Enhancing Collaborative Web Search with Personalization: Groupization, Smart Splitting, and Group Hit-Highlighting

Meredith Ringel Morris, Jaime Teevan, Steve Bush
Microsoft Research, Redmond, WA, USA
{merrie, teevan, stevebu}@microsoft.com

ABSTRACT
Collaboration on Web search is common in many domains, such as education and knowledge work; recently, HCI researchers have begun to introduce prototype collaborative search tools to support such scenarios. We analyze data from a collaborative search experiment, and based on these data we propose three techniques that can enhance the value of collaborative search tools using personalization: *groupization*, *smart splitting*, and *group hit-highlighting*.

Author Keywords
Collaborative search, web search, group search.

ACM Classification Keywords
H5.3. Information interfaces and presentation (e.g., HCI): Group and organization interfaces – *computer-supported cooperative work*.

INTRODUCTION
Web search is typically envisioned as a solitary activity, particularly since standard search tools, such as Web browsers and search engines, are not designed to support collaboration. However, previous research has shown that collaboration is an integral aspect of peoples' information retrieval practices in many domains, particularly education [1, 8, 17] and knowledge work [4, 6, 10]. For example, school children work together to find information for group homework assignments [1], and academics collaborate on literature searches for jointly-authored publications [10].

HCI (human-computer interaction) and IR (information retrieval) researchers have begun to design tools aimed at facilitating collaborative Web search [1, 2, 5, 11, 13]. These tools treat collaboration as a first-class citizen during search, providing support for activities such as group query histories, shared views of result lists, and facilities for partitioning result lists amongst group members.

We present a study of how personalization techniques can improve the group search experience. We look at collaborative Web searches issued by 10 groups, analyzing explicit search-result relevance judgments and implicit user profile information stored on participants' computers. Based on these data, we propose group-based algorithms and UI enhancements (*groupization*, *smart splitting*, and *group hit-highlighting*), which our data indicate can improve groups' ability to identify information relevant to their shared tasks.

CSCW'08, November 8–12, 2008, San Diego, California, USA.

RELATED WORK
Personalized search [3, 16] uses data such as browser history, query history, and term frequencies from locally-stored documents to improve the ranking of Web search results for a specific individual. Personalization algorithms are more effective when more data is available about the target individual [16]. Collaborative filtering techniques [3, 15] are one approach to identifying "similar" people, whose data can be combined with the target individual's in order to enhance personalization techniques.

Recently, researchers have explored the utility of exploiting group membership information as an additional implicit indicator of which users might be "similar" enough to each other to benefit from using each others' data for improved personalization of results. For example, Smyth [14] explored the similarity of query terms issued by members of "search communities" (*i.e.*, groups of people who use a special-interest Web portal). Mei and Church [9] investigated whether geographic location (as approximated by IP address) serves as a metric of user similarity.

In this paper, rather than studying the similarity of users who are implicitly related through use of a Web portal or through geography, we study explicitly formed groups of users, whose grouping is based on a shared task that they wish to accomplish through collaborative Web search.

METHODOLOGY
We conducted a study to gather data that would enable us to explore properties of group Web searching and to test our proposed algorithms and enhancements to the experience. Because we are interested in identifying general-purpose algorithms and UI enhancements for collaborative Web search, rather than enhancements specific to a particular collaborative Web search system, we designed a system-agnostic methodology for gathering data. While this methodology enables us to gather group members' query and relevance judgment data in the absence of widespread availability of collaborative search tools, it does not allow us to observe the impact of our proposed techniques on group dynamics and collaborative work styles.

We recruited 30 participants (19 male, 11 female) from within Microsoft. Participants came from a variety of occupational backgrounds, including sales/marketing, software development, and research. All participants rated themselves as having average (13/30) or above average (17/30) skill at Web search. Participants' ages ranged from 20 to 59 years.

Participants volunteered for the study in groups of three (for a total of 10 groups). Groups consisted of three colleagues who had a shared, work-related task that they hoped to accomplish via Web search. Group members had known each other an average of 3.3 years.

Each group provided a brief, 1 to 3 sentence statement summarizing their shared task. Example tasks included:

- "What is the state of the art of research combining speech interfaces and tabletops? This includes any prototypes or demos as well as studies conducted."
- "What techniques have been used to created semantic thumbnail representations of a web page?"
- "[We need] to search for information about companies offering learning services to corporate customers."

After receiving each group's task statement, we sent an e-mail questionnaire to each participant, which they were instructed to complete individually. The questionnaire reminded each participant of the group's chosen task, and then asked for a list of six queries the participant might type into a Web search engine in order to get information relevant to that task. Participants were instructed not to actually type the queries into a search engine, but rather to simply list them in the email. We then assembled a set of six unique queries specific to each group by using the first two non-overlapping queries each of the three group members had selected that returned more than 50 results.

For each of the six group queries, we asked participants to provide relevance judgments for a set of 21 results. We chose to collect judgments for 21 results because piloting revealed higher numbers led to significant fatigue, and because it is divisible by the group size, three, which is important for smart splitting, discussed later. The 21 results were selected by downloading and caching the odd numbered results from the top 42 search results returned by a major search engine. We used the odd-numbered half of the first 42 results, rather than simply using the top 21 results, in order to include results with varying relevance.

Participants judged relevance individually, using their own computers in their offices. Our study software presented participants with their list of six task-related queries. Clicking any query opened a web page showing the 21 previously cached search results for that query in random order. Each result consisted of the standard title + snippet + url trio. Participants were instructed to judge whether or not they personally found each search result relevant to the group's stated task. Next to each result were buttons marked "very relevant", "relevant", and "not relevant", that participants used to provide their judgments. Participants could also click on any search result in order to open the webpage it pointed to in a separate browser window, if they felt that viewing the page would assist them in making a relevance judgment.

For each search result, participants' relevance judgments were recorded by our software. Additionally, our software recorded user profile information necessary for personalization, such as the frequency with which each word that appeared in a search result appeared within documents on the participant's computer, and whether a url or domain associated with a search result was already stored in the participant's web history or bookmarks list.

RESULTS

Using the data collected, we explored how personalization techniques can improve the group search experience. In this section, we discuss three group personalization techniques: *groupization*, where group members' data is used to rank an individual's search results; *smart splitting*, where results are distributed to group members according to which members are most likely to find each relevant; and *group hit-highlighting*, where group members' queries are used to draw visual attention to potentially relevant results.

Groupization

We hypothesized that collaborative search tools could take advantage of a group's commonalities to further enhance previously-proposed personalization techniques (*e.g.*, [3, 16]) through a process we call *groupization*. The motivation behind groupization is similar to the motivation for extending standard recommender systems into group recommenders [12].

Groupization augments the personalization process by giving higher weights to pages that are relevant to more members of the group, based on matching against each group member's web history and local document term frequencies. To perform groupization on a set of search results, we first calculate a personalization score for each search result for each member of the group, using the process defined by Teevan *et al* [16]. For each search result, the groupization score is computed as the sum of the personalization scores of each group member. We then take a weighted combination of the groupization score and the search result's original rank so as to preserve important information used by Web search engines, such as the result's "authoritativeness."

For each participant in our study, we computed the normalized Discounted Cumulative Gain (DCG) [7] of the

Table 1. Normalized DCG for different result orderings.

Ordering type	Mean	Std. Dev.
Web	0.57	0.08
Personalization	0.65	0.05
Groupization	0.67	0.05

groupized search result lists for each of the six queries generated by that participant's group. DCG is a common IR metric that represents how good a particular ordering of search results is by comparing the order of the results to the participant's explicit relevance judgments for each result. An ordering in which results that were scored as highly relevant appear before results scored as not relevant would have a higher DCG than an ordering in which results scored as not relevant appear near the top. To facilitate cross-query comparison, the value is normalized to be between 0 (for the worst possible ranking) and 1 (the best possible).

For each participant, for each of the six group queries, we computed the quality of three different rankings: (1) the original ranking returned by the Web search engine, (2) the ranking found using a personalization algorithm [16], and (3) the ranking found using our groupization algorithm to combine personalization data from the participant's group.

As can be seen in Table 1, while personalization improved on the original Web ranking, the use of group data in addition to an individual's led to a greater improvement. ANOVA results show a significant difference among all three[1] (Wilks' Λ=0.35, $F(2, 7)$=6.61, p=0.02). Follow-up t-tests show all pairwise differences are significant, with personalization improving on the original web ordering ($t(8)$=3.60, p<0.01), and groupization improving on both the original web ordering ($t(8)$=3.88, p<0.01) and personalization ($t(8)$=2.50, p<0.04).

Smart Splitting

Based on reported obstacles to collaborative search using status quo tools [10], Morris and Horvitz [11] proposed that collaborative search tools should support *division of labor*. The SearchTogether system [11] introduced the concept of "split searching" as one mechanism by which a collaborative search tool could accomplish this goal. A split search is one in which one member of the group enters a query term, which is then sent to a search engine. The top results for this query are then divided up round-robin style amongst all of the group members, such that each person is given a non-overlapping portion of the results. Split searching can be used to allow group members to evaluate sets of results efficiently, without redundancy.

We hypothesized that the naïve splitting algorithm proposed by Morris and Horvitz could be improved by using personalization techniques to accomplish *smart splitting*. To accomplish smart splitting, we generate a personalized score [16] for each result of each participant's six group queries. The top Web results for a query are still distributed round-robin style, but rather than using the Web ranking to decide which result to distribute next, each participant receives the result that is most personally relevant, as evidenced by the result's personalized score.

[1] Group 3's data was removed from this analysis as an outlier, since the group's mean DCG scores were more than two standard deviations away from the overall means.

Table 2. Normalized DCG for different splitting algorithms.

Split Method	**Mean (group queries)**	**Std. Dev. (group queries)**	**Mean (group + control queries)**	**Std. Dev. (group + control queries)**
round robin	0.64	0.33	0.63	0.33
random	0.65	0.32	0.62	0.32
smart	0.68	0.31	0.66	0.30
ideal	0.71	0.30	0.68	0.30

Using group members' relevance judgments for each query generated by their group, we calculated the normalized DCG of results lists generated by naïve, round-robin splitting, as well as by an alternative naïve splitting method of randomly dividing results among participants. We also calculated DCG of the results lists generated by our smart-splitting algorithm, as well as calculating an ideal split where results are distributed round robin style using the participants' explicit relevance judgments. The trend of the resulting mean DCG scores (Table 2, "group queries" scores) shows smart splitting performing better than the two naïve methods, although not quite reaching ideal DCG. ANOVA results indicate a significant trend (Wilks' Λ=0.95, $F(3, 158)$=2.87, p=0.04), with pairwise follow-up tests indicating that the round robin ($t(160)$=2.65, p<0.01) and random ($t(160)$=1.99, p<0.05) algorithms resulted in significantly lower DCG than the ideal split, while the smart split and ideal split did not perform significantly differently from each other.

In addition to the six group-generated queries, all participants also evaluated the relevance of results for a set of nine control queries. Extending our analysis of the splitting algorithms to utilize both the group-generated queries and the control queries provides more statistical power, re-confirming with additional confidence the trend that smart splitting performs closer to the ideal splitting method than either naïve technique (Table 2, "group + control queries" scores): Wilks' Λ=0.96, $F(2=3, 369)$=5.73, p<0.001. Follow-up t-tests show that smart splitting resulted in significantly higher DCG than either the round robin ($t(371)$=2.21), p=0.03) or random ($t(371=2.51)$, p=0.04) splitting methods.

These findings suggest that personalization techniques can be used to personalize users' results lists *relative to* those of collaborators, by capitalizing upon the differences among group members' local content in order to determine areas of expertise. In a sense, *smart splitting* is the inverse of *groupization*, with the former capitalizing on the differences amongst group members in order to facilitate division of labor, and the latter amplifying the similarities among group members in order to produce an appropriately-ordered shared view of results.

By dividing labor among group members, smart splitting aims to focus each group member's attention on the results most similar to his/her expertise. This stands in contrast

Participant's query: snp disease data
Other group queries: genome-wide association; snp classification; synonymous snp; snps Wikipedia; computational analysis of snps data

Evaluating the Association of Mitochondrial **SNP** ...
An analysis of the Human_MitBASE **data** helped in the prediction of association between **SNP** haplotypes with **disease** phenotypes. A novel computational tool E-MIDAS was developed ...
http://csdl2.computer.org/persagen/

Figure 1. An example of group hit-highlighting. Terms from a participant's query are shown in bold, and terms from the group's other queries are underlined.

recommender techniques such as topic diversification [18], which aim to present a user with a diverse, rather than focused, set of items. Understanding the impact of this choice on group dynamics is left to future work.

Group Hit-Highlighting

Hit-highlighting is a technique used by most major search engines to help users understand how relevant a result is to their information need. Instances of the user's keywords that appear within the title, snippet, or url of each search result in the results list are emphasized (*e.g.*, bolded or colored distinctly). If collaborative search tools have access to a task-oriented, group query history (as proposed by Morris *et al.* [1, 11]), the system could perform *group hit-highlighting*, whereby all group members' keywords that appear within a search result are emphasized. Such a system could use distinct formatting to distinguish terms from the active query and the group's past searches (*e.g.*, Figure 1).

To understand whether group hit-highlighting could help our participants better identify relevant results, we looked at whether the result snippets for a group member's query that contained many terms from the other group queries were more likely to have been judged relevant. The value of the attention drawn to a snippet by the terms from the other five group queries was judged by the number of times those queries' terms appeared within the result. Since rank also affects how salient a result is, we also factored in ranking information to our attention-based score.

To evaluate the benefit of the group hit-highlighting, we calculated the normalized DCG of the result list based on this attention score rather than the ranking, so that the relevant results that draw the most attention contribute most highly to the DCG. We found the group hit-highlighting score yielded significant improvements in DCG (M=0.62, SD=0.26, $t(378)=5.02$, $p<0.001$) compared to the original ranking (M=0.59, SD=0.27). The DCG using the group hit-highlighting score also offered significant improvements over the DCG based on using only the participant's current query terms to generate a hit-highlighting score (M=0.60, SD=0.27, $t(378)=3.70$, $p<0.001$).

These findings suggest that employing group hit-highlighting in a collaborative search tool could help draw users' attention to highly-relevant results, and that using words from a collaborative system's group query history could be valuable for re-ranking search results.

CONCLUSION

We introduced three techniques that can enhance collaborative search tools, and demonstrated their value using empirical data. *Groupization* is a way to re-rank results in an order most relevant to group members. *Smart splitting* allows division of labor based on individual group members' specialized knowledge. *Group hit-highlighting* visually indicates particularly relevant results. Instantiating these concepts within a collaborative search tool is an important next step for verifying the utility and usability of these innovations, particularly for understanding their impact on group dynamics and collaboration strategies.

REFERENCES

1. Amershi, S. and Morris, M.R. CoSearch: A System for Co-located Collaborative Web Search. *CHI 2008*, 1647-1656.
2. Diamadis, E.T. and Polyzos, G.C. (2004). Efficient Cooperative Searching on the Web: System Design and Evaluation. *International Journal of Human-Computer Studies*, 61(5).
3. Dou, Z., Song, R., and Wen, J.R. A Large-Scale Evaluation and Analysis of Personalized Search Strategies. *WWW 2007*.
4. Fidel, R., Bruce, H., Pejterson, A., Dumais, S., Grudin, J., and Poltrock, S. (2000). Collaborative Information Retrieval. *New Review of Information Behavior Research*, 1(1): 235-247.
5. Freyne, J. and Smyth, B. Cooperating Search Communities. *Adaptive Hypermedia and Adaptive Web-Based Systems 2006*.
6. Hansen, P. and Järvelin, K. (2005). Collaborative Information Retrieval in an Information-Intensive Domain. *Information Processing and Management*, 41(5): 1101-1119.
7. Järvelin, K. and Kekäläinen, J. IR evaluation methods for retrieving highly relevant documents. *SIGIR 2000*, 41-48.
8. Large, A., Beheshti, J., and Rahman, T. (2002). Gender Differences in Collaborative Web Searching Behavior: An Elementary School Study. *Information Processing and Management*, 38: 427-433.
9. Mei, Q. and Church, K. Entropy of Search Logs: How Hard is Search? With Personalization? With Backoff? *WSDM 2008*.
10. Morris, M.R. A Survey of Collaborative Web Search Practices. *CHI 2008*, 1657-1660.
11. Morris, M.R. and Horvitz, E. SearchTogether: An Interface for Collaborative Web Search. *UIST 2007*, 3-12.
12. O'Conner, M., Cosley, D., Konstan, J., and Riedl, J. PolyLens: A Recommender System for Groups of Users. *ECSCW 2001*.
13. Pickens, J., Golovchinsky, G., Shah, C., Qvarfordt, P., and Back, M. Algorithmic Mediation for Collaborative Exploratory Search. *SIGIR 2008*, 315-322.
14. Smyth, B. (2007). A Community-Based Approach to Personalizing Web Search. *IEEE Computer*, 40(8): 42-50.
15. Sugiyama, K., Hatano, K., and Yoshikawa, M. Adaptive Web Search Based on User Profile Constructed Without Any Effort from Users. *WWW 2004*, 675-684.
16. Teevan, J., Dumais, S., and Horvitz, E. Personalizing Search via Automated Analysis of Interests and Activities. *SIGIR 2005*.
17. Twidale, M., Nichols, D., and Paice, C. (1997). Browsing is a Collaborative Process. *Information Processing and Management*, 33(6): 761-783.
18. Ziegler, C., McNee, S., Konstan, J., and Lausen, G. Improving Recommendation Lists Through Topic Diversification. *WWW 2005*, 22-32.