Design and Evaluation of a System to Support Collaborative Search

Robert Capra, Annie T. Chen, Katie Hawthorne, Jaime Arguello, Lee Shaw, Gary Marchionini
School of Information and Library Science
University of North Carolina at Chapel Hill
rcapra, atchen, kathryne, jarguell, ledshaw, march {@email.unc.edu}

ABSTRACT

We describe a collaborative search system called Results Space to support small groups of users in conducting asynchronous collaborative searches. We discuss the design of the system and present results from a laboratory evaluation. We also describe the development of an asynchronous collaborative task scenario (based on a task from the TREC Robust test collection) designed to elicit naturalistic behaviors with the system. Our results show that participants used the collaborative features not just to avoid duplication of effort, but also to check and refine collaborators' work, to gain a general understanding of collaborator's actions, and to get ideas for new queries. Although we expected participants to use the collaborative awareness mechanisms to find and rate new items, we found that participants were more likely to make ratings for result items that had been previously rated by their collaborators. Considered together, these results suggest a range of tactics and behaviors that collaborative search tools should support.

Keywords

Collaborative search, collaborative information seeking

INTRODUCTION

People work together to conduct searches for many reasons and in many situations. Over the past few years, researchers in information retrieval and information science have focused on the study of collaborative information retrieval (CIR) and collaborative information seeking (CIS) in which people work together to conduct searches and collect documents related to a shared information need. CIR and CIS have been the focus of recent workshops held at JCDL 2008, CSCW 2010, GROUP 2009, GROUP 2010, ASIS&T 2011, and CIKM 2011.

ASIST 2012, October 28-31, 2012, Baltimore, MD, USA.

There are many dimensions that affect CIS, including whether or not the collaborators are co-located or remote, working synchronously or asynchronously, the depth of the collaboration, and the roles that the collaborators may take in the search process (e.g., peers, power differential, prospector/miner). Supporting awareness and understanding of collaborators' actions and work in a collaborative system is a significant issue that is of interest to CIS researchers and designers (Paul and Morris, 2008; Shah and Marchionini, 2010; Golovchinsky et al., 2011) and is a focus area for our current work.

Many studies have investigated *synchronous* collaborative search between *two* collaborators (e.g., Morris and Horvitz, 2007; Shah and Marchionini, 2010; Pickens et al., 2008). However, several surveys have shown that people do a significant amount of *asynchronous* collaborative searching (Capra et al., 2011; Morris, 2007; Evans and Chi, 2008), often in small groups of between two to six collaborators (Capra et al. 2011). Studies have also reported that current tools do not support these activities well (Evans and Chi, 2008; Capra et al., 2010).

Another aspect of many prior laboratory studies of CIS is that they have used mainly recall-oriented tasks (e.g., find as many relevant documents as you can in the time allotted), and that many involve searching for information on the open Web, making computations of group-level precision and recall difficult.

In the work presented here, we set out to study CIS in the context of asynchronous collaboration among a small group of collaborators searching over a closed document set (e.g., a corpus of news articles). We designed a system, Results Space, to support this type of collaboration and developed a task scenario specifically designed to elicit a range of behaviors so that we could broadly examine participants' use of collaborative features.

Results Space includes collaborative awareness features that are embedded in the search results list and displayed within the interface. The system offers a rating mechanism and display of previous queries. In addition, it includes controls for reviewing and filtering results based on relevance ratings made by individual collaborators. These features provide users great control in reviewing and understanding the work that their collaborators have done.

Distinctions of our work include:

- Development of and evaluation using a task scenario designed to broadly consider user actions in an *asynchronous* collaborative search with multiple group members.
- Use of the TREC Robust corpus to support computation of standard recall and precision measures.
- Incorporating awareness displays and controls within the search results and document views to place them at the point of need.
- Highlighting and diminishing results in the search results display based on collective group ratings.

Our goals in this paper are to:

- Describe the Results Space system, its features, design rationale, and implementation.
- Present the asynchronous task scenario that we developed and describe aspects of the scenario designed to elicit naturalistic behaviors.
- Report results a laboratory study on the system with 14 participants. We describe observations about how participants used the collaborative awareness features, and also about how awareness of collaborators' actions may have influenced participants search behaviors and rating actions.

RELATED WORK

Many recent laboratory studies of collaborative search systems have focused scenarios involving two collaborators working synchronously, and many have focused on recall-oriented tasks. In this section, we briefly review prior studies, focusing on the collaboration configuration and the task types studied.

Morris and Horvitz (2007) studied use of their SearchTogether system in situations with two participants working synchronously in different locations, working on self-generated tasks such as joint planning and purchasing decisions, searching on the open Web. They noted the value of awareness and rating features and observed division of labor negotiations through a built-in chat mechanism. Shah and Marchionini (2010) and Shah and González-Ibáñez (2010) used a two-person synchronous configuration to study aspect of awareness and collaboration in the Coagmento system. Their task was recall-oriented and involved finding "all the relevant information" and to "collect as many relevant snippets as possible", again with searches conducted on the open Web. They documented the importance of providing group awareness features without increasing the cognitive load of users. Pickens et al. (2008) used a two-person synchronous scenario to examine algorithmically mediated collaboration. In their configuration, the collaborators were co-located and assumed specific roles of prospector and miner. They performed tasks on a set of TRECVID tasks+corpus and found that their collaborative system outperformed merging of individual users' searches.

Paul and Morris (2009) used an innovative configuration to study both synchronous and asynchronous collaboration within groups of three participants. They had two participants work together synchronously from different locations, and then at a later time had a third participant collaborate on the task asynchronously. Based on the results of their study, they designed the Co-Sense system to support sensemaking by providing timelines and other aggregated views of collaborators' queries and actions.

Golovchinsky et al. (2011) developed a system called Querium that incorporates a number of awareness and

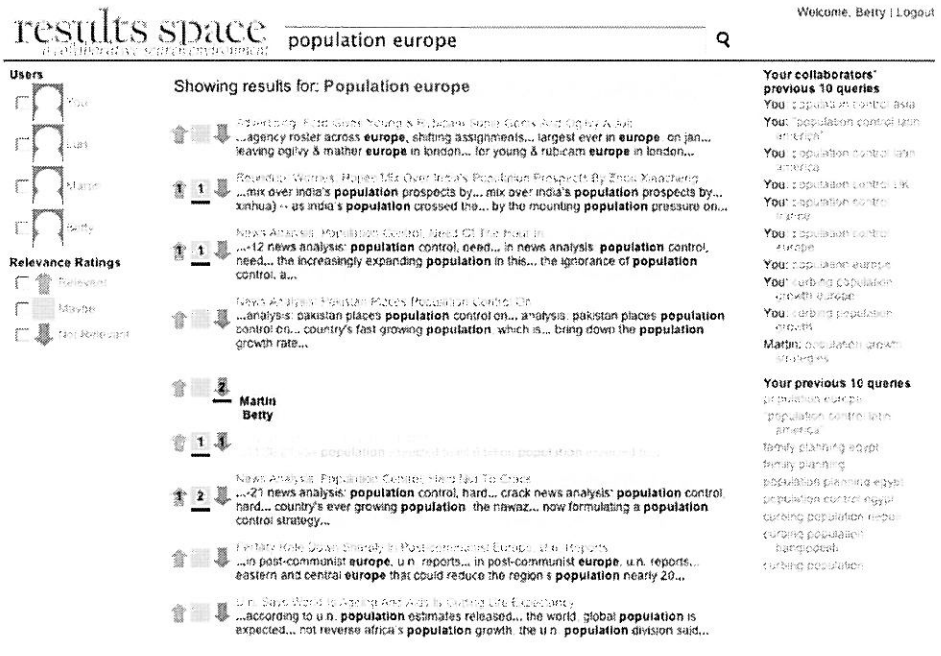

Figure 1. Results Space Collaborative Search Interface

collaborative communication mechanisms including embedding awareness information from prior searches within the search results and allowing users to selectively filter results using faceted controls.

TASK, CORPUS, AND COLLABORATIVE SCENARIO

Studying collaborative information seeking in a laboratory setting is challenging because the task and collaborative scenario description can have a significant impact on how participants approach the task. In this study, we sought to develop a task and scenario that would be ecologically valid for our participant population of university graduate students. We also wanted to craft a description that would inform participants about essential aspects of the task, but that would elicit as naturalistic behaviors as possible.

The scenario we developed situated the participant as a member of a team doing research for a group project for a university course assignment. This type of scenario has been described by participants in previous studies (e.g., Shelby and Capra, 2011) and similar scenarios have been used in studies of individuals doing exploratory searches (Kules and Capra, 2012). The text is shown below:

For this task, imagine that you are taking an Environmental Studies class here at UNC. As part of the class, your instructor has given you a research assignment to do in small groups. The goal of the assignment is to find articles that will help you write a research paper on an assigned topic (shown below). Your instructor has given you access to a database of news articles from 1996 to 2000 to be used for the assignment.

You are in a group with three other people (Luis, Martin, and Betty). Your team agreed that everyone would do some searches on the database to find articles that may be useful in writing your research paper. Your other team members may have already done some searches and the group has agreed to meet tomorrow to talk about what everyone found. Your task today is to find and rate articles that will help your group with the assignment.

In the scenario, the details of the status and goals of the collaborators were left intentionally vague to enable the participant to make choices about their approach to the collaboration. The task scenario situated the participant as an individual team member in the midst of a group conducting a collaborative search. This required us to "seed" the system with existing queries and ratings from the other team members. We describe the process we used to populate this seed data in the Method section. Paul and Morris (2008) used similar techniques in evaluating SearchTogether and Co-Sense. However, few other studies have used this type of scenario to evaluate collaborative search behaviors. We were careful in our wording (e.g., "team members may have done some searches", "your task is to… help your group") to try not to tilt the participants to stress recall over precision. By using a task that allowed participants to engage in a variety of collaborative behaviors, we hoped to gain insight into how the collaborative awareness features in our prototype would be used to support activities such as pruning, sensemaking, and finding new results.

For the information seeking goal, we wanted to use a well-defined task on a known corpus with ground-truth relevance judgments so that we could compute metrics such as group-level precision and recall. Group recall corresponds to the percentage of truly relevant documents collectively labeled as relevant by members of the group. Group precision corresponds to the percentage of documents collectively labeled as relevant that were truly relevant. While we did not explicitly ask participants to maximize either of these common IR evaluation measures, we wanted the ability to measure them.

To this end, we used the TREC Robust Track collection (Voorhees, 2006), including the AQUAINT IR test collection. TREC is a yearly workshop hosted by NIST to facilitate the benchmark comparison of IR systems. The AQUAINT collection consists of about one million English newswire articles published from 1996-2000. We did some cleaning on the dataset by removing duplicate articles, articles with missing titles, and articles with little narrative, resulting in a set of 856,941 documents. The TREC Robust collection includes relevance judgments made by NIST assessors based on the narrative for each of the search tasks.

For our study, we chose a TREC Robust task focused on finding articles about measures taken worldwide to curb population growth (Task 435). We selected this task for two primary reasons. First, we wanted to avoid a task with thousands of relevant documents (too easy) or only a handful (too difficult). The selected task contained 144 relevant documents in our curated version of the AQUAINT collection, which we considered to be a reasonable middle ground. Second, we wanted a search task that that would be engaging for our user study population: university students. In a study conducted by Bailey et al. (2009), this TREC task was rated by undergraduate student participants as being the second most interesting search task among 20 tasks that were evaluated. We carefully adapted the task wording (see below) to fit our scenario, being careful to preserve nuances of the task to preserve the integrity of NIST relevance judgments. We note that part of the complexity of this task is its specification of what *is* and *is not* considered relevant.

Use the news article database to find articles that will help you write a research paper on the topic below.

What measures have been taken worldwide and what countries have been effective in curbing population growth? While researching this topic, keep in mind that your paper will be stronger if you support your thesis with actual cases in which population measures have been taken and the results are known. For this assignment, reduction measures to control growth are defined as those that are being actively pursued. Passive events such as disease or famine that involuntarily reduce population should not be cited.

PROTOTYPE

In this section, we discuss the design and implementation of our collaborative search system. We present the major

components of the system and explain design rationale for each. We also describe the backend database and search engine that we use to support the interface functionality.

Design

Our prototype system, called Results Space, was designed to with two goals. First, it was designed to support small groups of two to six people in conducting collaborative searches. Second, we designed the system to support our goals for conducting research on the effects of various awareness mechanisms on dimensions of the collaborative search process. In our current version, it is configured to search and present results from the AQUAINT corpus.

The major components of the Results Space user interface are outlined below and shown in Figure 1.

Query Box

The query text box was placed at the top of the page and the search button is denoted a magnifying glass icon.

Results Displays

After issuing a query, results are shown in the middle of the page, with each result item consisting of the title of the news item and a snippet of text from the article with the search terms bolded. The overall presentation of the results is similar to many search engine results pages (SERPs).

We added two elements to the results presentation that are specifically designed to increase awareness of collaborators' prior actions and impressions of each result. First, we highlight or diminish each result based on the overall consensus of the group ratings. When ratings of the group are collectively a net positive, the background of the result in the SERP is highlighted green. Conversely, results with a net negative rating are greyed out to give a "faded" appearance. Multiple negative ratings result in an even more transparent display that makes the document surrogate difficult to read (i.e., there are two levels of fading). Our motivation for this feature was to emphasize the collective rating of each document. For example, in Figure 1, we can see document #2 and #3 have been rated as relevant and are highlighted green, whereas document #5 and #6 have been rated different levels of not relevant. Second, to the left of each result item, we display a set of three rating displays and controls. These will be described in more detail below.

Document View Page

Clicking a result item takes the user to a page that shows the text of the article along with a set of ratings controls similar to those in the results listing, and a back button to return to the SERP.

Relevance Rating Display and Controls

The Results Space system allows users to rate result items with one of three relevance ratings: "relevant", "maybe relevant", and "not relevant". A set of combined display/control icons for these ratings are shown to the left of each search result: a green up arrow ("relevant"), a red down arrow ("not relevant") , and a yellow box ("maybe relevant"). We display the rating icons in vertical columns to the left of each result to support easy scanning of results for each rating. Users can click on these icons to indicate their evaluation of a document. In addition, the icons serve as an awareness display of collaborators' ratings – numbers are displayed inside the icons to indicate the number of collaborators who have made that rating. Users get immediate feedback when they click their own ratings by the display of a black bar underneath the icon of their rating. This makes it easy for a user to see that they have made or changed a rating without having to rely on recognizing that the count has changed. Using the mouse to hover over one of the rating icons will cause a small box to be appear next to the icon with a display of the names of the collaborator(s) who had made that rating (e.g., "Martin, Betty" in Figure 1).

Filter Mechanisms

The filter mechanisms on the left side allow users to filter by users (i.e., individual collaborators) and relevance ratings. The system allows the option of applying the filters to narrow the results of a specific query or to apply them over the entire collection of documents (with an empty query). The currently applied filters are reinforced by a grey box at the top of the results that displays a textual representation of the query and filters (Figure 2).

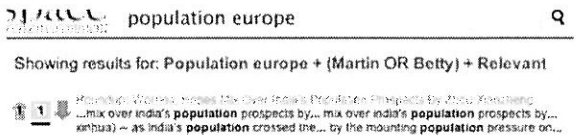

Figure 2. Current query and filter display above results

Query Histories

On the right hand side of the screen are two displays of query histories: the collaborators' previous queries (top), and the participant's own queries (bottom). These histories are provided to give users a topical overview and to provide awareness of where their collaborators had recently focused their searches. Each query history is ordered by time and displays the 10 most recent queries.

Design Summary

The prototype was designed to include features that would increase users' awareness of their collaborators' prior activities. We did not include a note-taking or messaging/chat feature between collaborators because in this phase of our research we wanted to focus on the design and use of the query history and rating mechanisms. However, we note that note/chat features are common communication methods in collaborative search systems.

Implementation

The Results Space system is written mainly in PHP. It relies on two primary data structures, a document index and a MySQL database, and utilizes the Indri search engine for

document indexing and retrieval. The MySQL database stores past search results, document ratings, user and group profiles, and settings for our experimental studies. User actions are also logged in this database.

When a user performs a query, the system determines whether the query involves filters or not. If it does not involve filters, the system runs the query in Indri and retrieves a ranked results list, which it then displays to the user. If the query involves filters, the system retrieves documents fitting filter specifications from the MySQL database, and then calls Indri to rank the documents.

METHOD / EVALUATION

To understand how users would make use of the collaborative features in an asynchronous task scenario, we conducted a laboratory study. Details of the task, corpus, and scenario we used are described previously in the paper.

During the task, the Results Space system logged user interactions including the queries issued, result items rated, result items viewed, filters applied, previous queries clicked on, and use of the back and next buttons to move through pages of the search results. After completing the task, we asked participants to complete an online questionnaire and conducted a verbal interview with questions about their experiences. In this paper, we report primarily on analysis of the log data and the post-session interview responses.

We recruited participants who were all graduate students at the University of North Carolina at Chapel Hill. A recruitment email was sent to an opt-in mass email distribution list and participants were selected based on the order in which they replied to our recruitment and their availability to schedule a session.

The first five participants were recruited from the School of Information and Library Science and played special roles. The first two (p11, p12) were treated as pilot participants (we started numbering at p11). The next three (p13-15) used the same experimental protocol as all participants, but they incrementally provided the "seed" data for the following participants. In other words, p13 started with no existing collaborator data, p14 started with p13's queries and ratings, and p15 started with data from both p13 and p14. The 11 subsequent participants (p16-p26) started with data from p13, p14, and p15 as their "teammates", referred to by the pseudonyms Betty, Luis, and Martin.

Upon arrival, participants were greeted, escorted to a quiet room in our lab, and seated at a computer workstation. The experimenter gave a brief overview of the study, explaining that the participant would be using an experimental search system designed to help small groups collaborate on searches and that we would ask them to think aloud while doing the task. We then gave the participant an informed consent form to read and sign, and administered a short demographic questionnaire. Next, the experimenter played a short video (~2.5 minutes) that introduced the Results Space system and described its main features. Then we gave the participant the task scenario and asked them to read it aloud to be sure they read all the parts. After a chance to ask questions, we then gave them 30 minutes to work on the task with a 5 minute warning at the 25 minute mark. Participants were informed that they could stop earlier if they reached a point where would normally stop.

RESULTS

Interactions with the Interface

To gain an overview of how participants interacted with the system, we calculated aggregate statistics based on actions logged by the system. The actions logged included: typing a query into the query box (Query), clicking on a link to a collaborators' prior query (CollabQ), clicking a link to one of the participants' own prior queries (PrevQ), clicking to go to the next or previous page of search results (NextSERP/PrevSERP), using the faceted controls to filter the current results (Filter), clicking a rating button next to a result on the SERP (Rate@SERP), clicking a rating button on the document view page (Rate@Doc), clicking one of the document links on the SERP to view the document (ViewDoc), and clicking the back button on the document to go back to the SERP (Back2SERP).

Action	n	Σ	M	SD	Min	Max
Query (type)	11	124	11.3	6.6	3	25
CollabQ (click)	4	19	4.8	1.3	3	6
PrevQ (click)	3	3	1.0	0.0	1	1
NextSERP (click)	11	143	13.0	10.3	4	40
PrevSERP (click)	2	19	9.5	7.5	2	17
FilterSERP (click)	6	38	6.3	2.7	3	10
Rate@SERP (clk)	7	77	11.0	13.0	2	42
Rate@Doc (click)	11	181	16.5	5.1	9	27
ViewDoc (click)	11	263	23.9	10.1	15	45
Back2SERP (clk)	11	253	23.0	8.9	14	45

Table 1 Descriptive Statistics for Participant Actions

Table 1 shows the number of participants who used each type of logged action (n), the total number of times each action appeared in the log (Σ), and the mean, standard deviation, min, and max for the actions across the participants who used them. As Table 1 indicates, some features were used by all the participants (e.g., n=11 for Query, NextSERP, Rate@Doc, ViewDoc and Back2SERP), while other interface components were clicked on by fewer participants (e.g., CollabQ, PrevQ, PrevSERP). We note that while only four participants clicked on their collaborators' prior queries, many made use of these queries visually as described later in the paper.

Rating and Viewing Documents

All 11 participants clicked on documents to view them and all 11 made ratings from a document view page. Seven participants made ratings directly on the SERP, indicating that having the ratings buttons in both locations was useful. When clicking ratings directly on the SERP, we observed participants basing their ratings on the title and snippets.

For example, in one case where a search had returned an entire page of poor results, the participant quickly marked all the results as not relevant and commented that they hoped that this would help their collaborators avoid having to review these results. On average, participants viewed 23.9 documents, rated 16.5 documents from the document view page, and rated 11.0 documents from the index page, indicating that the participants engaged with the task and found it useful to make ratings from both locations.

To get a better understanding of the ratings, we generated the plot in Figure 3. From the figure, we see variation in how people used the "maybe" and "not relevant" ratings. Some participants made very few "maybe" ratings (e.g., p26), and others made very little use of the "not relevant" rating (e.g., p17, p18, p22, and p23). Participant comments suggest variation in how participants interpreted the ratings. For example, for the "maybe" rating, some participants used it for items they were unsure about and others used it to mark items that they felt the group might need to consider. We also noticed some participants rating some relevant items as "not relevant" after reaching a point of saturation on the sub-topic of the item. For example, articles on China's efforts to manage population growth were plentiful in the corpus and some participants started rating them "not relevant" after finding many such articles.

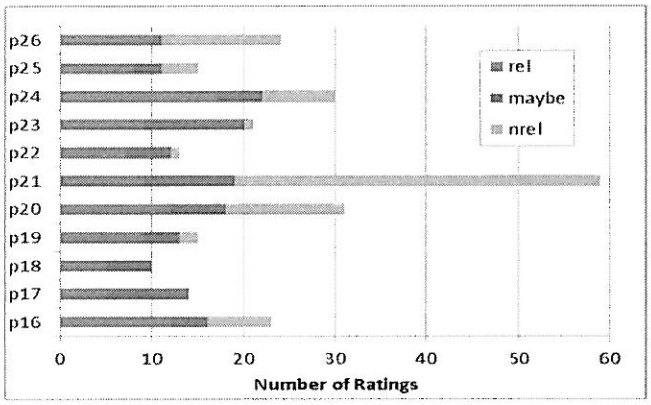

Figure 3. Participant Ratings by Type

Based on observations, we decided to explore a possible difference in how participants rated items when making the rating on the SERP versus making the rating on the document view page. We suspected that when rating on the SERP, participants made "not relevant" ratings more often, and when rating on the document view page, participants made "relevant" ratings more often (Table 2).

	Rated relevant	Rated not relevant	Σ
Rate@SERP	10	57	67
Rate@DOC	93	32	125
Σ	103	89	192

Table 2. Analysis of Place Rated vs Relevance Rating

Analysis of this cross-tabulation data shows a significant effect (χ^2=59.68, 1 d.f., p<0.001). When items were rated on the SERP, participants were more likely to rate the item as "not relevant". When items were rated on the document view page, participants were more likely to rate them as "relevant". This result seems logical – participants were likely to click on items in the SERP that appeared to be relevant and would then rate them as such from the document view page. It was easy to identify many documents as non-relevant from the title and snippet on the SERP and make the non-relevant rating there.

Filter Use

Six participants used the filtering mechanism. Each time an individual filter was applied or removed it resulted in a logged action. Participants often sequentially clicked several filters to apply them as part of one logical action. Filter use often occurred in clusters at the start, middle, or near the end of a session at points where a participant decided to investigate what their collaborators had found.

Actions across the Session

We note that the statistics reported in Table 1 all have large standard deviations representing the diversity in patterns of use of the features by our participants. To explore these variations in use, we generated visualization timelines to illustrate the actions that each participant took throughout the 30 minute session (Figure 4). By examining these patterns, we can gain insight into the search strategies. For example, Participant 18 spent most of their time running queries (black) and paging through pages of results (dark green), and made very few ratings (red). Participant 24 viewed documents throughout their search (grey) and early in the search process made ratings after viewing the documents, but partway through the session, stopped making ratings after viewing documents. Participants 19 and 24 both began clicking on collaborators' queries towards the end of their sessions (pink), perhaps to get new ideas. Participant 21, around the 11 minute mark, made numerous ratings (probably "not relevant" ratings) in a quick succession from the search results page.

Delta Precision and Delta Recall

While searching, users can adopt different strategies. For example, they can try to find as many relevant documents as possible (prioritize recall) or they can try to find only a subset, but minimize the number of false positives within this set (prioritize precision). Collaborative search is no different. A searcher might expend most of their effort finding relevant documents that were missed by their collaborators or might expend most of their effort pruning the set of documents previously rated relevant by their collaborators. Our collaborative awareness features have the potential to support both strategies—they provide the necessary information to avoid overlapping relevance judgments (prioritize recall) and they provide the necessary information to prune the documents already identified as relevant (prioritize precision). One distinctive aspect of the current study is that in our task scenario, we did not explicitly tell participants which strategy to adopt. This

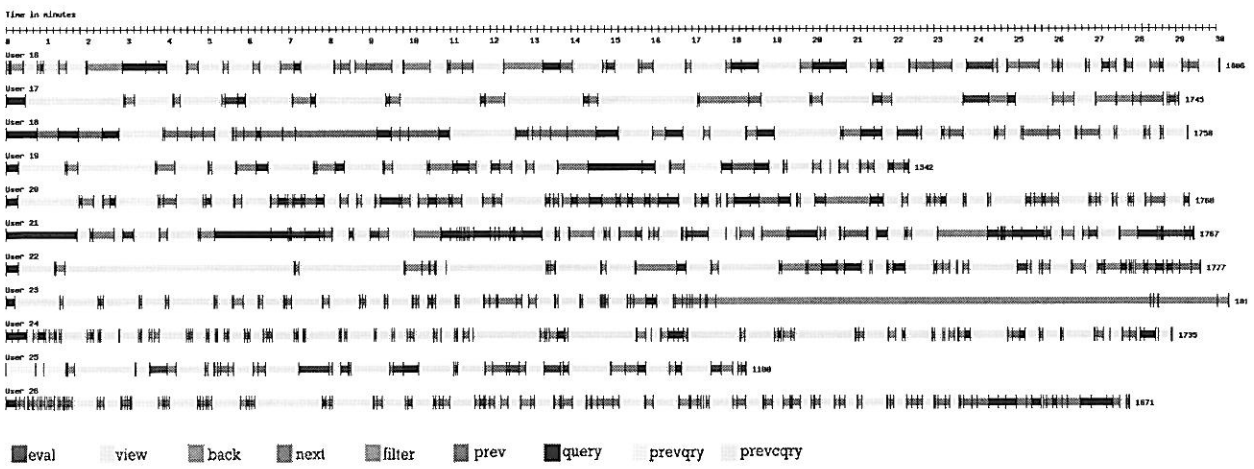

Figure 4. Action Sequence Timeline

flexibility was important so that we could see the variety of behaviors and strategies that participants used.

Table 3 shows each participant's percent improvement in group precision (P) and recall (R) over the precision and recall levels obtained by the three baseline collaborators (shown in gray), which were the same for all participants.

The set of documents collectively labeled relevant corresponds to the set of documents with a "relevant" majority vote. The voting scheme was operationalized as follows. Each "relevant" label was considered as +1, each "non-relevant" label was considered as -1, and each "maybe relevant" was considered as 0. The sum of these votes was aggregated across the four members of the group (the participant and the three baseline participants) and finally the document was considered "relevant" if the aggregate score was greater than 0 and "non-relevant" otherwise.

Pnum	Rated Rel.	Corr. Rated Rel.	True Rel.	P	R	ΔP %	ΔR %
Base.	30	13	144	0.43	0.09	--	--
16	39	17	144	0.44	0.12	2.33	33.33
17	31	14	144	0.45	0.10	4.65	11.11
18	35	17	144	0.49	0.12	13.95	33.33
19	33	15	144	0.45	0.10	4.65	11.11
20	42	16	144	0.38	0.11	11.63	22.22
21	34	15	144	0.44	0.10	2.33	11.11
22	32	15	144	0.47	0.10	9.30	11.11
23	33	14	144	0.42	0.10	-2.33	11.11
24	41	18	144	0.44	0.13	2.33	44.44
25	31	13	144	0.42	0.09	-2.33	0.00
26	35	16	144	0.46	0.11	6.98	22.22

Table 3. Participants' percent improvement in collaborative precision and recall

Table 3 shows several noteworthy trends. First, whatever strategies different participants adopted resulted in different improvements to precision and recall. For example, participant 24 increased recall by about 45% and precision by only about 2%. Conversely, participant 22 increased precision and recall roughly equally (i.e., about 9% and 11%, respectively). Thus, it seems to be important for collaborative search systems to support both strategies: to provide awareness features that allow users to avoid redundant relevance judgments (supporting an improvement in recall) as well as to provide features that allow users to arbitrate previously made collaborative judgments (supporting an improvement in precision)

Influence of Prior Ratings

Our task scenario gave participants flexibility in selecting how to balance their collaborative activities, keeping with our goal of observing naturalistic interactions. Some participants focused on more recall-oriented approaches, while other participants engaged in more sensemaking and pruning (i.e., precision-oriented) activities.

While watching the participants conduct the tasks and listening to their think-aloud comments, we observed that many participants seemed to gravitate toward items that had been rated by their teammates. These items were distinguished in the interface through the numbers displayed on the rating arrows, and through the greening-up (highlighting) and greying-out (diminishing) features.

Based on our observations, we suspected that participants might have been more likely to rate an item that had previously been rated by one of their teammates than to rate a previously unrated item. Based on the log data, for each participant, we listed all the result items they had been shown on SERPs, and for each of these placed it into one of four categories in a 2x2 cross-tabulation based on whether or not the item had been previously rated or not and whether or not the participant rated it. To get an overall picture, we summed each of these four counts over the 11 participants (Table 4). This gives us a view across all the opportunities participants had to rate or not rate documents.

	Teammates previous		Σ
Participant	rated	not rated	
rated	116	195	311
not rated	140	854	994
Σ	256	1049	1305

Table 4. Effects of Prior Ratings on Likelihood to Rate

Analysis of this data shows a significant effect (χ^2=79.49, 1 d.f., $p<0.001$). Across all rating opportunities, participants were more likely to rate items that had been previously rated than would be expected overall. We also considered whether or not *each individual participant* was more likely to rate previously rated items. Out of the 11 participants, 7 were more likely to rate previously rated items. We note that other factors besides collaborators' previous ratings could contribute to this observed effect. For example, the previously rated documents could be ones that were, in general, easily found. As part of a follow-on study (in progress) we are further investigating this aspect.

We also looked at the relationship between the valence (positive or negative) of a participant's rating, and of the group's existing rating, in two situations: when the group evaluation was positive, and when it was negative. In both cases, the chi-square was significant. In cases where the group evaluation was positive, participants were more likely to rate the document positive, (χ^2=22.35, 1 d.f., $p<0.001$), and in cases where the group evaluation was negative, participants were more likely to rate the document negative, (χ^2=6.90, 1 d.f., $p<0.01$).

Post-Session Interviews

Following task completion, we asked participants a series of semi-structured interview questions about their experience doing the task. Specifically, we asked: 1) Did they trust their collaborators' ratings?, 2) Did they make use of their collaborators' queries?, 3) Did they use their collaborators' ratings?, and 4) Did they intentionally try to write queries that were different than their collaborators? The questions were asked as yes/no questions with an additional "why or why not?" component. For each question, two coders independently listened to and coded the participants' responses from the recorded audio. The coders coded yes/no responses for each question and also each generated their own set of open codes for the "why" responses. The two coders and a third researcher then met and resolved the codes through a process of consensus. The independently generated codes had a good deal of commonality, so the code merging/consensus process was fairly straightforward. In this section, we present and describe the final classifications, and illustrate them with examples from the participants.

Trust in Collaborators' Ratings

We asked participants if they trusted their teammates' ratings. Ten of our 11 participants said that they generally trusted the teammate's ratings. Of the eight participants who provided additional comments, five noted that they "mostly" trusted their teammates' ratings, but that they had seen a few exceptions where they disagreed with a specific rating. One participant stated that they did not trust the ratings of a particular teammate. Two participants commented about not wanting to disagree with their teammates to avoid conflicts or to avoid having to redo work assessing the documents. We found these comments interesting from two perspectives. First, it provided evidence that participants took our task scenario seriously, engaged with the task, and had reactions that reflect real-world concerns. Second, it illustrates a well-known dynamic of group work (avoiding conflicts) that may have a significant impact on the quality of the search results gathered by the group. If teammates avoid disagreements in their ratings, this could impact group precision (and possibly group recall).

Intentionally Writing Different Queries From Collaborators

We asked participants about their use of a strategy that we anticipated would be used to improve group recall – intentionally writing different queries than their collaborators. This is an implicit method of division of labor that can be employed in collaborative search systems that allow collaborators to see each other's query histories. Since our task encouraged participants to find articles that would help their group, we anticipated that they might make use of this strategy.

All but one (10 of 11) of our participants said that they had used this strategy, and gave several reasons for doing so. Six participants talked about intentionally writing different queries to avoid overlaps with their collaborators' results. Three participants said they used this strategy to expand the results. We interpreted these two responses as representing different goals. We viewed avoiding overlaps as primarily motivated by expanding the breadth of the search, while expanding the results as focused on getting more depth.

Collaborator Query Usage

We asked participants if they used their collaborator's queries with a goal of gaining insights into how this type of collaborative awareness affected the strategies that our participants used to conduct their search. As described in the summary of logged actions, only four participants actually clicked on their collaborators' previous queries. However, ten out of our 11 participants reported consciously looking at and using their collaborator's queries from the query display. Based on our coding of the interview data, their motivations were grouped into four main categories:

- to write different queries from what their collaborators had already done (2 of 10)
- to get an overall familiarity of what collaborators had been looking for without an end goal in mind (3 of 10)

- to look at the train of thought their collaborators had been following and try figure out where to start their search (4 of 10)
- to get new ideas; participants reported turning to their collaborator's queries out of frustration with their own self-generated queries, and looking to their collaborator's searches for inspiration (4 of 10).

One user did not report using the collaborators' queries, explaining that they did not notice them.

These results reinforced that our participants had a collaborative mindset coming into the task, and they used the collaborative awareness mechanisms that we provided to accomplish this goal. This also reflects that our participants wanted to find their place within the group and how they could contribute to the success of the group based on what had already been accomplished. Looking at collaborator's queries also gave participants somewhere to turn when they were frustrated with their own results and needed new directions or keywords and also helped keep participants on track with the rest of their group.

Collaborator Rating Usage

We wanted to understand participants' self-reported impressions about how they used the ratings, and if they made use of collaborators' ratings. All eleven participants described using collaborator's ratings. Based on our coding, we classified responses into two main categories:

- to help select which documents to view (9 of 11)
- to focus on documents that had disparate ratings (4 of 11), with an intention of helping to resolve the ratings

Use of their collaborators' ratings to help select which documents to view suggests that participants focused effort on understanding what their peers had already found – an indication of sensemaking and trying to understand the task through the efforts of their collaborators. Several participants even commented that they needed to be "up to speed" on what their collaborators had found since the task scenario described a group meeting the next day.

The use of the collaborators' ratings seemed to heavily influence the paths that our participants took within their search, in some cases leading participants to greater overlap and a stronger drive to work within the results their collaborators had already found – efforts that favor precision over recall.

DISCUSSION

While searching, users can adopt various different strategies. One strategy might be to try to find as many relevant documents as possible (improving recall, possibly at the expense of precision). Another strategy might be to try to find only a fraction of the relevant documents, but to ensure that whichever documents are identified as relevant are truly relevant (improving precision, possibly at the expense of recall). Collaborative search is no different. A searcher might expend most their effort trying to find relevant documents missed by their collaborators or might expend most of their effort trying to prune the set of documents previously identified as relevant by their collaborators. In theory, collaborative awareness features support both strategies. They provide the information necessary to avoid redundant relevance judgments (improve recall) and they provide the information necessary to prune the documents already identified as relevant (improve precision). One unique aspect of our study is that we did not explicitly tell participants which strategy to adopt.

The results in Table 3 show that different participants adopted search strategies that resulted in different types of improvement. For example, participants 16 and 24 improved collaborative recall more than collaborative precision, while participants 18, 20, and 22 improved collaborative precision more than recall. Thus, it seems to be important for collaborative search systems to support both strategies: to provide awareness features that allow users to avoid redundant relevance judgments (supporting an improvement in recall) as well as to provide features that allow users to arbitrate previously made collaborative judgments (supporting an improvement in precision).

We were surprised to not observe a greater improvement in collaborative recall. In other words, we expected participants to use the collaborative awareness features primarily to find new relevant documents. However, we did not find this to be the case. Several aspects might have caused this behavior. First, across all experiments, participants started out with 30 documents already rated as relevant by their collaborators. A greater improvement in collaborative recall might have been observed if participants had started with fewer documents already rated as relevant. Second, it may be that collaborative awareness features kept participants from exploring entirely new territory. From a cognitive load perspective, it might be easier for participants to explore previous collaborator interactions (previously rated documents or previously issued queries) than to develop and evaluate their own search strategies. Indeed, our results indicate that participants were more likely to rate a previously rated document than a previously unrated document. Thus, it is possible that collaborative features can have the (possibly) undesired effect of limiting the range of content explored by different members of the group. These results contrast with a recent study by Shah and González-Ibáñez (2011) that found no positive impacts of collaboration on a recall-oriented collaborative search task. We believe this is an important question for future research.

Participants in our study used their collaborators' previous queries to help them know where to start their searching, to generally gain an understanding of their collaborators' work (sensemaking), to intentionally write queries that were different from their collaborators, and to get new ideas when they were frustrated. Participants used their collaborators' previous ratings to select documents to view and to find documents that had conflicting ratings in need

of resolution. Overall, participants reported trusting their collaborators' ratings, but they often mentioned exceptions.

Overall, the increase in group recall based on our participants' searches was lower than we had anticipated. We specifically designed our task scenario not to push participants toward "finding all relevant documents" and observed participants engaging in multiple strategies across their search sessions. We also found that among all the documents viewed, participants were more likely to rate documents that had already been rated by their collaborators, and were likely to assign the same rating as their collective collaborators. These results suggest that participants had goals other than simply finding additional relevant articles and found value in contributing ratings to a common set of results.

LIMITATIONS

As with all research, there are limitations to our system and experiment. Participants were asked to conduct searches as part of an artificially constructed group and did not meet or directly interact with their teammates. The searches were done over a TREC corpus of older news articles and the sessions were conducted in a laboratory. Future work could use more naturalistic settings and different corpora.

CONCLUSION

In this paper, we presented Results Space, a system to support small groups working on collaborative searches. We described the development of an asynchronous task scenario designed to elicit a range of behaviors involving the collaborative features and presented results from a laboratory study of the system. Our results suggest that collaborative search systems need to support a variety of collaboration and information seeking strategies and illustrate the importance of evaluating CIS systems using flexible scenarios that allow participants to engage in natural behaviors.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation grant IIS 0812363. We thank Gene Golovchinsky and Chirag Shah for helpful discussions.

REFERENCES

Bailey, E. W., Kelly, D., & Gyllstrom, K. (2009). Undergraduates' evaluations of assigned search topics. Proceedings of the 32th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '09), Boston, Massachusetts, 812-813.

Capra, R., Marchionini, G., Velasco-Martin, J., & Muller, K. (2010). Tools-at-hand and learning in multi-session, collaborative search. In Proc. CHI 2010. ACM, New York, 951-960.

Capra, R., Velasco-Martin, J. and Sams, B. (2011). Collaborative information seeking by the numbers. In Proc. of the 3rd International Workshop on Collaborative Information Retrieval (CIR '11). ACM, New York, 7-10.

Kules, B., and Capra, R. (2012). Influence of Training and Stage of Search on Gaze Behavior in a Library Catalog Faceted Search Interface. Journal of the American Society for Information Science and Technology, 63(1): 114-138.

Morris, M. R., & Horvitz, E. (2007). SearchTogether: an interface for collaborative web search. In Proc. UIST 2007. ACM, New York, 3-12.

Morris, M. R. (2008). A survey of collaborative web search practices. In Proc. CHI 2008. ACM, New York, 1657–1660.

Morris, M. R., & Teevan, J. (2009). Collaborative Web Search: Who, What, Where, When, and Why. Synthesis Lectures on Information Concepts, Retrieval, and Services #14. Morgan & Claypool.

Paul, S. A., & Morris, M. R. (2009). CoSense: enhancing sensemaking for collaborative web search. In Proc. CHI 2009. ACM, New York, 1771–1780.

Pickens, J., Golovchinsky, G., Shah, C., Qvarfordt, P., & Back, M. (2008). Algorithmic mediation for collaborative exploratory search. In Proc. SIGIR 2008. ACM, New York, 315-322.

Golovchinsky, G., Diriye, A., Pickens, J. (2011). Designing for collaboration in information seeking. In Proc. Fifth Workshop on Human-Computer Interaction and Information Retrieval (HCIR 2011), October 20, 2011, Mountain View, CA.

Shah, C., & González-Ibáñez, R. (2010). Exploring information seeking processes in collaborative search tasks. Proceedings of the American Society for Information Science and Technology (Vol. 47, pp. 1–7). Silver Spring, MD, USA: ASIS&T.

Shah, C., & González-Ibáñez, R. (2011). Evaluating the Synergic Effect of Collaboration in Information Seeking. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information (SIGIR '11), 913-922.

Shah, C., & Marchionini, G. (2010). Awareness in collaborative information seeking. Journal of theournal of the American Society for Information Science and Technology, 61(10), 1970-1986.

Shelby, J., and Capra, R. (2011). Sensemaking in Collaborative Exploratory Search. In Proceedings of the 74th American Society for Information Science and Technology Annual Meeting (Vol. 48, Poster 73).

Voorhees, E. M. (2006). Overview of the TREC 2005 Robust Retrieval Track. Proceedings of TREC-14.