Characterizing the Value of Personalizing Search

Jaime Teevan
Microsoft Research
Redmond, WA 98052 USA
teevan@microsoft.com

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

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

ABSTRACT

We investigate the diverse goals that people have when they issue the same query to a search engine, and the ability of current search engines to address such diversity. We quantify the potential value of personalizing search results based on this analysis. Great variance was found in the results that different individuals rated as relevant for the same query—even when the same information goal was expressed. Our analysis suggests that while search engines do a good job of ranking results to maximize global happiness, they do not do a very good job for specific individuals.

Categories and Subject Descriptors

H.3.3 [**Information storage and retrieval**]: Information search and Retrieval

General Terms

Measurement, Experimentation, Human Factors.

Keywords

Personalized search, Web search, individual differences.

1. INTRODUCTION

Users tend to issue short queries when searching, resulting in tremendous ambiguity about their informational goals. For example, a query for *IR* could imply an interest in information retrieval, but may instead refer to Iran, infrared light, or the Ingersoll-Rand Company. To investigate how query ambiguity affects result quality, we conducted a study to examine the consistency of relevance judgments assigned by different individuals to the results of the same query. Our analysis is aimed at assessing the relationship between the rank of a search result as reported by a Web search engine and the perceived relevance of the result to an individual.

2. METHODS

We conducted a study in which 15 participants evaluated the top 50 Web search results for approximately 10 queries each. Queries were selected two different ways. In one approach (*self-selected*), participants chose a query from a recently performed search, based on a diary of searches kept during the day. In the other approach (*pre-selected*), participants selected a query from a list of ten queries that were of general interest (e.g., *cancer, Bush, Web search*). The mixture of self-selected and pre-selected queries was left to the participants' discretion.

For both self-selected and pre-selected queries, participants were asked to write a detailed description of the informational goal they associated with the query before rating the results. Because the pre-selected queries were given to the participants, they had to create intents for these queries. However, since participants could select which (if any) of the pre-selected queries to evaluate, all evaluated queries were mapped to a personal goal.

For each query the top 50 results were presented in a random order. Relevance was rated on a 3-point scale (*highly relevant*; *relevant*; *not relevant*). Rather than instructing participants to select results that were "relevant to the query" in general, we asked them to indicate results that were "*personally relevant to them,*" (i.e., what they meant by the query). We then quantified the variability in information goals associated with the same query.

We collected 131 queries, of which 53 were pre-selected and 78 self-selected. The pre-selected queries enabled us 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 require different participants to coincidentally issue the same query on their own. The conclusions drawn from pre-selected queries are validated with data from the self-selected queries.

3. ARTICULATING SEARCH INTENTS

We observed a great deal of variation in participants' rating of results. One reason for the variability in ratings is that participants associated different intents with the same query. This was evident in the detailed descriptions of the information goals written by participants for their queries. For example, the descriptions of the query *cancer* ranged from "information about cancer treatments" to "information about the astronomical/ astrological sign of cancer". Such variation appeared both for the pre-selected queries, where participants had to come up with an intent based on the query, and the self-selected queries, where the query was generated to describe a pre-existing intent. Although we did not observe duplicate self-selected queries, many self-selected queries like *rice* (described as "information about rice university"), and *rancho seco date* ("date rancho seco power plant was opened") were clearly ambiguous specifications of intent.

Even when the detailed descriptions were very similar, ratings varied. This suggests that the descriptions were not at a level of detail required to distinguish different needs. For example, three people stated these similar intents for the query *Microsoft*:

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

Despite similarity of intent for these individuals, only one result (http://www.microsoft.com) was given the same rating by all. 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 descriptions of their intent was 62%. It was clearly hard for participants to accurately describe their intent, not just for pre-selected queries like *Microsoft*, but also for self-selected queries. Searches for self-selected query terms were elaborated as "information on *query term*" (*UW* → "information

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

about UW", leaving open whether they meant the University of Washington, the University of Wisconsin, or something else).

The overall inter-rater agreement for queries evaluated by more than one participant was 56%. This agreement is lower than that observed for TREC [4] and previous studies of the Web [2]. However, agreement cannot be directly compared due to variation in the number of possible ratings and the size of the result set evaluated. More importantly, the differences we observed are likely a result of our focus on understanding personal intentions.

The ratings for some queries showed more agreement than others, suggesting that some queries may be intrinsically less ambiguous. Some participants gave ratings that were similar to other's ratings, suggesting the potential for cluster people, However, even the most highly correlated individuals showed significant differences.

4. POTENTIAL FOR PERSONALIZATION

We shall now investigate how closely the Web ranking matched the best possible rankings based on the ratings we collected. We found that the Web ranking tended to be closer to the ranking that is best for the group than the ranking that is best for the individual, and that a considerable gap in list quality is created by requiring result lists to be the same for everyone.

To measure the quality of a ranking, we used *Discounted Cumulative Gain* (DCG), a well-known measure of the quality of a ranked list of results [3]. DCG measures the result set quality by counting the number of relevant results returned. It incorporates the idea that early-ranked documents are more important by weighting the value of a document's occurrence in the list inversely proportional to the log of its rank. DCG also allows us to incorporate the notion of two relevance levels by giving *highly relevant* documents a different gain value than *relevant* documents. For *relevant* results, we used a gain of one, and for *highly relevant* results, two, reflecting their relative importance.

The best possible ranking is the one with the highest total DCG. For queries where one participant evaluated 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 puts results that have the highest collective gain first.

We compared how close the best possible rankings were to the rankings that the search engine returned using 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 to the best possible ranking for each individual. On average, the individual's best ranking was slightly closer to the Web ranking (KT distance = 0.47) than to a random ranking (KT distance = 0.50). The average group ranking was closer to the Web ranking (KT distance = 0.44), and this difference was significant ($t(9) = 2.14$, $p<0.05$). It appears that Web rankings satisfy a community of intents better than they satisfy the goals of individuals.

We also found a significant difference in DCG for an individual's rankings compared to the best group ranking. Figure 1 shows the average normalized DCG for group and personalized 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 queries as well. 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 added additional people and re-computed the DCG for each individual's best rankings and for the best group ranking. As people were added, the gap between individualized rankings and the group ranking grew. On average, the best group ranking yielded a 35% improvement in normalized DCG over what the current Web ranking, while the best individual ranking led to a 66% improvement. 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. We take the gap between the individual and group personalized ranking as an indication of the potential gain that can be achieved by personalizing rankings.

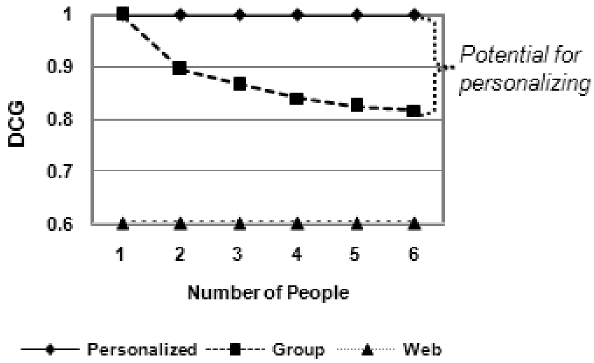

Figure 1. As more people are taken into account, the average DCG for each individual drops for the ideal group ranking, but is constant for the ideal personalized ranking.

5. CONCLUSION

Improving core search algorithms is difficult, with research typically leading to small improvements. The results of our experiment highlight the promise of providing users with better result quality through procedures that enable personalization. We hope to extend this work with a large log study in which we quantify the potential for personalization for different queries and groups of searchers.

6. 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] 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.

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

[4] 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.

[5] Teevan, J., Dumais, S. T. and Horvitz, E.: Beyond the Commons: Investigating the Value of Personalizing Web Search. Presented at *PIA '05* (2005).