Discovering and Using Groups to Improve Personalized Search

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

Meredith Ringel Morris
Microsoft Research
Redmond, WA 98052 USA
merrie@microsoft.com

Steve Bush
Microsoft Research
Redmond, WA 98052 USA
stevebu@microsoft.com

ABSTRACT
Personalized Web search takes advantage of information about an individual to identify the most relevant results for that person. A challenge for personalization lies in collecting user profiles that are rich enough to do this successfully. One way an individual's profile can be augmented is by using data from other people. To better understand whether groups of people can be used to benefit personalized search, we explore the similarity of query selection, desktop information, and explicit relevance judgments across people grouped in different ways. The groupings we explore fall along two dimensions: the longevity of the group members' relationship, and how explicitly the group is formed. We find that some groupings provide valuable insight into what members consider relevant to queries related to the group focus, but that it can be difficult to identify valuable groups implicitly. Building on these findings, we explore an algorithm to "*group*ize" (versus "*personal*ize") Web search results that leads to a significant improvement in result ranking on group-relevant queries.

Categories and Subject Descriptors
H.3.3 [**Information storage and retrieval**]: Information Search and Retrieval – *query formulation*; H.5.3 [**Information interfaces and presentation**]: Group and Organization Interfaces – *computer supported cooperative work.*

General Terms
Algorithms, Measurement, Experimentation, Human Factors.

Keywords
Personalization, collaborative filtering, collaborative search.

1. INTRODUCTION
Web search personalization algorithms improve the Web search experience by using an individual's data (e.g., topical categories marked interesting, query history, or term vectors of previously viewed content) to identify the results that are the most relevant to that individual. This can be done in several ways. For example, a searcher's query can be modified to reflect a particular interest, or results may be re-ranked so that personally relevant results appear higher in the list [15]. Previous research suggests personalization algorithms perform best when there is a large amount of data available about an individual [22]. For this reason, we propose combining an individual's data with that of other related people to enhance the performance of personalized search. We call the use of group information for personalization "*group*ization."

One challenge in the use of group data for personalization lies in the identification of related groups of people. To develop an understanding of what factors are important for building groups for groupization, we conducted two studies of a total of 140 people. The data we collected enabled us to understand whether people grouped by various properties were similar in the queries they selected, the information they had on their desktop, or the relevance judgments they assigned to search results. We explored groupings that varied based on the longevity of the relationship and on whether the group was formed explicitly or implicitly. Specific grouping criteria included task, interests, demographics, geographic location, occupation, work group, query selection, and the content on their desktop computers. By correlating group membership with the similarities of the group members' explicit relevance judgments, we are able to understand what types of groups are most likely to receive value from groupization.

It appears that some attributes are more useful than others for identifying people who find the same results relevant, and, in particular, that group membership provides information about what members consider relevant to group-related queries. Using the data we collected to understand group properties, we explore combining information about group members to produce a groupized (versus *personal*ized) result list. We find that it is possible to aggregate personalization scores from different group members to create a groupized result list that is of higher quality than each individual's personalized list. Consistent with the understanding we develop of the different attributes, groupization appears most useful for queries related to the group.

We begin the paper with a discussion of related work in the areas of personalization, collaborative filtering, and collaborative Web search. We then describe our data collection methodology. By analyzing the collected data, we explore the within-group variation of relevance judgments, query selection, and user profile information. We then describe a groupization algorithm that extends personalization techniques to include group data. We analyze the value of groupization for the groups represented in our study, and conclude with a discussion of practical issues, including techniques for identifying groups and group-related queries outside of experimental settings.

2. RELATED WORK
Research on personalizing search results [4, 16, 22] has found that implicitly gathered information such as browser history, query history, and desktop information, can be used to improve the

WSDM'09, February 9–12, 2009, Barcelona, Spain.

ranking of search results on a per-user basis. Teevan et al. [22] found that the performance of the personalization algorithm they studied improved as more data became available about the target user. This finding suggests that additional data from similar people may be useful in enhancing personalization systems.

Collaborative filtering is one way that data from similar people is identified for use in improving the search experience for an individual. As an example, Sugiyama et al. [20] filled in sparse user term-weight profiles by applying collaborative filtering techniques to provide term weights based on those of users with similar profiles. Sun et al.'s CubeSVD approach [21] used click-through data (represented as a user+query+URL triple) to generate personalized Web rankings; they used collaborative filtering techniques to generate missing click-through triples, thereby enhancing their technique's performance. Dou et al. [3] compared several personalization strategies and found that the use of click-through data and *k*-nearest neighbor collaborative filtering was a promising approach. Almeida and Almeida [1] used Bayesian algorithms to cluster users of an online bookstore's search service into communities based on links clicked within the site and found that the popularity of different links within different communities could be used to customize search result rankings. VisSearch [8] uses data mining to uncover patterns in users' queries and browsing in order to generate recommendations for users with similar queries. Some recommender systems, such as the movie recommender system PolyLens [13], attempt to generate recommendation lists for groups of users.

Smyth [17] suggested that click-through data from users in the same "search community" (e.g., a group of people who use a special-interest Web portal or who work at the same company) could enhance search result lists. Smyth provided evidence for the existence of search communities by showing that a group of employees from a single company had a higher query similarity threshold than general Web users. Freyne and Smyth's I-SPY system [5] expanded the notion of search communities to include related communities, measuring intercommunity similarity based on the degree to which communities' queries and result click through overlap. Mei and Church [9] found that geographic location might serve as a reasonable proxy for community, since they observed that grouping users into classes based on the similarity of their IP addresses could improve search results.

As with the above studies, in this paper we investigate enhancements to personalization techniques that supplement data about a target user with other people's data. We use explicit relevance judgments, user profile information, and query selection to explore group similarity. We then explore how group similarity impacts the performance of personalization algorithms that use group data; we compare the value of group personalization for several relationship classes, including general demographics, geography, occupation, interest group membership, online behavior, and shared task.

Although many of the aforementioned personalization techniques (e.g., collaborative filtering and recommender systems) are "collaborative" in the sense that many users' data is combined to produce a final result, such techniques represent a passive form of collaboration; in contrast, several researchers have studied collaborative searching, a more active process wherein a group of users actively work together on a shared search task. Collaborative searching has been reported amongst students [24] and professionals [11] for tasks like travel planning, online shopping, and researching business or school-related topics.

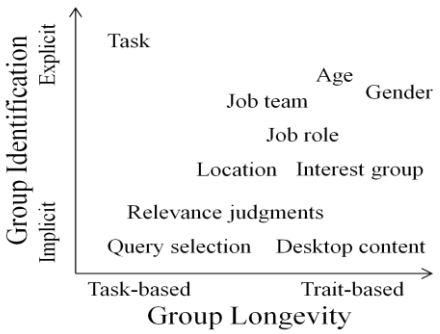

Figure 1. Groups studied in this paper, broken down by group longevity and how the group was identified.

Recently, several new systems have been developed to assist with collaborative searches. For example, SearchTogether [12] and Cerchiamo [14] support real-time collaboration on search tasks amongst a group of users who each have their own computer. CoSearch [2] supports co-located collaboration by allowing users to type queries on their mobile phones and send them to a shared computer to be queued with other group members' queries. Our findings suggest that task-based communities, as defined through participation in a collaborative search session, could provide a valuable source of data for enhancing personalization algorithms.

3. GROUP TYPES

In this paper we study people grouped along two axes. The first axis relates to the longevity of group membership: groupings can be relatively short term (*task-based*), or last longer (*trait-based*). The second axis relates to how group membership is determined: either by information provided by group members (*explicit*) or inferred from member activity (*implicit*). Figure 1 illustrates how the groups we studied fell along these axes.

3.1 Group Longevity

The most short-term grouping we studied is of people with a shared goal. Since group members work together to accomplish a shared task we refer to these groups as *task-based* groups. Common tasks that may motivate groups to collaborate on Web search include travel planning, shopping, work- or school-related projects and reports, social planning, or medical searches [11].

Longer-term groups can be comprised of people who are related through shared traits or long-term interests. We refer to these groups as *trait-based* groups. Members may not be consciously collaborating on the same task, but may be highly likely to repeat or augment tasks already accomplished by other group members, have interests in the same queries and results as other group members, or possess information relevant to another group member's task. We explore a variety of trait-based groups built from shared interests, occupations, geography, or demographics.

Interest groups' members share an interest in a particular topic. We used e-mail distribution lists as one means of studying interest groups (i.e., users who have subscribed to email discussion groups on topics such as photography). Another approach we used to group people by interest is to compare the similarity of the content on an individual's computer to that on others' computers.

Occupational groups are comprised of people with related jobs. We explore two main classes of occupational groups: *Job-role* groups consist of people with similar job titles who may work on different products or even in different companies (e.g., a group of

software engineers or a group of pediatricians). *Job-team* groups consist of people who work on the same product or for the same company; such groups may consist of people with heterogeneous job roles (e.g., the team of people that works on Microsoft Word, including engineers, managers, and marketing specialists).

Geographic groups are comprised of people who live or work in a particular region. The relationship can be based on a city, county, state, etc., and may be hierarchical. Demographic groups are made of people who share characteristics such as gender or age.

3.2 Group Identification

In addition to group longevity, another axis we explored is how groups are identified. One way group membership can be determined is by information provided directly from the members. We consider such groups to be *explicit*. For example, an explicit task-based group is one where group members are overtly collaborating on a specific task. Group membership can also be inferred. We call these groups *implicit* groups. An implicit task-based group may be formed from people who appear, based on their actions, to be conducting the same task.

We explore several methods of group identification with explicit data, including explicit task-based collaboration, and self-reported gender, age, geographic location, job-role, and job-team. We also explore several methods of implicit group identification, including grouping people with similar desktop indices, who issue similar queries, and with similar relevance judgments. Some methods fall in between. For example, we identify interest-based groups based on mailing list membership; the membership is explicit, but the inference that members share an interest is implicit.

4. METHODS

To understand group similarity and groupization, we collected two different data sets consisting of user profile information and explicit relevance judgments. Both data sets include implicit and explicit properties about each participant that can be used to determine group membership, as well as each participant's explicit relevance judgments for a number of overlapping queries.

4.1 Dataset 1: Trait-Based Groups

In the first dataset, we gathered information about 110 participants by targeting people with shared long-term traits who worked at Microsoft. We studied several types of trait-based groups at the explicit end of the group-identification continuum (demographic, geographic, occupational, and interest-based), as well as more implicitly-defined groups (behavior- and profile-based).

People were assigned to demographic groups based on gender and age. Each participant was assigned to either of the groups *male* or *female*, and to one of the groups *twenties*, *thirties*, and *over forty* depending on their self-reported gender and age data.

Geographic groups were assigned using residential zip code data provided by participants. As all participants worked for Microsoft, most lived in Seattle or the surrounding area; we used this data to assign participants to the groups *Seattle* and *Seattle's suburbs*. More distant geographic groupings may reveal larger differences.

Occupational groups were of two types: job-role and job-team. Job-role groups were formed on the basis of participants' self-reported job title. The most common job roles were *developer* (technical job roles primarily involving software development or debugging), *program manager* (job roles involving defining a vision for products and facilitating communication between developers, marketing, and customers), or *researcher* (job roles not tied to specific products but rather to conducting academic research). Job-team groups were formed on the basis of membership on mailing lists for specific teams within Microsoft (different teams work on different products). The three job-team groups with reasonable participation levels in our experiment were *Product group I*, *Product group II*, and a research group, since the experiment was advertised to these teams' e-mail lists.

Table 1. Number of Dataset 1 participants comprising the membership of each explicit group.

Group Type		Group Name	# People	# Queries
Demographics	Gender	Male	80	462
		Female	30	174
	Age	20-30	38	221
		30-40	45	253
		40+	12	83
	Location	Seattle	29	175
		Seattle's suburbs	73	420
Interest	Mailing list	Pets	16	99
		Photography	21	123
		Vegetarianism	34	196
Job function	Team	Product group I	16	131
		Product group II	11	64
		Research group	14	79
	Job role	Developer	54	321
		Program Manager	19	117
		Researcher	6	45
		Total	110	624

Interest groups were defined by membership in opt-in e-mail distribution lists within Microsoft. The experiment was advertised on three of these e-mail lists, hence our interest groups are based on those three lists' topics: *pets*, *photography*, and *vegetarianism.*

The number of people represented in each of these groups can be found in Table 1. The number of people in each group ranged from six (*researcher*) to 80 (*male*).

We also explored three methods of implicit grouping: grouping participants with similar desktop indices, grouping participants who chose to evaluate similar queries in our experiment, and grouping participants with similar relevance judgments for the set of search results in our experiment.

Note that many of the groups studied are not mutually exclusive, and membership in a group of one category is not necessarily independent from membership in another category. For example, in our participant pool we found there were high correlations between some of the demographic, geographic, occupational, and interest group memberships. Job role, for example, was correlated with gender: in our sample the developers tended to be male (r=0.30) while program managers tended to be female (r=0.10). Gender also correlated with interest group membership: photo enthusiasts were mostly male (r=0.25) while pet lovers tended to be female (r=0.51). Age, geography, and interests were sometimes related in complex ways: people who lived in Seattle were more likely to be in their twenties (r=0.10) and to be vegetarians (r=0.12), while pet lovers tended to be in their fifties (r=0.12) and live in the suburbs (r=0.13).

Table 2. Trait-based relevance judgments were collected by asking participants to select three queries from each of two query categories (social and work).

Category		Query	# Queries
Social	Pets	cat on computer	42
		toilet train dog	56
	Photography	black & white photography	35
		slr digital camera	66
	Vegetarianism	bread recipes	59
		Redmond restaurant	81
Work	Product related	photosynth	53
		live meeting	51
	Management	business intelligence	30
		Microsoft new technologies	54
	Programming	c# delegates	50
		powershell tutorial	47

4.2 Dataset 2: Task-Based Groups

The second dataset consists of 30 participants, all Microsoft employees. Participants were recruited in task-based groups of three people each, for a total of ten groups. Group members all knew each other prior to the data collection, and had work-related tasks that resulted in a shared information need that they wished to address through Web search. Each group provided a brief description of its shared task, such as "Find the economic pros and cons of telecommuting" (Table 3, Group 1) or "Search for information about companies offering learning services to corporate customers" (Table 3, Group 2). Participants also provided demographic information, such as age and gender.

4.3 Data Collection

In order to understand the variation in what different people considered relevant to the same query, we needed to collect relevance judgments for overlapping queries. For this reason, our study software asked participants to evaluate queries from a pre-populated list (query selection is described in Section 4.4).

Selecting a query displayed a list of pre-cached search results (40 for Dataset 1, and 20 for Dataset 2 due to the larger number of queries participants needed to judge in the latter case). Results were ordered randomly so that participants' judgments were not biased by rank. All results were displayed in standard search result format, with title, snippet, and URL. Clicking on a result opened the target page in a new browser window. Next to each result were buttons that allowed the participant to mark whether they deemed that result to be *highly relevant*, *relevant*, or *not relevant* to the current query. In the background, we also collected information about how well each result fared according to various personalization metrics [22], such as whether the target URL had been previously visited by the participant and how frequently the terms in the result appeared in the participant's desktop index.

The explicit relevance judgments we gathered differed from those typically gathered as part of the TREC benchmark collections used in the evaluation of information retrieval systems [25]. In TREC, expert judges are asked to judge a result's relevance based on a detailed description of an information need. This scenario is unrealistic for Web search, where people issue very short queries to describe their information needs [19]. The same short Web query issued by two different people is unlikely to have the same unambiguous information goal behind it. In this paper our focus is on which results are personally relevant to different group members. For this reason, rather than have people evaluate results for fully defined information goals, we encouraged participants to judge what they *personally would consider relevant*.

Table 3. Task-based judgments were collected for the task-based queries generated by the participant's group, as well as for one query from every other group. The queries shown are those that members of Group 1 were asked to evaluate.

Group	Query	# Queries In	# Queries Out
1	economic comparison telecommuting versus office	3	23
	Pros and cons of telecommuting	3	0
	Pros and cons of working in an office	3	0
	social comparison telecommuting versus office	3	0
	telecommuting	3	0
	working at home cost benefit	3	0
2	cisco university services	3	21
3	Diablo ii character guide	2	23
4	apple spotlight file indexing	3	23
5	group Instant Messaging	3	21
6	NewPage Corporation	2	22
7	snp disease data	3	21
8	Social Computing	2	22
9	tabletop computing studies	3	21
10	thumbnail information retrieval	3	21

All participants ran the study software on their own computer, enabling us to collect user profile information relating to each query, in addition to the group membership information and explicit relevance judgments.

4.4 Query Selection

4.4.1 Trait-Based Dataset

Participants in the trait-based dataset were asked to select six queries to evaluate from a list of 12 pre-generated queries (see Table 2). So as to be of general interest to participants, the queries were chosen via examination of logs of queries issued from within Microsoft. Six queries related to business functions were chosen, and six to social topics. Our study design asked subjects to select three queries from the set of work-related queries and three from the set of social queries. Queries were not labeled as "work-related" or "social" in the rating interface used by participants.

Although the queries were intended to be interesting, they were not created by the participants. Instead, participants had to create an intent for each query evaluated. By allowing people to decide whether they wanted to evaluate a particular query, we sought to allow them to only work with queries and associated results that were meaningful. Personally relevant explicit judgments for the same queries would have been difficult to collect using only self-generated queries, since it would require different participants to coincidentally issue the same query on their own.

While allowing participants to select queries to evaluate meant the queries were more personally relevant than they might have been otherwise, selection presents a challenge for data analysis. The resulting query/user matrix is sparse and values cannot be directly

compared across cells. To calculate whether group differences are significant we either compare the average values for each participant, or do a pair-wise comparison using the average values for each of the twelve queries. This results in some loss of power.

In total, for the trait-based dataset we collected in total judgments from 110 people for 624 queries. The number of people who evaluated the same set of results for each of the 12 queries ranged from 30 to 81. The number of people from a particular group who evaluated the same set of query results was often much smaller.

4.4.2 Task-Based Data

To select the queries for the task-based dataset, each participant individually completed an email questionnaire. The questionnaire reminded the participant of the group's chosen task and asked for six queries the participant might submit to a search engine to find relevant information. We then chose the first two unique queries from each of the three group members' lists to create a set of six distinct queries for the group members to evaluate.

Group members were also asked to evaluate results for queries selected from each of the other task-based groups. The queries evaluated by all participants were selected by asking three independent judges to rank each group's six queries according to how meaningful the queries were to them and selecting from each group the one with the highest common rank. An example of the queries evaluated by members of Group 1 can be seen in Table 3. In total we collected judgments for 380 queries for Dataset 2.

5. VARIATION WITHIN GROUPS

We analyzed the data we collected to explore how similar the members of groups of different types were in the queries selected to evaluate, their user profiles, and their relevance judgments.

5.1 Variation in Query Selection

Query selection provides information about the topics a person is interested in. We characterized individuals in the trait-based dataset based on which six of the 12 queries they chose to evaluate. We were unable to perform similar analysis for the task-based dataset because participants did not have any choice in the queries they evaluated. In general people with similar traits appear to make similar query choices, with group members choosing queries related to their personal and/or professional interests. We found that participants' choice of work-related queries was influenced by their occupation groups, while their choice of social queries was influenced by membership in special-interest groups.

Our analysis of query selection was done by representing the queries each individual selected as a vector. We then computed Pearson's correlation coefficient for the work-related and social query-choice vectors for each pair of participants in the same group, and averaged across pairs to produce a mean correlation.

When considering all of the trait-based data collected, the mean correlation in choice of work-related queries was 0.03 (no correlation). Interest-related groups also showed no correlation for work-related query choices, with vegetarians having a mean correlation of 0.01, pet lovers 0.02, and photo enthusiasts 0.03. However, occupational groups had significantly higher mean correlations among their members' work-related query choices, as measured by independent-samples *t*-tests comparing each group's mean correlation to the mean for the group of all participants. Job role was one type of occupational grouping we examined: developers had a mean correlation of 0.13, which was significantly higher than the participant pool as a whole (p<0.01), and program managers had a mean correlation of 0.14, which was marginally higher than the entire subject pool (p=0.07*)*. Job team groups also chose similar work-related queries: participants who worked in Product group I had a mean work-related query choice correlation of 0.21, which was significantly higher than the pool at large (p<0.01), and those who worked in the research group had a mean correlation of 0.29 (p<0.05).

The mean correlation in choice of social queries for the entire trait-based dataset was 0.06 (again, no correlation). Occupational groups also showed no correlation regarding social query choices, with developers having a mean correlation of 0.06, program managers -0.01, members of Product group I 0.18, and members of the research group 0.16, none of which were significantly different from the mean of the entire subject pool. In contrast, pet lovers had a mean correlation of 0.37 for their social query choices, which was significantly higher than the pool as a whole (p<0.01) and photographers had a mean correlation of 0.25, also higher than the pool as a whole (p<0.01). In contrast, vegetarians had a mean correlation of 0.06, which was not significantly different than the pool of all participants; we suspect this is because the queries in the social set did not hold any options that would appeal more to vegetarians than to non-vegetarians – the two food-related queries, "redmond restaurant" and "bread recipes", were not specifically vegetarian-oriented.

In natural interactions with search engines users do not select queries from a list of pre-selected choices, but rather generate queries based on their current information need. However, our data provides some support for the hypothesis that users with similar interests are more likely to generate queries on similar topics (related to those interests) than the population at large.

The relationship between group membership and query choice suggests query choice might be a good method for determining group membership when membership is not known a priori; in particular, two people with a history of querying on a topic may share an occupational or social interest group membership and benefit from group personalization when issuing queries on that topic. We explored using query selection to create implicit groups by identifying sets of people who selected four queries in common. We compared the average inter-rater reliability over those four queries for the implicit group compared with the entire study population. People who selected similar queries did not appear to have highly correlated relevance judgments. We also clustered people based on their queries (using *k*-means and agglomerative clustering algorithms for two to six clusters), and compared the inter-rater reliability of cluster members with all participants. Again, we found no significant difference. A longer query history may be necessary to identify query-based groups.

5.2 Variation in User Profile

We also characterized group members' variation by exploring the similarity of members' user profiles. To do this we used a rich, implicitly constructed user profile that consisted of the content stored on the different group members' desktop computers. The hypothesis is that if group members have similar standing interests, this will be reflected in the emails they receive, the Web pages they visit, and the content they author – all of which is available in an index of the person's desktop computer. If our hypothesis is true, there is potential to mine relevant information from other members of a group in the support of an individual.

To understand how user profiles differed within and across groups, we created term vectors for each participant in our study representing how many times a term occurred in their profile (i.e.,

Table 4. The similarity of user profiles for interest-based groups, work groups, and task-based groups. There is a larger difference in group similarity for group-related queries.

Group		Queries	
		All	***Group***
Pets	In	0.52	0.40
	Out	0.47	0.31
	Diff.	10%	26%
Photography	In	0.56	0.64
	Out	0.46	0.35
	Diff.	23%	82%
Vegetarianism	In	0.49	0.65
	Out	0.48	0.41
	Diff.	1%	58%
Work groups	In	0.52	0.60
	Out	0.47	0.54
	Diff.	9%	11%
Task groups	In	0.42	0.77
	Out	0.31	0.35
	Diff.	34%	120%

in their desktop index). Each term in the vector is given a BM25 weight [18]. To protect our participants' privacy, rather than collect a person's entire desktop index we gathered term counts only as pertained to the queries they evaluated. Each item in the vector represents that term's co-occurrence with the query term. As a result, we can only compare profiles across individuals for queries that different individuals issued in common. To measure how close any two individuals' profiles were for a query, we computed the cosine similarity between their corresponding term vectors. We used the average individual's distance from a group's centroid to represent the variation within a group.

5.2.1 Group Index Similarity

In general, across both the task-based and trait-based datasets, the average participant's desktop index had a cosine similarity of 0.40 to the median of the entire study population. The indices of our participants as a whole were particularly similar for some queries, such as "Microsoft new technologies" (cosine similarity of 0.77), and particularly dissimilar for others, such as "economic comparison telecommuting versus office" (0.07). On average, there was more similarity across work queries (0.54) than social queries (0.41). The fact that participants had more similar work-related content on their computer than social content likely reflects the fact that all participants worked for a common company. There was less similarity for task-based queries (0.31) than for either the work or social queries. This may reflect that the task-based queries were very specific and focused.

As can be seen in Table 4, the user profiles of interest-based group members were more similar to each other than the profiles of people not in the group. On average, interest-based groups had a cosine similarity of 0.52, while people not related by interest had a cosine similarity of 0.47. The difference was significant ($p<0.01$). Photo enthusiasts' indices were 23% more similar than non-photo enthusiasts ($p<0.05$), and pet lovers were 10% more similar than non-pet lovers. These differences were particularly pronounced for the group-relevant queries: photo enthusiast profiles were 82% more similar for photography queries, pet lovers were 26% more similar for pet queries, and vegetarians were 58% more similar for food-based queries. The similarities may be because the interest-based groups were selected based on mailing list participation, and thus members were likely to have a set of similar emails saved. But as these emails likely represent only a small amount of the content on their desktop computer, it could also represent a broader correlation of interests.

Work-related groups also tended to have similar indices compared to people not in the group (9% more similar, $p<0.01$), and were particularly similar for work-related queries. This could reflect the fact that similar emails and documents are often common amongst members of work groups. A notable exception was for researchers, who appeared to have very dissimilar indices, perhaps due to the independent, self-defined nature of many of their jobs.

We also found that task-based group members' indices were more similar than for people not in the task-based groups. When comparing across the 10 common task-based queries, task-based groups had an average cosine similarity of 0.42, compared with 0.31 with task-based participants not in the same group ($p<0.01$). Task-based groups' indices were most similar for their task-related queries. Within a group, the average index similarity was 0.77 for task-based queries, and 0.35 for other queries ($p<0.01$).

5.2.2 Using Index Similarity to Create Groups

Given the observed similarity in indices within groups, it is not surprising that we found index similarity does a good job of predicting explicit group membership. For each query, we applied *k*-means clustering to the index similarity data to group the participants who evaluated the query into two clusters. We found many strong correlations between the memberships of these clusters with the memberships in our known groups.

For example, one of the clusters of participants created using the similarity of people's desktop indices for terms from the "black & white photography" query results correlated strongly with membership in the photo enthusiast e-mail list ($r=0.49$). The other cluster was correlated with non-membership in the list. Clusters associated with query results for "photosynth" (a piece of software for digital photography effects) were also correlated with being a photo enthusiast ($r=0.35$) as well as being highly correlated with being male ($r=0.44$). The clustering for the results from "redmond restaurant" correlated with membership in the vegetarian e-mail list ($r=0.33$), as did the clustering from "bread recipes" ($r=0.41$). Note that correlations with groups based on e-mail list membership are likely strengthened by the fact that most list members have some identical content contained within their desktop indices (namely the messages sent to the distribution list).

However, the clustering also showed correlations with groups that do not share e-mail distribution lists; for instance, the clustering for index contents associated with the query "business intelligence" correlated with the list of study participants whose job role was program manager ($r=0.40$), as did the clustering for the query "live meeting" ($r=0.25$). The desktop similarity clustering for the query "cat on computer" was most closely correlated with the list of participants who lived in suburban, rather than urban, areas ($r=0.25$) as opposed to members of the pet lovers e-mail list ($r=0.16$).

Clustering indices did a particularly good job of identifying task-based groups. The mean correlation coefficient for task-based groups with the group's associated common query was 0.42. The clusters created using the more unusual task-based queries, such as "snp disease data" ($r=0.74$), "diablo ii character guide" ($r=1.00$), and "newpage corporation" ($r=0.59$), were particularly

accurate. For the task-based dataset, queries were identified by the task group members. It may be particularly easy to identify the task-based groups because participants chose specialized queries that were reflected in their desktop indices and not in others'.

In general, the fact that *k*-means clustering based on the cosine similarity of users' desktop indices does a reasonable job of re-creating group membership lists suggests that profile similarity could be used to create groups for group personalization in cases when explicit group membership data is not available.

5.2.3 Index Similarity and Judgment Similarity

However, we also observed a challenge to using index similarity to implicitly identify groups. When we studied whether desktop index similarity led to similarities in explicit relevance judgments, we found little evidence that it does. To measure the similarity in relevance judgments between pairs of individuals, we used Cohen's Kappa to calculate the inter-rater reliability. The mean correlation coefficient between the cosine similarity of any two individuals' indices for a query and those two individuals' inter-rater reliability for that query was only 0.09.

The fact that similarity in desktop indices does not predict similarity in judgments suggests index similarity may not be a good way to identify people with similar judgments. Instead, the dissimilarity in indices between people with similar judgments may be a benefit during ranking, when the unique content from a different group member's index may benefit the ranking in a way that content from the individual's own profile cannot.

5.3 Variation in Relevance Judgments

The earlier analyses of query selection and index similarity give some understanding of how people's interests varied by group. In addition, we explored the variation in explicit relevance judgments among groups of people. The cohesiveness of a group's relevance judgments allows us to understand whether the members of a group would benefit from all having their results ranked in the same, group-specific order. We explored this by calculating the inter-rater reliability between sets of judged results. We use Cohen's Kappa to represent the inter-rater reliability between pairs of individuals, and Fleiss's Kappa for groups. Both measure the extent to which the observed probability of agreement exceeds the expected probability of agreement if all raters were to make their ratings randomly.

Across both data sets, the average group studied had a Fleiss' Kappa (averaged across queries within a group) of 0.08 for their judgments. This indicates group members generally only agreed slightly about what was relevant to a query. Previous research [23] suggests that there is a lot of variation in what different individuals consider relevant to the same query, and our findings provide evidence that there is a large amount of variation in what people consider relevant even when those people are very similar.

Nonetheless, some groups had more similar relevance judgments within group than out of group, and these are the groups that are most likely to benefit from group personalization. On average, task-based groups had an inter-rater reliability of 0.16, compared with an inter-rater reliability of 0.11 for non-group members ($p<0.05$). This difference is most pronounced when only the task-based queries are considered ($\kappa=0.22$ v. $\kappa=0.11$, $p<0.01$). For these queries, the task-based groups have fair agreement.

We also looked at whether two individuals' similarity in judgments for one query predicted similarity in judgments for others. For example, if two people found the same results relevant to the query "slr digital camera" they may also be more likely to find the same results relevant to "black & white photography". To test this hypothesis, we looked at how correlated the inter-rater reliability was across pairs of individuals who issued two queries in common. We did not find evidence of a relationship, with the correlation coefficient being 0.05. We similarly did not find a relationship between judgments across queries when we looked only at pairs of individuals who fell within the same explicit group (the average correlation coefficient for a group was 0.00).

However, when we restricted our analysis to the pairs of related queries shown in Table 2, or to queries related to the same task in our task-based dataset, the average correlation coefficient increased to 0.09. It may be that similarities in judgments across people apply only to queries that are strongly related. The implication of this is consistent with our earlier findings that the most relevant group of people to use in group personalization are likely to be topic or task dependent.

Throughout Section 5, we have looked at understanding groups created by a variety of different means. It appears that the groups we studied primarily look cohesive for the queries related to the group topic. For example, work-based groups look cohesive for work-based queries but not social queries, social-based groups look cohesive for social queries but not work queries, and task-based groups looked cohesive for task-based queries but not for non-task-based queries. Short term, task-based groups seem particularly valuable to identify, since they have more similar notions of relevance than other groups. Query selection may be one way of identifying useful, task-based groups, while index similarity may be useful for interest-based or occupational groups. However, implicit measures such as query choice, desktop similarity, and similar relevance judgments do not appear to serve as good predictors of future relevance judgment agreement.

6. USING GROUPS TO IMPROVE RANK

We now look at using group data to improve personalization through groupization. In earlier work [10] we introduced the idea of expanding search result personalization to include user profile information from task-based groups of users. Here we explore the groupization algorithm in greater depth, and discuss how groupization performed for the different group types we studied. Consistent with the analysis in Section 5, groupization performs particularly well for group-related queries and task-based groups.

6.1 Groupization Algorithm

We built the groupization algorithm on top of an existing Web search personalization system. A groupized score for each of the top 40 results to each query (20 results for the task-based dataset) was computed for a group by summing the personalized score for that query across each group member.

The personalization system scores a result along two dimensions: 1) a content dimension that represents the similarity of the result's text content to the text content of the user's profile, and 2) a behavior dimension that represents the similarity of the result's URL to previous URLs visited by the individual. As was done by Teevan et al. [22], our content-based portion of the personalization score is computed by using the individual's desktop index as relevance feedback within the BM25 [18] framework. Algorithms that boost previously clicked URLs in search result rankings have been implemented with some success [4, 16]. For this reason the behavior-based portion was implemented by progressively matching the text of the result's URL to each URL previously visited by the individual.

Results are ranked according to the groupized score. Note that the groupization algorithm produces the same ranking for each group member. However, if producing the same list for all members is not a requirement, a combination of the personalization and groupization scores may lead to even larger performance gains.

6.2 Groupization Performance

We tested the groupization algorithm on the set of explicit relevance judgments we collected for the two datasets, using the different groups studied as input. This section begins with a discussion of how groupization performed using our entire study population as a group, and then discusses performance for groups that vary in terms of group longevity (task- versus trait-based) and in terms of group identification (explicit versus implicit).

We use normalized Discounted Cumulative Gain (DCG) [6] as our measure of performance. DCG gives more weight to highly ranked documents and allows different relevance levels to be incorporated by giving them different gain values. We used a gain of 2 for highly relevant documents, and 1 for relevant documents. Because queries associated with a higher number of relevant documents have a higher DCG, we normalized to a value between 0 (the worst possible DCG) and 1 (the best possible DCG given the ratings) to facilitate averaging across queries.

6.2.1 Group: All Participants

Overall, using the aggregate group data collected across all of our study participants yielded a significant improvement over individual personalization alone. Using the optimum balance between content and behavior-based ranking (found using leave one out cross-validation, 0.9 for behavior, 0.1 for content in each case), groupization resulted in a normalized DCG of 0.61, which is significantly better (p<0.01) than the average personalization performance of 0.55. What this means is that the groupization algorithm finds one query result ranking that can be returned to all participants that is on average better than each of the individual rankings found via personalization.

These results are consistent with previous research showing the combination of relevance indicators from different sources leads to better overall performance [3]. Additional improvement may derive from the fact that our participants were very similar, even when not divided into fine-grained groups. A large majority of the participants lived in the Seattle metropolitan area, and all worked for Microsoft. Lots of information about an individual has been shown to yield better performance for personalization algorithms than smaller amounts of higher quality information [22]; in this case information from many similar individuals may similarly yield improvements because of its quantity. In particular, we believe that some of the benefit may come from the general lack of correlation observed in index similarity compared to explicit judgment similarity (Section 5.2.3). People with similar ratings contributed very different user profile data to the groupization scoring and this was used to the group's advantage.

Also consistent with previous research [22], personalization that ignored the Web ranking as a prior did not improve on the Web ranking. The average normalized DCG for personalized rankings was 0.55, while for the Web ranking it was 0.57. In contrast, the rich user profile information used by the groupization algorithm was enough for the profile information alone to significantly improve on the Web ranking (to 0.61, p<0.01).

Even larger improvement gains were made by treating the Web ranking as a prior, and incorporating that prior information into the groupization ranking. When this was done, the group ranking yielded a normalized DCG of 0.63 and the individual personalized ranking 0.60. Both are significantly (p<0.01) better than the Web ranking, and groupization is significantly (p<0.01) better than personalization. Note that the relative improvement of including the Web ranking in the individual personalized ranking is greater than in the group personalized ranking (9.2% v. 3.5%). Web search engine rankings are based in part on large amounts of aggregate group data (using, for example, PageRank), and it may be that some of the benefit added by the Web ranking is not as valuable for a ranking that already uses group data, albeit of a different flavor and constructed on a smaller scale.

6.2.2 Explicit, Trait-based Groups

We next look at the performance of the groupization algorithm using the finer-grained trait-based groups described earlier (e.g., demographics, geography, occupation, and interest). We found that it significantly helps (or does not hurt) query performance for all of our explicitly-defined groups. The difference in performance between personalization and groupization is shown in Figure 2, broken down by group and query type.

Like with the group made of all participants, groupization with trait-based groups performs better than personalization. Trait-based groups received an average 3.5% boost in DCG over personalization by using groupization (0.59 to 0.61). For every group type studied the quality of the personalized and groupized rankings are higher for work queries (improvement of 4.4%, from 0.66 to 0.68) than social queries (improvement 2.9%, from 0.53 to 0.55). The difference between work and social queries may reflect the fact that participants were related through work (they all worked at the same company). In fact, we see a relatively larger gain in improvement for social queries for the interest-based groups compared with the work-related groups (4.4% versus 1.0%), and it is this combined boost in both social queries and work queries that may account for why groupization improves general performance particularly well for interest-based groups.

Mirroring a trend we observed in our analysis of the groups in Section 5, we see that for many of the work-related groups (Product Group I, program manager, developer), the observed improvement is significant (p<0.01) for work related queries, but not significant for non-work related queries. Again, researchers are unusual in that they don't benefit from using the data of other researchers to groupize the results, likely because researchers' foci were more diverse than for other job-team groups (i.e., most researchers have non-overlapping areas of specialization).

6.2.3 Explicit, Task-based Groups

In addition to looking at how well groupization performed using trait-based groups, we also explored how well it performed using task-based groups. In the analysis presented in Section 5, we found these groups to be particular cohesive, and thus it is not surprising that we similarly find the groupization algorithm is particularly useful for such groups.

The quality of the current Web ranking of the search results for task-based queries was viewed as better by people who were not members of the associated task-based group that it was by the people who were. For example, members of the group that was researching telecommuting thought, on average, that the results to "economic comparison telecommuting versus office" merited a normalized DCG of 0.45, while non-members of that group thought it merited a 0.62. On average, the Web ranking had a DCG of 0.58 for non-task-based groups and 0.51 for task-based

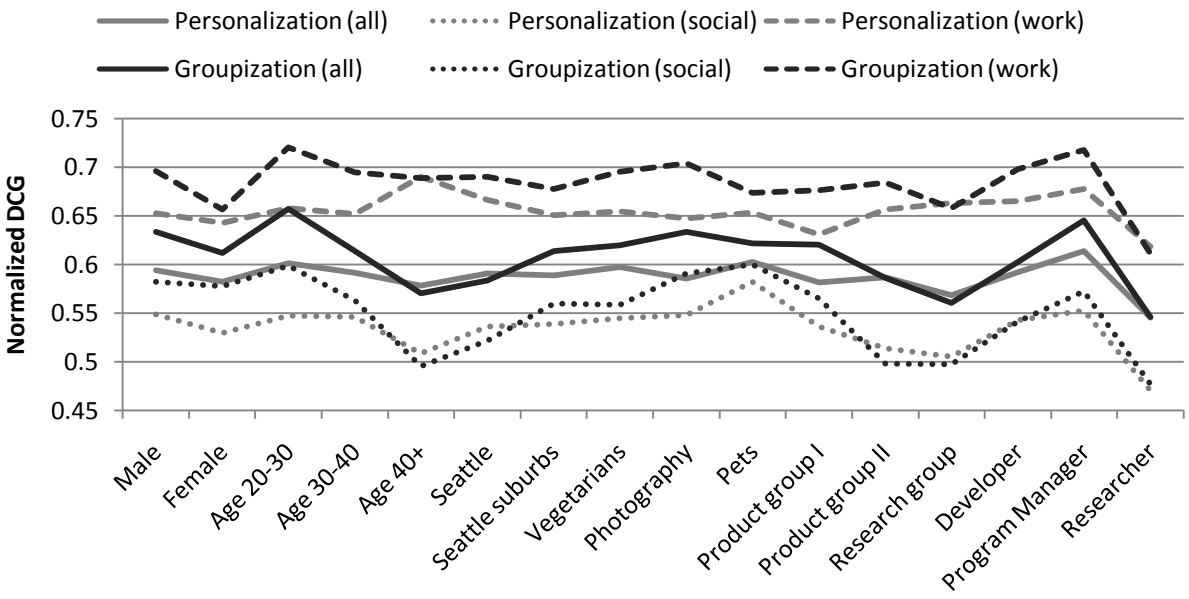

Figure 2. The performance of the groupization algorithm for trait-based groups, compared with personalization. Results are shown for all queries, social queries, and work queries.

groups (p<0.01). It may be that task-based groups all shared a similar, but atypical, interest in their chosen topics.

However, as can be seen in Table 4, groupization using the best parameter setting for task-based groups (again selected via leave-one-out cross validation) is able to bring the quality of the ranking up to a significantly better quality level than it does for non-task-based groups (to 6% better, p<0.01, even though it started out 13% worse). Within-group groupization improves the result quality 32% over the Web, to 0.67, while out-of-group groupization only improves the results 8%, to 0.63. This jump is much higher than we observed for the trait-based groups, and reflects the relative value of using short-term task-based groups.

We also found that using an explicit, task-based group to rank the results performed particularly well for the task-related queries. The performance within a group on task-related queries is 11% better than on non-task-related queries. This difference is significant (p<0.05).

6.2.4 Implicit Groups

In Section 5 we observed that implicit groups with similar notions of result relevance appear to be hard to identify. To explore the value of using implicit indicators of interest to identify individuals who would benefit from groupization, we looked at whether the cosine similarity of a group member's index to the group's centroid was correlated with the relative improvement gained by using groupization. If it is correlated, then the user profile may be useful in determining the value of applying the groupization algorithm.

Table 4. The performance of the groupization algorithm within task-based groups compared with outside of them.

	Normalized DCG	
Ranking algorithm	***Within group***	***Out of group***
Web	0.51	0.58
Groupization	0.67	0.63

However, we found that there was no correlation (correlation coefficient was 0.04). This echoes our earlier observation that participants with similar indices were not particularly likely to have similar relevance judgments. It may also support the hypothesis that groupization provides the most benefit when indices are different because it allows different (but related) content than is on the individual's own computer to contribute to the ranking. And, in fact, when the correlation was broken down by group, we observed that the similarity of the user profiles to the centroid for task-based groups were less correlated with relevance judgments than for other group types (-0.02 vs. 0.06).

In summary, we found that groupization provided improved results ordering as compared to personalization for several types of explicit groups, including task-based groups, interest-based groups, and occupational groups. This effect was more pronounced for group-related queries (such as work-related queries for occupational groups). However, it may be a challenge for groupization to add value for implicitly identified groups.

7. IDENTIFYING GROUPS *IN SITU*

In this study, we gained an initial understanding of the potential for using information about users' group associations to enhance personalization of their Web search results. The information we used to define our explicit group categories was gathered by a combination of participant self-reporting plus lookup in corporate databases. While some of this information (age, gender, occupation, interest groups, and zip code) might be available in enterprise search situations, it would not typically be available to commercial search engines. The ability to identify groups "in the wild" is important to the success of groupization techniques. We believe that this is an attainable goal; in this section we propose techniques for identifying our target group types.

Explicit, task-based groups can be identified through use of a collaborative search tool (e.g., [2, 12, 14]). Explicit, trait-based

information can be gathered from profiles that some (though not all) users fill out in order to register with and access custom features of many popular search engines. E-mail and instant-messaging contacts could also be used to construct group membership information. Additionally, search systems for the enterprise, for use on corporate intranets, would likely have access to employee directories with a variety of demographic information including job titles and hierarchies.

In this paper we examined a few implicit measures for identifying groups that would likely extend to less controlled settings, such as grouping users with similar term frequencies in their desktop indices, similar domains in their prior Web histories, or a history of issuing semantically or categorically similar queries. Implicit groups can also be identified via the use of server-side metrics that many search companies typically gather (in some cases only for users who have opted in to special services such as search toolbars, query histories, or personalization). Geographic data can be gleaned from IP addresses [9], and it may be possible to infer gender from query history [7]. Use of special-topic websites or portals may indicate interest-based groups [17]. Relevance-judgment similarity could be approximated using click-through data, or data from social bookmarking tools like http://del.icio.us. One can imagine that collaborative search tools might evolve as part of social networking sites (e.g., Facebook, MySpace) in which users' profiles and network structures could provide information relevant to several trait-based grouping categories.

8. CONCLUSION

In this paper we explored the potential for using information from a group of related users to enhance the personalization of Web search results. We analyzed the similarity of query choices, relevance judgments, and personal content for several categories of explicitly- and implicitly-defined task-based and trait-based groups. We found that several of the groups at the explicit end of the spectrum (task, occupation, and interest) were similar in many respects when considering queries related to their group's theme, but that for off-theme queries such groups were less cohesive. Implicitly-defined groups also lacked cohesion with respect to our three similarity metrics. We then analyzed the effectiveness of *groupization*, a personalization technique that combines personal and group content to improve Web rankings, for different group/query combinations. Our findings demonstrate that groupization improves upon personalization for several group types, particularly for explicit groups and group-related queries. These contributions suggest promising directions for future research into group identification and groupization methods.

9. ACKNOWLEDGEMENTS

We appreciate the time and effort contributed by our anonymous participants, and are grateful to Ashish Kapoor for his assistance with the algorithms used to identify various implicit groups. Susan T. Dumais and Edward Cutrell provided valuable feedback.

10. REFERENCES

[1] Almeida, R. and Almeida, V. (2004). A community-aware search engine. In *Proc. of WWW '04*, 413-421.

[2] Amershi, S. and Morris, M. R. (2008). CoSearch: A system for co-located collaborative Web search. In *Proc. of CHI '08*.

[3] Belkin, N. J., Cool, C., Croft, W. B., and Callan, J. P. (1993). The effect of multiple query representations on information retrieval system performance. In *Proc. Of SIGIR '93*, 339-346.

[4] Dou, Z., Song, R., and Wen, J. R. (2007). A large-scale evaluation and analysis of personalized search strategies. In *Proc. of WWW '07*, 581-590.

[5] Freyne, J. and Smyth, B. (2006). Cooperating search communities. In *Proc. of AH '06*, 101-110.

[6] Järvelin, K. and Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In *Proc. of SIGIR '00*.

[7] Jones, R., Kumar, R., Pang, B., and Tomkins, A. (2007). "I know what you did last summer": Query logs and user privacy. In *Proc. CIKM '07*, 909-914.

[8] Lee, Y-J. (2005). VizSearch: A collaborative Web searching environment. *Computers & Education*, 44(4): 423-439.

[9] Mei, Q. and Church, K. (2008). Entropy of search logs: How hard is search? With personalization? With backoff? In *Proc. of WSDM '08*.

[10] Morris, M. R., Teevan, J., and Bush, S. (2008). Enhancing collaborative Web search with personalization: Groupization, smart splitting, and group hit-highlighting. *Proc. of CSCW '08*.

[11] Morris, M. R. (2008). A survey of collaborative Web search practices. In *Proc. of CHI '08*, 1657-1660.

[12] Morris, M. R. and Horvitz, E. (2007). SearchTogether: An interface for collaborative Web search. In *Proc. of UIST '07*.

[13] O'Conner, M., Cosley, D., Konstan, J., and Riedl, J. (2001). PolyLens: A recommender system for groups of users. In *Proc. of ESCW '01*, 199-218.

[14] Pickens, J., Golovchinsky, G., Shah, C., Qvarfordt, P. , Back, M. (2008). Algorithmic mediation for collaborative exploratory search. In *Proc. of SIGIR 2008*.

[15] Pitkow, J., Schutze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E., and Breuel, T. (2002). Personalized search. *Communications of the ACM*, 45(9): 50-55.

[16] Shen, X., Tan, B., and Zhai, C. X. (2005). Implicit user modeling for personalized search. *Proc. of CIKM '05*, 824–831.

[17] Smyth, B. (2007). A community-based approach to personalizing Web search. *IEEE Computer*, 40(8): 42-50.

[18] Sparck Jones, K., Walker, S., and Robertson, S. A. (1998). Probabilistic model of information retrieval: Development and status. TR-446, Cambridge University Computer Laboratory.

[19] Spink, A. and Jansen, B. (2004). *Web Search: Public Searching of the Web*. Kluwer Academic Publishers.

[20] Sugiyama, K., Hatano, K., and Yoshikawa, M. (2004). Adaptive Web search based on user profile constructed without any effort from users. In *Proc. of WWW '04*, 675-684.

[21] Sun, J.-T., Zeng, H.-J., Liu, H., Lu, Y., and Chen, Z. (2005). CubeSVD: A novel approach to personalized Web search. In *Proc. of WWW '05*, 382-390.

[22] Teevan, J., Dumais, S. T., and Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. In *Proc. of SIGIR '05*, 449-456.

[23] Teevan, J., Dumais, S. T., and Horvitz, E. (2005). Beyond the commons: On the value of personalizing Web search. In *Proc. of PIA '05 Workshop*, 84-92.

[24] Twidale, M., Nichols, D. and Paice, C. (1997). Browsing is a collaborative process. *IP&M*, 33(6): 761-783.

[25] Voorhees, E. and Harman, D. (Eds.) (2005). *TREC: Experimental Evaluation of Information Retrieval*. MIT Press.