Can People Collaborate to Improve the Relevance of Search Results?

Arun Kumar Agrahri
Institute of Technology,
University of Minnesota, Twin Cities
Computer Science
agrahri@cs.umn.edu

Divya Anand T.M.
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham,
Cyber Security
anandivya@gmail.com

John Riedl
Institute of Technology,
University of Minnesota, Twin Cities
Computer Science
riedl@cs.umn.edu

ABSTRACT
Search engines are among the most-used resources on the internet. However, even today's most successful search engines struggle to provide high quality search results. According to recent studies as many as 50 percent of web search sessions fail to find any relevant results for the searcher. Researchers have proposed *social search* techniques, in which early searchers provide feedback that is used to improve relevance for later searchers. In this paper we investigate foundational questions of social search. In particular, we directly assess the degree of agreement among users about the relevance ranking of search results. We developed a simulated search engine interface that systematically randomizes Google's normal relevance ordering of the items presented to users. Our results show that (a) people are biased toward items in the top of the search lists, even if the list is randomized; (b) people explicit feedback is not biased and (c) people's shared preferences do not always agree with Google's result order. These results suggest that social search techniques might improve the effectiveness of web search engines.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering, relevance feedback, search process, selection process;*

H.5.3 **[Information Interfaces and Presentation]**: Group and Organization Interfaces – *collaborative computing, computer-supported collaborative work, evaluation/methodology.*

General Terms
Measurement, Design, Experimentation, Human Factors

Keywords
Social access patterns, social search, browsing, explicit feedback, recommender systems

RecSys'08, October 23-25, 2008, Lausanne, Switzerland.

1. INTRODUCTION
In a typical web search session, a rational searcher might be expected to assess each of the page summaries against their information needs and click on the one that appears to be the most relevant while a novice searcher may follow search engines ranking order. In practice, people preferentially select items near the top of the list [6], based on their expectations of the quality of the search engine results, or on laziness. In recent studies researchers have found that because search engines repeatedly return currently popular pages at the top of search results, popular pages tend to get even more popular, while unpopular pages get ignored by an average user [2, 3]. This "rich-get-richer" phenomenon is particularly problematic for new and high-quality pages which may never get the chance to get top listing despite being highly relevant for the requested query. In this paper we address foundational questions that underlie these approaches, in an attempt to develop insight into the directions likely to improve the relevance of search results.

Recent studies have addressed the "rich-get-richer" phenomenon. Cho et al. proposed a new ranking metric by considering not just the current link structure, but also the *evolution* and *change* in the link structure [4]. Pandey et al. proposed partial randomization of rank positions such that new pages can get noticed by the users [5]. Smyth et al. has shown that people are biased in selecting top results items in both normal versus reverse order [6]. The study suggests that people are biased in their use of search engines. Thus, search engines could misleadingly over-promote an initially popular page which might not be very relevant for users information needs. In our attempt to further investigate this behavior, we designed a simulated search engine environment to capture user response by presenting Google top ten results in randomized order.

H1: We hypothesize that users will prefer to rate results at the top of the results list, whether the results are randomized, or in the order that Google presents them

Studies have shown that people are biased in their search activities [1, 6]. Typically relevance is built by aggregating user's implicit feedback. However, implicit feedbacks are biased towards the top ranked items, laziness or random click behavior. Thus, such inferences for relevance will produce biased ranking; decreasing the overall quality of search results. Here, we attempt to test if users are biased in their explicit feedback for relevance.

H2: We hypothesize that users explicit relevance ranking are not biased by the rank of items in the result list

In order to test this hypothesis we presented Google results items in randomized order. A Normal order will list the Google's most relevant items on the top; probably user will also rate these items as relevant. Thus, we don't know if this rating was a result of biasing or the result item is actually relevant. Randomization will bring some low ranked results with high relevant content to top and thus, different ratings are expected for top selected items.

Search on the web has been a tedious task for novice users. They often fail to provide the right set of keywords (when a typical query length is two words long [7]) to aptly express their information needs and hence, the results obtained in a ranking order by the search engine will have to be filtered again by the user to find the content more relevant to his query needs. Recent studies have found that as many as 50 percent of the web search sessions fail to find any relevant results for the searcher [1, 7]. Also, more than 90 percent of search session doesn't go beyond the first results page [1]. Thus, ordering of search results play a vital role in realizing a query fetch into a successful search. This study assesses if people can collaborate to improve the relevance ranking of search results. We examined the effects of two different ways of reordering search results, to understand the impact on users.

H3: We hypothesize that for some types of query people collaborative effort can produce better ordering of search results.

2. METHODS

2.1 Collaborative Search Ranking

The study was conducted in two phases. During the first phase, 145 e-mail invitations were sent to participants to rate query search results for their relevance. In the second phase, novice users evaluated user-based versus Google ordering. In order to make the study as accessible as possible we chose popular queries. We perused the AOL search logs and selected six categories of queries which appeared most frequently: *shopping, health, technology, business, computers and arts.* We then collected the most popular queries from three social search sites namely Eurekster, Mahalo, del.icio.us[1].

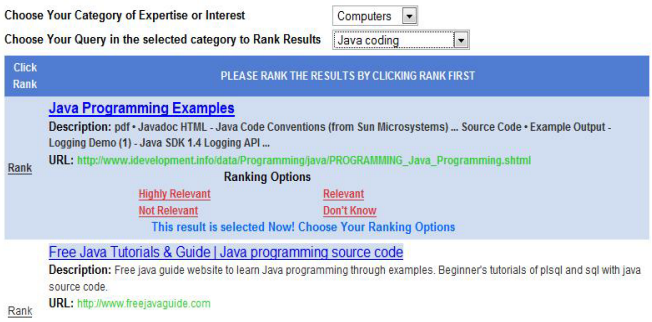

Figure 1: A screenshot showing rating interface for chosen query

Recent studies [3, 4, 5, 6] have shown that presenting result items in normal order will be unfair approach for improving the relevance of new or high quality pages. Thus, presenting in Google normal order will affect our ranking algorithm and results may be biased if raters always choose to rate the top results items. So we randomized our result lists so that each page gets equal chance to appear in top list. We created a simulated search engine environment systematically randomizing Google's normal relevance ordering of the items. For each query the simulated system was built by extracting top ten results with their result title, description and source URLs from Google. Participants in the first phase were asked to rate as many of the queries as they wished. Rather than asking participants to rate any specific query we asked them to rate their own preferred ones. Giving control to users for query selection has an advantage that they could choose their area of interest or expertise resulting in better rating of results items. Participants began the study by visiting the website www.searchranking.qsh.in, which they could do on their own preferred time from their home or workplace. Participants could base their ranking on the title of the Web page, the short blurb provided by Google, or they could visit the web page itself to determine the relevancy of the item. Participants could rate any number of items for the selected query and relevance was measured on a 4-point scale (highly relevant (3); relevant (2); don't know (1); not relevant (-1)). For each item rated we recorded the participant response and updated its rating in the database.

After rating the queries each participant was asked to take a short survey to understand their interaction with search engines. The purpose of this survey is to study how experience in using search engines affects the perception of relevance.

2.2 Users versus Google Ordering

After receiving ratings from at least one contributor for a given query, we began our evaluation phase of comparing user versus search engine ordering of search results. We chose queries which were rated by a least three users, such that queries have enough ratings for fair comparison. 15 of 30 queries met this requirement. We randomly selected 20 of the participants who had self-identified as novice searchers. Participants were sent email invitations to take part in the evaluation phase. Participants were asked to choose their own preferred query to compare rank ordering of the two approaches. The interface (Figure 2) shows two result lists, with the left side showing a user-based ordering and the right side showing Google ordering. We computed the score of a result item as

$$Score = 3\times HRcount + 2\times Rcount + DNcount + (-1)\times NRcount$$

Where *HRcount* is the number of times the result item was rated as "highly relevant"; *Rcount* is the number of times the item was rated as "relevant"; *DNcount* is the number of times the item was rated as "don't know" and *NRcount* is the number of times the item was rated as "not relevant".

Figure 2: A screenshot showing evaluation interface left side with expert-based ranking and right side with Google ranking

[1] All of the datasets used in this study, including the queries and survey questions, are available for other researchers at http://www.searchranking.qsh.in/.

3. RESULTS AND DISCUSSION

Survey on Search Habits

In presenting the results, it is useful to distinguish among three groups of participants. We refer to the 20 inexperienced searchers as *novices*. We refer to the 90 participants who rated search results as *raters*. In addition to the ratings from the users, we also carried out a survey of their subjective impressions. The survey was carried out immediately after the users completed rating the search results. Of a total of 145 participants 60 completed the survey. We refer to the 60 raters who completed the survey as *respondents*. One of the questions in the survey asked the respondents to explain their rationale in determining the relevance of the search results. We offered respondents four choices: *Trust Search Engine Ordering; Don't Trust Search Engine Ordering, Prefer Trusted and Popular Web Sources; Random Click Behavior*, and they could choose more than one option to express their search behavior. 35% of the respondents reported that they trust the ordering of the result list returned by the search engine. Of this 35% only 15% reported that they consider trusted and popular web sources to be more relevant than unfamiliar sources. Again, from this 35%, 20% reported random click behavior among the top result items expecting the top results to be most relevant. Of the whole group of respondents, 42% reported that they do not trust search engine ordering and select results items based on item's title, snippet and URLs source information. Of this 42%, nine percent also reported that the popularity of web source didn't add much value to the relevance of result item. 10% of all respondents reported that they choose random results in expectation that one of them would lead to the targeted document

.

We also asked our respondents for their most frequent search activities. Again, contributors were given four choices (*Re-Find:* To re-find previously visited page*; URLs Search:* like bookmark applications*; Web docs Search:* Searching information online, typical search queries*; don't use search engine*) Figure 3 lists the different applications for which respondents use a search engine. 32% of the respondents reported that re-find activity is very common in their interaction with search engines. 38% of the respondents reported that they frequently use search engines to locate URLs of webpage like a bookmark application while more than 63% of the respondents reported that most of the time they use search engines to find web documents that meets their information needs. While the results of our survey are consistent with prior analysis of search logs [1]; they also motivate the development of several search applications. Several search interfaces are beginning to be developed that takes advantages of these results [8, 9].

In our attempt to test the hypotheses, our first measure was to capture the link chosen by the user to rank in a returned results list to a given query. We analyzed the data separately with queries that had fewer than 10 results items rated by people than queries that had all ten result items rated. For each of these queries we recorded the results item rank in our randomized search ordering environment. We hypothesize that if the rating of the chosen item is always high for top items; then people are biased in rating while if the ratings vary we can conclude people are not biased. Our result shows that people have preference for selecting top items for their search activity but their explicit feedback for relevance is not biased.

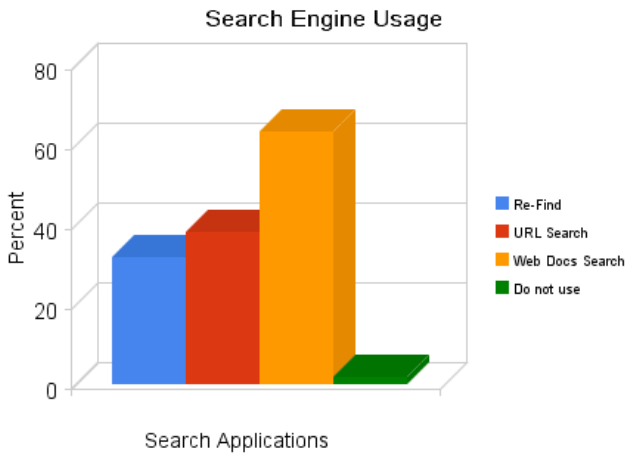

Figure 3: depicts respondents most frequent activities at search engines

The results are summarized as: among the queries which had fewer results item rated, 85% of the time both first and second item of the randomized search results were chosen for rating; while only 20% of the time people went beyond the fifth item in the list for rating. Figure 4 depicts that users were more likely to rate items near the top of the list than other items and shows the frequency that they rated the results item highly relevant given it was chosen from top list. For instance, whenever first item was chosen, 70% of the times it was rated *highly relevant* while for second ranked item only 30% of times it was rated as *highly relevant.* This clearly indicates that people were not biased in their explicit search relevance feedback though partially biased with top result items.

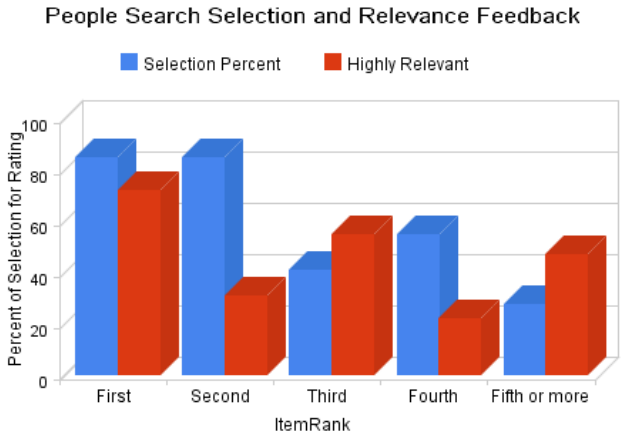

Figure 4: first bar represents percent of selection of the item for rating; second bar depicts the percent of times when item was rated as "highly relevant"

The significance of what is happening is apparent if we consider the items by position in the results lists. It is interesting, this finding occurred even when people were told both in the invitation and FAQs section that the results list will be randomized to avoid bias in their selection. Either participants did not read the instructions carefully or they preferred to rate top items even though they were randomly selected.

Our second attempt was to measure the difference in ranking of search results items between raters versus Google. We plotted this difference versus number of items having the same difference value. Figure 5 depicts that the difference follows a distribution which is close to normal distribution with ~ N (0, 3.14). On further investigating the results, we found that people have a shared preference of search ordering for some types of queries which are not consistent with Google ordering. While it is challenging to predict which type of queries works best with social search, we attempt to answer this with our limited dataset. In our attempt to answer this question, we collected items which were rated by more than three users. We found that users have an very different perception of relevance of search results with queries of category *shopping*:Digital Cameras, Walking Shoes (mean difference in ranking = 4.2); while users had a consistent view with Google for queries of category *Business:*Microsoft Bid for Yahoo, Online Advertisement (mean ranking difference = 0.8)

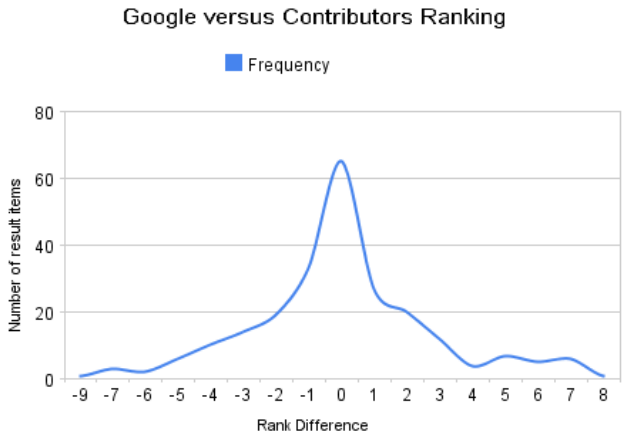

Figure 5: depicts the difference in ranking of two approaches versus number of items having the same difference value

To test our hypothesis if people can collaborate to provide better search ordering, we ran an evaluation study where we asked 20 *novices* to evaluate the users versus Google ranking of results lists for their prefered query. The results is summarized as: 70 percent of the *novices* reported that users-based ordering of results items outperformed the Google search ordering . These 70% users choose queries from categories Shopping, Computers, Arts. While 30 percent of the *novices* reported that Google ordering was more relevant than users-recommended ordering for the chosen query; most of their queries were chosen from categories Business, Technology.

4. CONCLUSION

On the whole, our study suggests that people prefer top items in the result list, whether the results are randomized or in the order Google presents them. Our results also suggests that users explicit feedback for relevance of results items are not biased. Finally, we also found that Google predicted search ordering is inconsistent with users perception of relevance for some types of queries. While our data is limited in predicting types of queries which works best with social search, a fruitful area for future work is to further explore the predicion of such query types. Though our study only considers the first 10 results for each query which may have approximately equal relevance, we were still able to present the difference in the perception in ordering of search results by the users and Google. Our survey results suggests that future information delivery system have to learn from users search behavior.

5. REFERENCES

[1] Pass, G., Chowdhury, A., Torgeson, C. 2006. Picture of Search. *In Proceedings of InfoScale 2006*

[2] Baeza-Yates, R., Saint-Jean, F., Castillo, C. 2002. Web dynamics age and page quality. *In Proceedings of SPIRE 2002*

[3] Cho, J., Roy, S. 2004. Impact of search engines on page popularity. *In Proceedings of WWW 2004*

[4] Cho, J., Adams, R.E. 2005. Page quality: In search of an unbiased Web ranking. *In Proceedings of ACM SIGMOD 2005*

[5] Pandey, S., Roy, S., Olston, C., Cho, J., Chakrabarti, S. 2005. Shuffling a Stacked Deck: The Case for partially randomized ranking of search engine results. *In Proceedings of VLDB 2005*

[6] Keane, T., O'Brien, M., Smyth, B. 2008. Are people biased in their use of search-engines? *Communications of the ACM, 51(2)*

[7] Spink, A., Wolfram, D., Jansen, B.J., Aracevic T. 2001. Searching the Web: The public and their queries. *Journal of the American Society of Information Science and Technology,* 52(3)

[8] Amershi, S., Morris, M. 2008. CoSearch: System for co-located collaborative Web search. *In Proceedings of 26th CHI Conference*

[9] Morris, D., Venolia, G. 2008. SearchBar: A search centric Web history for task resumption and information re-finding, *In Proceedings of 26th CHI Conference*

[10] Smyth, B. 2007. A community-based approach to personalizing Web search. *IEEE Computer*, 8(40)

[11] Teevan, J., Adar, E., Jones, R., Potts, M.A.S. 2007. Information re-retrieval: repeat queries in Yahoo's logs. *In Proceedings of ACM SIGIR 2007*

[12] Aula, A., Jhaveri, N., Kaki, M. 2005. Information search and re-access strategies of experienced Web users. *In Proceedings of WWW 2005*