant user preference for search results retrieved using the implicit feedback algorithm is also apparent when the user had zero positive clickthrough actions on the first 10 results. After skipping the first 10 results and asking for a subsequent set of search links, the users are ∼35% more likely to click on the top ranked Surf Canyon result compared to result # 11 from Google. Clearly, the searcher is not so interested in search results produced by the identical algorithm that produced the 10 skipped links and an update of the user intent model for this query is appropriate.

6. CONCLUSIONS AND FUTURE DIRECTIONS

Surf Canyon is an interactive information retrieval system that dynamically modifies the SERP from major search engines based on implicit relevance feedback. This was built with the goal of relieving the growing user frustration with the search experience and to help searchers "find what they need right now". The system presents recommended search results based on an instantaneous user-intent model. By comparing clickthrough rates, it was shown that real-time implicit personalization can dramatically increase the relevancy of presented search results.

Users of web search engines learn to think like the search engines they are using. As an example, searchers tend to select words with high IDF (inverse document frequency) when formulating queries – they naturally select the rarest terms that they can think of that would be in all documents they desire. Excellent searchers can often formulate sufficiently specific queries after multiple iterations such that they eventually find what they need. Properly implemented implicit relevance feedback would reduce the need for query reformulations, but it should be noted that in the current study most users had not yet adjusted their browsing habits to the modified behavior of the search engine. By tracking the current users in the future we hope to see changes in user behavior that can further improve the utility of this technology. As the user-intent model is cumulative, more interaction will produce better recommendations *once* the users learn to trust the system.

7. REFERENCES

[1] J.J. Rocchio. *The Smart Retrieval System Experiments in Automatic Document Processing.* Prentice Hall, 1971.

[2] D. Harman. Relevance feedback revisited. In *Proceedings of the Fifteenth International ACM SIGIR Conference*, pages 1–10, 1992.

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

[4] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR '05*, 2005.

[5] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23(2):147–168, April 2005.

[6] Xuehua Shen, Bin Tan, and ChengXiang Zhai. Context-sensitive information retrieval using implicit feedback. In *SIGIR '05*, 2005.

[7] Xuehua Shen, Bin Tan, and ChengXiang Zhai. Implicit user modelling for personalized search. In *CIKM '05*, 2005.

[8] B. Jansen and A. Spink. How are we searching the world wide web?: a comparison of nine search engine transaction logs. *Information Processing and Management*, 42(1):248–263, 2006.

[9] B. Jansen and A. Spink. An analysis of web documents retrieved and viewed. In *The 4th International Conference on Internet Computing*, pages 65–69, 2003.

[10] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *The First International Conference on Scalable Information Systems*, 2006.

[11] D. Downey, S. Dumais, and E. Horvitz. Studies of web search with common and rare queries. In *SIGIR '07*, 2007.

[12] T. Joachims. Unbiased evaluation of retrieval quality using clickthrough data. In *SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*, 2002.