Pitt@TREC08: An Initial Study of Collaborative Information Behavior in E-Discovery

Zhen Yue, Jon Walker, Yi-Ling Lin, Daqing He*

School of Information Sciences

University of Pittsburgh

Abstract

The University of Pittsburgh team participated in the interactive task of Legal Track in TREC 2008. We designed an experiment to investigate into the collaborative information behavior (CIB) of the group of people working on e-discovery task provided by Legal Track in TREC 2008. Through the study, we identified three major characteristics of CIB in e-discovery. 1) Frequent communication among participants 2) division of labor is common; and 3) "awareness" is important among participants. Based on these insights, we also propose a set of essential technologies and functions for retrieval systems that support CIB.

1.1 Collaborative Information Behavior

It is well recognized that people act in a social and organizational context when trying to solve information seeking problems (Karamuftuoglu, 1998; Munro, Hook, & Benyon 1999; Soininen & Suikola, 2000). However, in most collaboration environments, information behavior is still commonly perceived as operating at the individual level (Sonnenwald & Pierce, 2000). Therefore, in order to understand aspects of collaborative activities from an information seeking and retrieval perspective, we need to investigate the manifestations of collaboration (Hansen & Jarvenlin, 2005).

Although no definition of collaborative information behavior (CIB) has been universally accepted, for the sake of our discussion, we adopt a definition of CIB as "activities that a group or team of people undertake to identify and resolve a shared information need" (Poltrock, Dumais, Fidel, Bruce, & Pejtersen, 2003). Two important aspects of CIB are revealed in this definition. The first is collaboration where people working together to seek information. The second aspect is about resolving one common information need, which includes seeking, retrieving, and using information to solve a problem (Reddy & Jansen 2008). This definition helps us to establish that CIB is as common and natural as individual information behavior (IIB) (Hansen & Jarvenlin, 2005; McKenzie, 2003; Talja, 2002)

CIB studies have been conducted in many different settings, including academia, industry, medicine, military and everyday life. Talja (2002) conducted a study through empirical observation and showed that there were different types of information sharing in academic communities. A study of two teams engaged in the design of computer-related products focused on how team members collectively sought and shared external information acquired within the team (Poltrock, Grudin, Dumais, Fidel,

* Corresponding author: email: dah44@pitt.edu

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Pitt@TREC08: An Initial Study of Collaborative Information Behavior in E-Discovery	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Pittsburgh,School of Information Sciences,Pittsburgh,PA,15260**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Seventeenth Text REtrieval Conference (TREC 2008) held in Gaithersburg, Maryland, November 18-21, 2008. The conference was co-sponsored bythe National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**10**	

Bruce, & Pejtersen, 2003). In a study of information behavior in military command and control teams, Sonnenwald and Pirerce (2000) studied collaboration in dynamic situations with rapidly changing information and a need for continuous information exchange. Through a study of CIB of two healthcare teams, Reddy and Jansen (2008) proposed a model for understanding CIB in context. Within everyday life information seeking (ELIS) studies, McKenzie (2003) found that people routinely assist each other in solving information problems.

To the best of our knowledge, there are few studies of CIB that have been conducted in a legal search or related setting. Therefore, our study presented in this paper, which was conducted under the e-discovery task in Legal track of TREC 2008, will be an important contribution to the CIB literature in legal domain.

1.2 E-Discovery and TREC Legal Track

Discovery of the relevant information gathered about a topic in dispute is at the core of the litigation process (Hickman, 1947). With the help of modern information technology and the common practices of using digital documents, the discovery process need to identify and produce relevant information has evolved from the manual review of paper documents to the one involving vast volumes of electronic documents (Baron, 2007).

It is under this circumstance that TREC Legal track started in 2006. The legal track has two tasks. The ad hoc task is designed to facilitate comparison of retrieval systems, whereas the other task – the interactive challenge task – was to model more realistically the way in which retrieval queries might be generated, refined, and applied in the e-discovery domain [TREC-2008 legal track guidelines].

In real e-discovery processes, there is a lead attorney who is in charge of overseeing a large document-review effort and for vouching for the completeness and accuracy of the produced collection. The attorney often hires an e-discovery firm or team to gather all the relevant documents from the full document collection implicated by the matter. The design of interactive challenge task resembles this situation, and the research team involving in the task acts as the hired e-discovery firm, and the track organizers act as the lead attorney to provide authority on the issues related to the relevance and the scope of the discovery process.

Therefore, e-discovery tasks in general, and the interactive challenge task in specific, are challenging and complex information seeking tasks that requires a group of people to collaborate. The goal of our participation in this year's TREC Legal Track, therefore, is to investigate CIB in e-discovery, and generate a set of insights into CIB, and a set of recommendations for tools that truly support CIB in e-discovery and in information seeking in general.

2 Methodology and Experimental Setting

2.1 Tasks and Experiment Procedure

Our investigation of CIB activities in e-discovery involves a group of people working on the e-discovery tasks provided by Legal Track. The objective of each task is to find all the relevant documents from a Tobacco document collection. The Legacy Tobacco Documents Library (LTDL) contains more than 9.7 million documents (50+ million pages) created by major tobacco companies

related to their advertising, manufacturing, marketing, sales, and scientific research activities. The participants used the search tool provided by LTDL[1] to search for relevant documents.

We designed our study to have two phases. In the first phase, two participants worked on the following e-discovery topic from the TREC-2008 Legal Track Interactive task:

> *Documents referring to marketing or advertising restrictions proposed for inclusion in, or actually included in, the Master Settlement Agreement ("MSA"), including, but not limited to, restrictions on advertising on billboards, stadiums, arenas, shopping malls, buses, taxis, or any other outdoor advertising*[2].

In this first phase, the participants worked separately in different places, and there was no time limitation of finishing the task. Participants could do whatever they want to complete the task.

It was in the second phase, the same participants worked collaboratively on a second e-discovery task from the TREC-2008 Legal Track Interactive task which follows:

> *All documents which describe, refer to, report on, or mention any "in-store," "on-counter," "point of sale," or other retail marketing campaigns for cigarettes*[3].

The study in this phase was structured to observe and document the types of interactions between the participants when working as a team. Specifically we were interested in the flow of information between the two participants which contribute to accomplishing the team's task. The two participants were working at their own desks which are approximately two feet apart, facing each other. In order for one participant to see the other's computer screen, the screen had to be turned around (easily done) or the person had to walk around to that side of the line of desks and look at it. No one else was working in the office when the study was taking place, so verbal communications also played a large role in the information exchange process.

Our experiments actually focused on the second phase which involved CIB. However, phase 1 was important because it allowed participants getting familiar with the e-discovery tasks and the collection. It also helped the participants make comparison to the experience between IIB and CIB.

2.2 Participants

Three people participated in the experiment. Two of them worked as the actual searchers. They were PhD students majoring in Information Science. Both of the searchers were experienced in accessing information through a search engine, however, neither of them had experience in an e-discovery environment. The third participant had more knowledge about the e-discovery task. He did not conduct search directly, but acted as a topic authority. Whenever the searchers had questions about the task, they would ask him for answers or confirmations. The third person also acted as an observer during the

---

[1] http://legacy.library.ucsf.edu/

[2] TREC-2008 legal track interactive task topics

[3] TREC-2008 legal track interactive task topics

experiment. He took notes about the actions of the participants and any issues he found significant during the experiment.

2.3 Qualitative Data and Analysis Method

In this study, we used qualitative method to analyze the experiment data. The data include the notes that participants took and the reports recorded from a focus group study after the experiments.

After the experiments, the participants were asked to write an account of the whole experiment process, in which the following areas were to be addressed: how did they complete the task step by step and what kind of issues they had in each step? How did they deal with those issues? How did they feel about the difference between individual and collaborative task? How did they collaborate with each other and was there any tool help them to collaborate?

At the same time, the participants and the observer formed a focus group with regular weekly meeting. The goal of the focus group was to review the notes, and initiate discussions to review any particular point in the experiment and got response from each other. In this way, a better understanding of various behaviors in the experiment was obtained. The outcomes of the focus group were recorded for the qualitative analysis.

3 Data analysis and Findings[4]

Overall we found that the participants were more satisfied with the completion of collaborative task. Since e-discovery is a complex information seeking task, no individual participant could have comprehensive knowledge of it. That's why the collaboration made the completion of task more efficient and effective and the participants felt more confident and satisfied with the collaboration task result.

Through the study, we also identified three major characteristic of CIB in e-discovery. These are 1) frequent communication is an essential component of CIB; 2) the division of labor is common in the collaborative task of e-discovery; and 3) it is important for collaborators to keep an "awareness" of each other's activities to make sure the collaboration goes well. In the remainder of this section, we will talk about the findings in these three characteristics of CIB in details.

---

[4] As we were preparing this version of the paper for submission, we received the results from our study from TREC. In the individual phase, our two participants identified 4,505 documents and 25,816 in the joint phase. Recall increased from 0.7% to 2.0% and precision decreased from 86.6% to 70.0% between the 1st and 2sd phases. It would be very difficult to attribute any difference in recall or precision to the joint as opposed to individual searching. In this particular case, the learning effect from the individual trial to the joint trail cannot be ignored. (Campbell, 1963)

3.1 Communication

Frequent communication was common between the searchers as well as between the searchers and the "authority"[5]. For searchers, the content of their communication varied from search strategies to keywords and queries and information about relevancy. There were two forms for this type of communication, including talking face to face and chatting via instant messenger.

> *At the beginning of collaboration task, we sat together and discussed about what strategy should be used for the task. During the search, we exchanged keywords that we found should be included in the query via instant messenger*[6].

The communication between searchers was often triggered by a need to share information. We observed two types of such needs. One is that an individual did not have enough information to continue with the task, and needed to obtain more information to continue. Discussion about search strategy at the beginning of the search task, for example, was an instance of this type of needs. Both searchers had some ideas about the strategy but they want to communicate with each other to make sure they choose the best one.

The other type of need is that one searcher had obtained a valuable piece of information and wanted to share with the rest of the team. Exchanging keywords was an instance of the second type of needs. Keywords were often exchanged during our experiments. Another example was to share the search tactics that had been demonstrated to work well. After the end of the first session and before the beginning of the second session, one participant emailed the others in the group about a 'better' way to define the queries being used to determine the quality of the documents being added by the additional keywords. The route being used was add the query term to the query and then compare the counts of documents returned and to see if any new documents were more likely relevant or not. The recommended change was to add in an 'AND NOT' followed by the old query which would leave only the newly returned documents to scan. This would have made the evaluation of the new set of documents easier. However, the other participant did not pursue this suggestion. This may reflect that email was not an appropriate channel to send such information. A live session showing how this new query structure would facilitate the process may have lead to its being adopted.

For the communication between the searchers and the authority, the content was often about the relevancy of certain document. For most of the time, the form of this type of communication was the searcher asked questions and the authority provided her with answers.

> *When assessing the relevancy of the document, one searcher was not sure whether one document talking about advertisement research was relevant or not. She asked the authority to take a look at the document. After examining the document for a while, the authority told*

---

[5] The "authority" in this context was not the TREC's topic authority, but the observer.

[6] Searcher's note

her that this document was relevant. Also, the searcher would ask the authority for his opinion about whether certain keywords should be included in the query[7].

This communication was probably triggered by a lack of confidence. When the searcher lacked confidence about the relevancy of documents; she would ask the authority for help. Another example is that when a searcher found a keyword in the document but she was not confident whether this keyword should be added to the query, she would ask the authority for his opinion instead of sharing with the other searcher.

3.2 Division of Labor

Although in the second phase of the experiment the participants were willing to collaborate, they did not want to spend all the time together working on the task. Sometime they would rather divide the labor and work alone. We find that whenever there was a need to interact with the search engine system, the searchers wanted to work separately on two machines. It was clear that each individual wanted to have full control of any physical devices or the system.

There are two ways that the searchers divided their labor. The first one was to split the task. Following is an instance of splitting the task:

The two searchers agreed on their preliminary query and decided to divide up the returned documents in order to see what percentage of documents were relevant and what additional potential query terms could be gleamed from the relevant documents. They decided that one participant would view documents from screens 1, 3, 5, 7, 9 and the other would look at screens 2, 4, 6, 8, 10 and at this point they basically started working in parallel[8].

The trigger of splitting the task was that an agreement had been reached on how to proceed. In the above example, the two searchers agreed on the preliminary query. At this point, both of them were very clear with the following steps. So they found a way to divide the task so that each took a half.

In contrast, the second way of division of labor without splitting the task was due to a disagreement. Here is an example:

The two searchers began to work on two machines to find out how to incorporate 'cigarette' into the search, as 'cigarette' or as 'cigarette'. One searcher started looking at the help of the search engine. The other searcher tried the both searches using the two formats. It turns out 'cigarette*' was including a lot of documents in German. Finally they decide to use "cigarette or cigarettes" instead of "cigarette*"*[9].

---

[7] Report of focus group

[8] Observer's note

[9] Observer's note

The above example shows that the two searchers came to a point where they couldn't reach an agreement on how to proceed. This was how division of labor without splitting task was triggered. There wasn't any explicit division on the task for who should work on which part. The two searchers took actions according to their own way of understanding the problem. They wouldn't reach an agreement until one of them showed the other the evidence to convince the other.

Morris (2008) describes the results of a survey study of different techniques that her subjects used when involved in joint search tasks. Morris identified two strategies commonly used, "divide-and-conquer" where the task is partitioned and divided up and the participants are involved in parallel but different sub-tasks and 'brute-force' where participants are working in parallel on the same sub-task. This second strategy sometimes leads to a competitive situation where it becomes a race to find the object of the search first or the most interesting object. The two scenarios that we found resemble Morris' two strategy, but our results tell that when there is an agreement, a "divide-and-conquer" strategy would be adopted, whereas when there is a disagreement, "brute-force" strategy would be used to compete for convincing evidence.

In fact, the division of labor can sometimes come from frustration. After working on one machine together, and moving the keyboard and mouse back and forth between them for a while, one participant thought that they could not make any further progress. So after a statement "This isn't working. Why don't you work on your computer….", the participants separated and worked on their own tasks. The "This isn't working." seems to imply that a level of frustration had been reached that made working separately seem more attractive than continuing to work jointly. This would seem then to indicate that there needs to be an additional qualification to the task – tool division. Because this tool was working appropriately for the task until that level of frustration was reached. And when the participants started working separately, they were using the same tool as before the separation doing the same tasks. Now, what could have been the issue at this point could have been no more than individual preference: one participant's preference was to continue exploring what happens whereas the other participant wanted to pursue a course which was going to provide more immediate payback. Obviously, we cannot determine all the causes of the two participants stopping working on the same machine and starting working separately from the data we collected.

We also find that trust is an important factor that affects the division of labor. The two searchers didn't always trust each other during the whole process. After they had split the task into parts, they trusted each other's relevance judgment and wouldn't check it again because they had reached an agreement before splitting the task. However, when they distrusted each other on certain issue, they could not divide the labor by splitting the task. In this circumstance, they would try whatever method that they thought could solve the problem until one of them found good evidence to make the other person agree.

3.3 Awareness

In the study, we find that maintaining an awareness of each other's activity is essential for the collaboration. The activities that are related to awareness include other participant's interaction with

the system and the communication with the rest participants. Following is a problem caused by lacking of awareness of other's interaction with the system:

> *The search engine that they were using allowed the user to specify the numbers of items you wanted to include on each screen. One user had changed her number to 50 the maximum allowed and the other user was using the default value of 10. When they decided that one searcher would view documents from screens 1, 3, 5, 7, 9 and the other would look at screens 2, 4, 6, 8, 10, they actually don't have the same document on each page*[10].

Fortunately, they realized the problem very soon and changed the setting to the same number. This problem could be avoided if the participants had a tool that supported awareness of each other's activity. Actually, there were several cases that one person had to turn her screen to the other person to show her what was going on with the system. In this way, they maintained an awareness of each other's progress. Besides maintaining an awareness of other's interaction with the system, awareness of other's communication is important too. In our experiment, the authority had been asked the same question by both searchers. If one searcher knew the answers of the questions asked by the other searcher, she didn't need to ask that question again.

The participants became aware of an additional relationship during sharing of keywords. During the experiments, each participant opened a file and was keeping track of keywords she wanted to add to the query. Whenever a keyword was added to a participant's list, it was also messaged to the other user. This activity in real time lead to creating a norm for how fast one should be adding keywords to the query. Although there was no discussion along the lines of "You're adding lots more terms than I am.", one participant acknowledged that the rate at which the other user was adding keywords lead her to make up her mind about a document being relevant or not faster than she had been doing and to be more lenient in deciding if a term deserved to be added as a key word or not.

4 Suggestions for designing collaborative retrieval systems

Through our study, we find that collaboration is necessary and common for complex information seeking task like e-discovery. Although the participants in our study were physically co-located in one room, certain insights still can be drawn for retrieval systems that support completing information seeking tasks collaboratively. We think that collaborative information retrieval (CIR) systems should support the following collaboration information behaviors and have the described essential features:

Verbal communication: Our study indicates that verbal communication plays an important role in the collaboration. It allows participants to interact with each other and get feedbacks immediately. Our study also reveals that verbal communication is more frequent at the beginning stage when participants discuss about the search strategy and information need.

Text exchanging: Sending text message is another important form of communication among participants. Some information flow could not be well described by verbal. For example, participants in

[10] Report of focus group

our experiment sometimes need to sending keywords and Boolean queries to each other. To avoid any error, they would rather send this kind of information via instant messenger. However, we do find some limitation of current instant messengers. Most current instant messengers has a size limit to each message, whereas we find that the Boolean queries exchanged between participants during the experiments were often cut off because of the size limits. It caused inconvenience and sometimes errors. Supporting unlimited text exchanging is necessary for CIR systems.

Synchronous collaboration: When people work collaboratively on the same task, they do need to spend some time together and work on things in synchronous mode. In our experiment, it was easy for the participants to collaborate synchronously because they were in the same room. When they discussed the search strategy, they brainstormed keywords and wrote the keywords on a piece of paper. Using the paper as the share space, the participants kept on adding and removing words to change the query structure until they reach an agreement on a preliminary query. In a CIR system, the function of the paper could be implemented by using an electronic white board. People who collaborate could share this white board and edit the content synchronously. Another important feature that the CIR system should support in synchronous collaboration mode is screen sharing. Supported by this function, participants could make the same screen displayed on the monitors of all participants. This is useful when some important information has to be shared visually with the rest of the team. However this feature should be used cautiously. Sometimes people don't want to be interrupted. Therefore, permission should be asked before this feature takes control of all the screens.

Asynchronous collaboration: Besides synchronous collaboration, CIR system should also support asynchronous collaboration. As identified in our study, participants in the team sometimes need to work separately to complete different parts of the task. To maintain collaboration, they need to share various information, which includes search strategies, domain knowledge, relevant document, personal and subjective opinions and so on. A function of supporting various types of information sharing is necessary and important for CIR system. The shared information should be well organized and easily be located by collaborators.

5 Conclusion

In this report, we present an initial study of collaborative information behavior in the context of TREC Legal Track. Through participating in the interactive challenging tasks, we studied various aspects of CIB by simulating e-discovery tasks. We obtained many interesting and important insights about CIB. First, communication is frequent and an essential component of CIB. Second, the division of labor is common in the collaborative task of e-discovery. Third, it is important for collaborators to keep "awareness" of each other's activities to make sure the collaboration goes well.

Based on these insights, we also propose a set of essential technologies and functions for retrieval systems that support CIB. Collaborative information retrieval technologies should support collaborative information behaviors including verbal communication and text exchanging. More importantly, a well designed CIR system should support both synchronous collaboration and asynchronous collaboration.

Reference

Baron, J. (2007) The Sedona Conference® Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery. *The Sedona Conference Journal*, 189-223.

Campbell, D.T. & Stanley, J.C. (1963), Experimental and Quasi-Experimental Designs for Research, Houghton Mifflin, Boston.

Hansen, P. and K. Jarvelin (2005). Collaborative Information Retrieval in an information-intensive domain. *Information Processing & Management*, 41(5): 1101-1119.

Hickman T. (1947). Mutual knowledge of all the relevant facts gathered by both parties is essential to proper litigation.

Karamuftuoglu, M. (1998). Collaborative Information Retrieval: Towards a Social Informatics View of IR Interaction. *Journal of the American Society for Information Science*, 49(12): 1070-1080.

Reddy, M. C. & Jansen, B. J. (2008) A model for understanding collaborative information behavior in context: A study of two healthcare teams. *Information Processing and Management*, 44(1): 256–273.

McKenzie, P. (2003). A model of information practices in accounts of every-day life information seeking. *Journal of documentation*, 59(1), 19-40.

Morris, M. R. 2008. A survey of collaborative web search practices. In *Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy, April 05 - 10, 2008). CHI '08. ACM, New York, NY, 1657-1660.

Munro, A. J. Höök, K. & Benyon, D. (1999). (Eds). Social navigation of information Space. *Springer*, London.

Soininen, K. & Suikola, E. (2000). Information Seeking is Social. Proceedings of the First Nordic Conference on Computer-Human Interaction, *NordCHI* 2000, (pp. 1-9), 23-25 October 2000, Stockholm, Sweden: Royal Institute of Technology.

Poltrock, S., Dumais, S., Fidel, R., Bruce, H., & Pejtersen, A.M. (2003). Information seeking and sharing in design teams (pp. 239–247).Paper presented at the ACM conference on supporting group work (GROUP'03), Sanibal Island, FL.

Sonnenwald, D. H. & Pierce, L.G. (2000). Information behaviour in dynamic group work contexts: interwoven situational awareness, dense social networks and contested collaboration in command and control. *Information Processing and Management* 36(3):461-479.

Talja & Hansen, (2006) New Directions in Human Information behavior. Chapter 7: Information Sharing: 113-134.