Beyond the Commons: Investigating the Value of Personalizing Web Search

Jaime Teevan[1], Susan T. Dumais[2], and Eric Horvitz[2]

[1] MIT, CSAIL, 32 Vassar St., G472,
Cambridge, MA, 02139 USA
teevan@csail.mit.edu
http://www.csail.mit.edu/~teevan
[2] Microsoft Research, One Redmond Way,
Redmond, WA, 94114, USA
{sdumais,horvitz}@microsoft.com

Abstract. We investigate the diverse goals people have when they issue the same query to a Web search engine, and the ability of current search tools to address such diversity, in order to understand the potential value of personalizing search results. Great variance was found in the results different individuals rated as relevant for the same query—even when those users expressed their underlying informational goal in the same way. The analysis suggests that, while current Web search tools do a good job of retrieving results to satisfy the range of intentions people may associate with a query, they do not do a very good job of discerning an individual's unique search goal. We discuss the implications of this study on the design of search systems and suggest areas for additional research.

1 Introduction

Traditional search engines are designed to return a set of documents that match a query. Studies of search engine quality have tended to be based on the ability of search engines to return the set of results that its users want as a population, as opposed to the results that match each individual's unique search goal. For example, at the DARPA Text REtrieval Conference (TREC), relevant documents to a particular query are identified by an expert judge, based on a detailed description of an information need. Ideally the description is explicit enough and the rater skilled enough that the documents selected as relevant are the same ones that another rater would consider relevant.

However, Web search behavior suggests that providing results to an unambiguous query might not be the most appropriate design target for a search engine. Web queries are very short, and it is unlikely that a two- or three-word query can unambiguously describe a user's informational goal. What one person considers relevant to a query like "jaguar" is not necessarily the same as what someone else considers relevant to the same query. Even a seemingly precise query like "PIA 2005" returns Web pages about the Personal Information Access workshop, the Parachute Industry Asso-

ciation, Professional Insurance Agents, the Pacific Institute of Aromatherapy, etc. Further, if Web searchers are not skilled at stating their goal, even longer descriptions may not reliably disambiguate intent.

We report on a study of the ability of current Web search engines to provide relevant documents to users, in order to understand how future search tools can be built to best meet the needs of their users. Understanding relevance is a complex problem [11, 13], and we address only a small portion of it in our work. Our analysis is aimed at assessing the relationship between the rank of a search result as returned by a Web search engine and the individual's perceived relevancy of the result. We find a considerable mismatch due to a variation in the informational goals of users issuing similar queries. The study suggests personalization of results via re-ranking would provide significant benefit for users. We conclude with a discussion of how the results of this study should triage future research.

2 Methods

We conducted a study in which 15 participants evaluated the top 50 Web search results for approximately 10 queries of their choosing. Participants were employees of a large corporation. Their job functions included administrators, program managers, software engineers and researchers. All were computer literate and familiar with Web search.

Web search results were collected from a "Top Choice" search engine, as listed by Search Engine Watch. For each search result, the participant was asked to determine whether they personally found the result *highly relevant*, *relevant*, or *irrelevant*. So as not to bias the participants, the results were presented in a random order.

The queries evaluated were selected in two different manners, at the participants' discretion. In one approach (*self-selected queries*), users were asked to choose a query to mimic a recently performed search, based on a diary of searches they were asked to keep during the day. Thus, we believe that the self-selected queries closely mirrored the searches that the participants conducted in the real world.

In another approach (*pre-selected queries*), users were asked to select a query from a list of queries that were formulated to be of general interest (e.g., *cancer, Bush, Web search*). Although users did not generate these queries themselves, they were free to choose the pre-selected queries they found most interesting, and thus presumably only chose queries that had some meaning to them. By using pre-selected queries, we were able to explore the consistency with which different individuals evaluated the same results. Such data would have been difficult to collect using only self-selected queries, as it would have required us to wait until different participants coincidentally issued the same query on their own. We validate the conclusions drawn from pre-selected queries with data from the self-selected queries.

For both the self-selected queries and the pre-selected 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. Because the pre-selected queries were given to the user, the user had to create some intent for these queries. However, by allowing

them to decide whether or not they wanted to evaluate a particular query, we sought to provide them with a query and associated results that would have some meaning for them.

We collected a total of 137 queries. Of those, 53 were pre-selected queries and 85 were self-selected. The number of users evaluating the same set of results for the pre-selected query ranged from two to nine. Thus we had evaluations by different people for the same queries drawn from the pre-selected set of queries, as well as a number of evaluations for the searches that users had defined themselves.

3 Rank and Rating

We used the data we collected to study how the results that the Web search engine returned matched our participants' search goals. We expected them to match relatively closely, as current search engines seem to be doing well, and in recent years satisfaction with result quality has climbed.

Fig. 1 shows the average result's relevancy score as a function of rank. To compute the relevancy score, the rating *irrelevant* was given a score of 0, *relevant* a score of 1, and *highly relevant* a score of 2. Values were averaged across all queries and all users. Separate curves are shown for the pre-selected (solid line) and self-selected (dashed line) queries. Clearly there is some relationship between rank and relevance. Both curves show higher than average relevance for results ranked at the top of the result list. The correlation between rank and relevance is -0.66. This correlation coefficient is significantly different from 0 ($t(48) = 6.10$, $p < 0.01$). However, the slope of the curves flattens out with increasing rank. When considering only ranks 21-50, the correlation coefficient is -0.07, which is not significantly different from 0. Importantly, there are still many relevant results at ranks 11-50, well beyond what users typically see. This suggests the search result ordering could be improved.

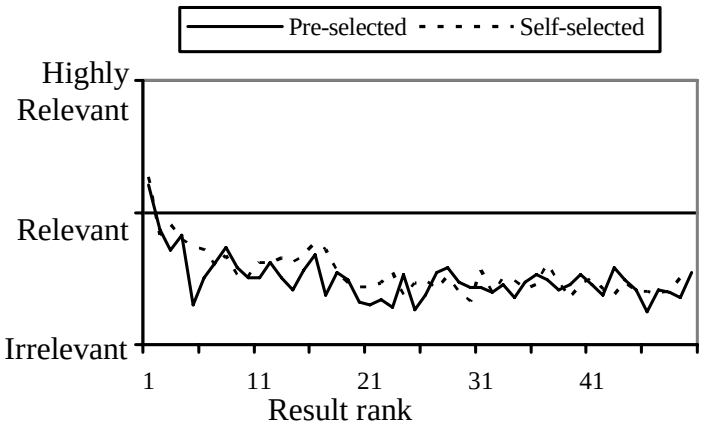

Fig. 1. Average ratings for Web search engine results as a function of rank. There are many relevant results that do not rank in the top ten

The general pattern of results seen in Fig. 1 is not unique to our sample of users or queries. A reanalysis of data from the TREC Web track [4] shows a similar pattern. In the TREC-9 Web track, the top 100 results from 50 Web queries were rated using a similar three-valued scale, *highly relevant*, *relevant* and *not relevant*. Results for one top-performing search systems, uwmt9w10g3, yielded an overall correlation between rank and relevance of -0.81, which drops off substantially to -0.30 for positions 21-50.

4 Same Query, Different Intents

Our analysis shows that rank and rating were not perfectly correlated. While Web search engines do a good job of ranking results to maximize their users' global happiness, they do not do a very good job for specific individuals. If everyone rated the same currently 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), our study demonstrated a great deal of variation in their rating of results.

As will be discussed in the following sections, we found that people rated the same results differently because they had different information goals or intentions associated with the same queries. This was evidenced by the variation in the explicit intents our participants wrote for their queries. Even when the intents they wrote were very similar, we observed variation in ratings, suggesting that the participants did not describe their intent to the level of detail required to distinguish their different goals.

4.1 Individuals Rate the Same Results Differently

Participants did not rate the same documents as relevant. The average inter-rater agreement for queries evaluated by more than one participant evaluated was 56%. This disparity in ratings stands in contrast to previous work. Although numbers can't be directly compared, due to variation in the number of possible ratings and the size of the result set evaluated, inter-rater agreement appears to be substantially higher for TREC (*e.g.*, greater than 94% [8]) and previous studies of the Web (*e.g.*, 85% [3]). The differences we observed are likely based in our focus on understanding personal intentions; instead of instructing our participants to select what they thought was "relevant to the query," we asked them to select the results they would want to see personally.

The ratings for some queries agreed more than others, suggesting some queries might be less ambiguous to our population than others. Similarly, some participants gave ratings that were similar to other participants' ratings. It might be possible to cluster individuals, but even the most highly correlated individuals showed significant differences.

4.2 Same Intent, Different Evaluations

We found that our participants sometimes used the same query to mean very different things. 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". This was evident both for the pre-selected, where the user had to come up with an intent based on the query, and self-selected queries, where the query was generated to describe the intent. Although we did not observe any duplicate self-selected queries, many self-selected 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.

Interestingly, even when our participants expressed the same intent for the same query, they often rated the query results very differently. 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 their intent, only one URL (www.microsoft.com) was given the same rating by all three individuals. Thirty-one of the 50 results were rated *relevant* or *highly relevant* by one of these three people, and for only six of those 31 did more than one rating agree. The average inter-rater agreement among these three users with similar intentions was 62%.

This 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*" ("UW" → "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-selected queries, where we expected they might have some difficulty creating an intent (mitigated by the fact that they only rated pre-selected queries by choice), but also for the self-selected queries.

Although explicit intents generally did not fully explain the query term, they did 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 include this additional information in their original query, but it does suggest that they could perhaps be encouraged to provide more information about their target when searching. However, even if they did this, they would probably still not be able to construct queries that expressed 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 searchers either need help communicating their intent or that search systems should try to infer it.

5 Search Engines are for the Masses

The previous sections showed that our participants ranked things very differently, in ways that did not correspond closely with the Web search engine ranking. We now describe analyses that show that the Web ranking did a better job of satisfying all of our participants than any individual.

5.1 Web Ranking the Best for the Group

In this section, we investigate the best possible ranking we could construct based on the relevance assessments we collected, and compare this ideal ranking with the original Web ranking. For scoring the quality of a ranking, we use *Discounted Cumulative Gain* (DCG), a measure of the quality of a ranked list of results commonly used in information retrieval research [5]. DCG measures the result set quality by counting the number of relevant results returned. It 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). DCG also allows us to incorporate the notion of two relevance levels by giving *highly relevant* documents a different gain value than *relevant* documents.

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

For *relevant* results, we used G(i) = 1, and for *highly relevant* results, G(i)=2, reflecting their relative importance.

The best possible ranking for a query given the data we collected is the ranking with the highest DCG. For queries where only one participant evaluated the results, this means ranking *highly relevant* documents first, *relevant* documents next, and *irrelevant* documents last. When there are more than one set of ratings for a result list, the best ranking ranks first those results that have the highest collective gain.

We compared how close the best possible rankings were to the rankings the search engine returned. To measure "closeness," we computed the Kendall-Tau distance for partially ordered lists [1]. The Kendall-Tau distance counts the number of pair-wise disagreements between two lists, and normalizes by the maximum possible disagreements. When the Kendall-Tau distance is 0, the two lists are exactly the same, and when it is 1, they are in reverse order. Two random lists have, on average, a distance of 0.5.

We found that for eight of the ten queries where multiple people evaluated the same result set, the Web ranking was more similar to best possible ranking for the group than it was, on average, to the best possible ranking for each individual. The average individual's best ranking was slightly closer to the Web ranking than random (0.5), with a distance of 0.469. The average group ranking was significantly closer ($t(9) = 2.14$, $p < 0.05$) to the Web ranking, with a distance of 0.440. The Web rankings seem to satisfy the group better than they do the individual.

5.2 Gains of Personalization via Re-ranking

Again taking DCG as an approximation of user satisfaction, we found a sizeable difference between our participants' satisfaction when given exactly what they wanted rather than the best group ranking for that query. On average, the best group ranking yielded a 23% improvement in DCG over what the current Web ranking, while the best individual ranking led to a 38% improvement.

The graph depicted in Fig. 2 shows the average DCG for group (dashed line) or personalized (solid line) rankings. These data were derived from the five pre-selected queries for which we collected six or more individual evaluations of the results, although the pattern held for other sets of queries. To compute the values shown, for each query we first randomly selected one person and found the DCG for that individual's best ranking. We then continued to add the additional people, at each step recomputing the DCG for each individual's best rankings and for the best group ranking. As can be seen in Fig. 2, as additional people were added to the analysis, the gap between user satisfaction with the individualized rankings and the group ranking grew. Our sample is small, and it is likely that the best group ranking for a larger sample of users would result in even lower DCG values.

These analyses underscore the promise of providing users with better search result quality by personalizing results. Improving core search algorithms has been difficult, with research leading typically to very small improvements. We have learned that, rather than improving the results to a particular query, we can obtain significant boosts by working to improve results to match the intentions behind it.

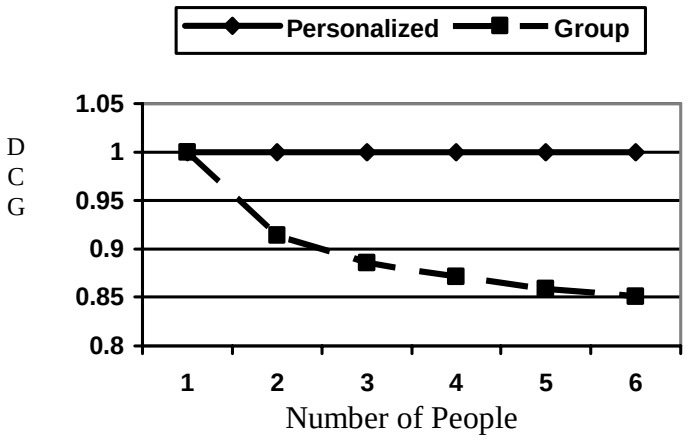

Fig. 2. As more people are taken into account, the average DCG for each individual drops for the ideal group ranking, but remains constant for the ideal personalized ranking

6 Directions in Personalized Search

We believe that Web search tools could be enhanced significantly by considering the variation in relevancy of results for users. We shall now touch on several directions for doing such personalization suggested by the above analysis.

We observed that our participants rated the results to the same queries differently because they had different intents. One solution to ambiguity is to aid users in better specifying their interests and intents. As an example, Google Personal [4] asks users to build a profile of themselves by specifying their interests. Other search systems have tried to help users better express their informational goals through techniques such as relevance feedback or query expansion. While it appears people can learn to use these techniques [2, 8], in practice, on the Web they do not appear to improve overall success [2, 3], and such features have been found to be used rarely. We agree with Nielsen [10], who cites the importance of not putting extra work on the users for personalization. Also, even with additional work, it is not clear that users can be sufficiently expressive. Participants in our study had trouble fully expressing their intent even when asked explicitly to elaborate on their query. In related work, people were found to prefer long search paths to expending the effort to fully specify their query [11].

We believe that another promising approach to personalizing search is to infer users' information goals automatically. Kelly and Teevan [7] give an overview of research done in information retrieval on how implicit measures can be used to help search, highlighting prior contributions focused on helping to improve results for individuals, versus for the general population. In a related paper [15], we describe a search personalization prototype that we have developed which builds on the lessons learned from the study described in this paper. The prototype, named *PS*, uses a person's prior interactions with a wide variety of content to personalize that person's current Web search in an automated manner.

Our study suggests that the results returned by Web search engines represent a range of intentions that people associate with queries. Thus, we believe that personalized search systems could take current Web search results as a starting point for user-centric refinement via re-ranking (e.g., [9, 15]). The original ranking of results by a Web search engine is a useful source of information for a more personalized ranking, and, as we discovered, the first several results are particularly likely to be relevant.

We found that not all queries should be handled in the same manner. For example, we observed that some queries appeared less ambiguous than others and showed less variation among individuals. For such queries, the group ranking (i.e., the current Web search ranking) might be sufficient. A search system that allows users to control how much personalization they receive would improve search relevance while following Neilson's [10] suggestion that users be given control of their content instead of having personalization imposed on them.

7 Conclusion

We have found that there is promise in building tools that perform personalization via re-ranking the results currently provided by current search engines. We have not discussed specific methods to automatically identify users' intentions. Instead we have worked to characterize the range of informational goals associated with queries, and investigated the potential value that can be seen by users via methods that re-rank the list of results provided by search engines.

References

1. Adler, L. M.: A modification of Kendall's tau for the case of arbitrary ties in both rankings. Journal of the American Statistical Society, Vol. 52 (1957) 33–35
2. Anick, P.: Using terminological feedback for Web search refinement: a log-based study. In Proceedings of WWW '04 (2004) 88–95
3. Eastman, C. M. and Jansen, B. J.: Coverage, relevance and ranking: The impact of query operators on Web search engine results. TOIS, Vol. 21(4) (2003) 383–411
4. Google Personal: http://labs.google.com/personalized
5. Hawking, D.: Overview of the TREC-9 Web track. In Proceedings of TREC '00 (2000) 87–102
6. Järvelin, K. and Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. In Proceedings of SIGIR '00 (2000) 41–48
7. Jeh, G. and Widom, J.: Scaling personalized Web search. In Proceedings of WWW '03 (2003) 271–279
8. Kelly, D. and Teevan, J.: Implicit feedback for inferring user preference: A bibliography. SIGIR Forum, Vol. 37(2) (2003) 18–28
9. Koenmann, J. and Belkin, N.: A case for interaction: A study of interactive information retrieval behavior and effectiveness. In Proceedings of CHI '96 (1996) 205–212
10. Kritikopoulos, A. and Sideri, M.: The Compass Filter: Search engine result personalization using Web communities. In Proceedings of ITWP '03 (2003)
11. Mizzaro, S.: Relevance: The whole history. JASIST, Vol. 48(9) (1997) 810–832
12. Nielsen, J.: Personalization is overrated. In Jakob Nielsen's Alertbox for October 4 (1998) http://www.useit.com/alertbox/981004.html
13. Schamber, L.: Relevance and information behavior. ARIST, Vol. 29 (1994) 3–48
14. Teevan, J., Alvarado, C., Ackerman, M. S. and Karger, D. R.: The perfect search engine is not enough: A study of orienteering behavior in directed search. In Proceedings of CHI '04 (2004) 415–422
15. Teevan, J., Dumais, S.T. and Horvitz, E.: Personalizing search via automated analysis of interests and activities. To appear in Proceedings of SIGIR '05 (2005)