A Model for Understanding Collaborative Information Behavior in E-Discovery

Zhen Yue, 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 2009. We designed an experiment to investigate into the collaborative information behavior (CIB) of the group of people working on e-discovery tasks provided by Legal Track in TREC 2009. Through the studies, we proposed a model for understanding CIB in e-discovery.

1   Introduction

It is common for people to collaborate when dealing with complex problem. In information seeking environment, collaboration is important and useful when the task is too complex for a single individual to tackle. E-discovery can be viewed as such type of complex information seeking task that requires the help from collaboration. 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 task in legal track resembles this situation. Therefore, we conducted our study based on the interactive task in order to investigate the manifestation of collaboration in e-discovery. In TREC legal track 2008, we identified several major characteristics of CIB in e-discovery. The goal of our participation in this year's TREC Legal Track is to explore the essence of collaboration in information seeking process and then toward building a model for understanding CIB in e-discovery.

2   Methodology and Experimental Setting

2.1   Collaborative Experiment Design

Shah (2008) proposed a four layers model for collaborative information seeking. His model incorporates two users who can access and organize information individually, or decide to collaborate with each other. Each of the users can use the search tool to access the collection and obtaining results. In the mean time, they could use the collaborative tools to communicate with each other or they may have a common or shared interface. Also, they have a shared space where they can store and organize

---

[*] Corresponding author: email: dah44@pitt.edu

their results. This model represents a typical scenario of collaborative information seeking.

In order to apply this model in the e-discovery scenario, we modified it to incorporate three users. Although Shah claimed that his model can be extended to incorporate any number of users, this third user we add into the model is actually a different type of user. It represents the topic authority in e-discovery while the other two users are actually searchers. The topic authority will share information need with the searchers and communicate about relevancy information. However, the topic authority won't conduct search himself. Therefore, this third user we added to the model doesn't have access to the search service. Nevertheless, this third user plays an important role because he/she defines the information need and has the right to judge the relevancy of the results.
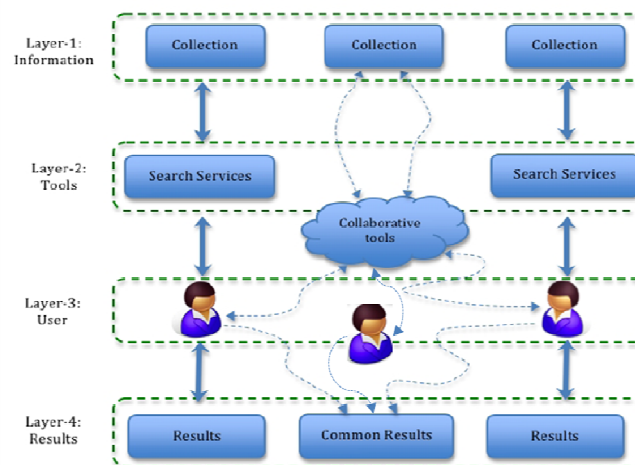

Figure 1 Collaborative Information Seeking Model in E-Discovery

2.2 Tasks and Experiment Procedure

Our investigation of CIB activities in e-discovery involves a group of people working on the e-discovery task provided by Legal Track 2009. The objective of the task is to find all the relevant documents from Enron email collection. The Enron collection is a collection of email messages, with attachments, produced by Enron in response to requests by the Federal Energy Regulatory Commission. In our experiment, the participants worked on the following e-discovery topic from the TREC-2009 Legal Track Interactive task:

> *All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in structured commodity transactions known as "prepay transactions"*[1]

We designed our experiment to have several sessions. There is no time limit for the participants to finish the task. They could do whatever they could to complete the task. Once they feel that they have reached certain point and wanted to stop, we would call it a session. The participants could use as many sessions as they want until they could finally complete the task. In the end, it takes our

---

[1] TREC Legal Track 2009 Topic 201

participants four search sessions and one final session of wrapping up the results to complete task. Each search session last about two hours and the wrapping up session took about half an hour. Throughout each session, we observed the behaviors of our participants and took notes about the actions of the participants and any issue we found significant during the experiment. At the end of each session, the participants worked collaboratively to write the subtasks they had done during the session and then filled out questionnaires separately about their satisfaction to the search experience and search results. Also the participants were asked to answer several open-ended questions about the details of conducting the task and their feelings and thoughts involved after each session. The experiment was structured to observe and record the types of interactions between the participants when they worked as a team. Specifically we were interested in the flow of information between the two participants which contributed to accomplishing the team's task.

2.3 Participants

In our experiment, the topic authority is assigned by the track organizer based on the topic we chose. Besides the topic authority, we recruited two participants to work collaboratively as searchers in the experiment. Surowiecki (2004) emphasized that one of the requirements for a successful collaboration is the "diversity of opinion", which means that the participants of a team should come with different background and expertise. Therefore, we recruit one participant from the law school who is an expert with legal background and the other participant from school of information science who is an information retrieval expert. Both of the searchers were familiar and had experiences in e-discovery.

2.4 Tools

According to our collaborative experiment design, participants will be provided with a search tool and a set of collaborative tools. The search tool we use is "email explorer", which is a search tool specially designed for email searching with graphic user interface.[2] There are two collaborative tools used in this experiment. One is Skype and two functions are used: 1) users can send and receive text messages with unlimited size; 2) a screen sharing function is available for user to show what is going on in their screen to the other user. The other collaborative tool is wiki. It is a space for users to sharing search strategies, results and other information that the users want to post and share. The two participants were in the same room, back to back, about 10 feet apart. Verbal communication is available between them. However, for the interaction with the topic authority, they had to use the conference phone. Each of the searchers has a computer and they have full control of their own physical devices such as keyboard and mouse.

---

[2] Developed by UMD

2.5 Data Analysis Method

In this study, we used both quantitative and qualitative method to analyze the experiment data. The quantitative data are those collected in post-session questionnaires about the satisfaction on search experience and search results. Qualitative data includes the participants' answers to open-ended questions and the reports recorded from a focus group study after the experiments.

After the whole experiment, the participants and the observer formed a focus group. The goal of the focus group was to review the experiment, discuss 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

Overall the participants thought that the e-discovery task is a complex information seeking task for which collaboration is necessary. They claimed that they discovered more than they would have had if they worked entirely separately, which makes it a good collaboration. However, the participants were not completely satisfied with the final results because they didn't have access to the attachment (limitation of the search tool) which greatly limits the number of documents produced for the final results. They also mentioned that if they had access to the attachment, then it wouldn't be a task that they could finish within 10 hours for two people. It may require much more time and even more people involved in the collaboration in order to complete the task effectively and efficiently.

In last year's experiment that we designed for Legal Track 2008, we identified three major characteristic of CIB in e-discovery. Those 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 this experiment, besides these characteristics which we still found existing and important, we also focused on other aspects that reveal the essence of collaboration throughout the whole information seeking process.

3.1 Subtasks Analysis

In total, there are four experiment sessions for information seeking and one last session for final results wrapping up. In the first two sessions, the two participants got familiar with the e-discovery task and the background information by reading the materials[3] and conference phone call with the topic authority. They explored the collection and consulted the topic authority to make sure that they had a comprehensive understanding of the information need and chose the right strategy and path. At

---

[3] Background information provided by Legal Track 2009 and the topic authority

the end of the second session, they developed two search strategies. Based on the communication with the topic authority, they thought both of the strategies are correct and they decided to pursue independent paths. Therefore, in the third and fourth session, each of the participants focused on their own path and tried their best to produce results in good quality. In the final stage, the two participants worked together to combine their results.

At the end of each session, we asked our participants to summarize the subtasks they had been doing during each session. For each subtask, they need to identify whether it's an individual subtask or collaborative subtask and then assign complexity level (easy, medium or difficult) to each subtask. For example, "reading prelim documents on the topic and notes from topic authority" is recognized as an easy individual subtask. "Looking for special purpose entities" is another individual subtask of which the complexity level is difficult. "Conference phone call to topic authority" is an instance of easy collaborative subtask. For complex collaborative subtask we have "Setting up new search approaches for relevant documents". The distribution of subtasks according to sessions is indicated in the following table:

Table 1 Subtasks distribution

Subtask Type		Session 1	Session 2	Session 3	Session 4	Final Session
Individual Subtasks	Easy	2	1	1	0	0
	Medium	0	1	2	4	0
	Difficult	1	0	2	2	0
Collaborative Subtasks	Easy	2	2	0	0	1
	Medium	1	0	0	0	1
	Difficult	0	1	0	0	0
Total		6	5	5	6	2

From the table, we can see that all the collaborative subtasks are in the first two sessions and the final session. Most of the individual subtasks are in the third and fourth session. Although there are no collaborative subtasks indentified in session 3 and 4. The communication between participants were still frequent as they constantly contribute to each other with valuable information. The subtasks were relatively easy in the first two sessions because they were working with the general characteristics of the task. In general the concept behind what was happening was difficult. However, once they started to have to make decisions about was this or that document exactly dealing the issue, it became more difficult. This wasn't a collaborative issue but a judgment call about what they thought the topic authority wanted. Additionally, as they exploring the collection more and more in the later sessions, they found that there were indications that there were many more relevant documents in the collection, but they didn't have the information needed to evaluate them (i.e., can't access to the attachments). This makes them feel the subtasks becoming more difficult.

For each subtask, we asked the participants to evaluate their satisfaction on the result and search experience respectively on a 1-10 scale. ANOVA analysis was performed on result satisfaction (RS) and experience satisfaction (ES) respectively as a function of subtask type (individual, collaborative) and task complexity level (easy, medium, difficult).

We found a significant effect of subtask type on result satisfaction ($F(1,28)=7.83$, $P=0.009$).

Similaryly, there was a significant effect of subtask type on experience satisfaction ($F(1,28)=4.22$, $P=0.049$). Generally speaking, the participants are more satisfied with the result and experience when conducting collaborative subtasks than individual subtasks. we can see in figure 2 that the result satisfaction and experience satisfaction for collaborative subtask is higher than that of the individual subtasks. In particulaly, the score difference for difficult subtasks is the biggest. This indicates that for complex task, collaboration produce much better result and experience.
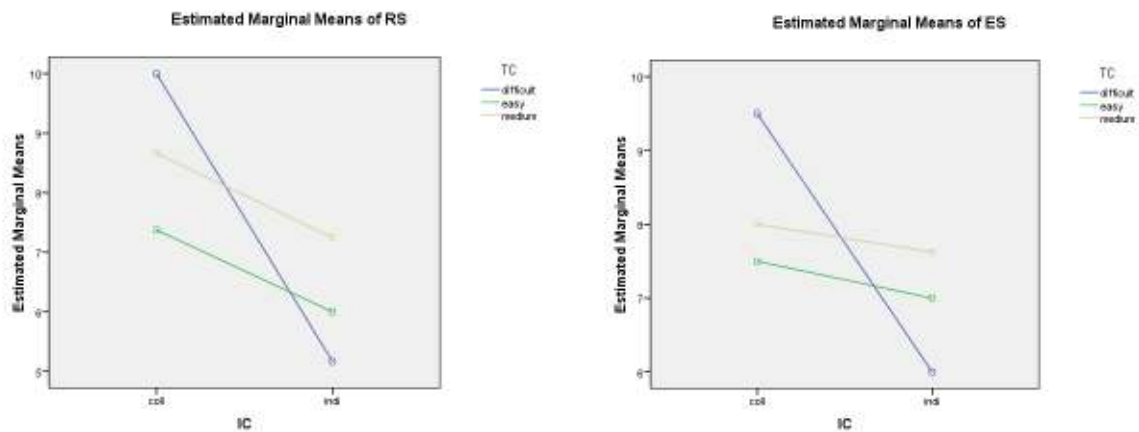

Figure 2 Interaction between subtask type and complexity level

3.2 Cognitive load and Collaboration cost

In general, the e-discovery task is a complex information seeking task that requires collaboration in order to be complete efficiently and effectively. No individual participant could have comprehensive knowledge of it. That's why the collaboration made the completion of task efficient and effective and the participants felt more satisfied with the collaborative results and experience. One participant mentioned that the collaboration made the learning of search lessons twice as fast. No matter either one learned anything first, the other one could be benefit by just knowing the fact. Also, the different background of the two participants contributes to the success of collaboration. For example, one of them is more familiar with legal terminologies and it is easy for him to come up with synonyms. The other participant felt that those legal terminologies helped him a lot during the search. Also, the participant who is an information retrieval expert using network analysis to assist the information searching and this strategy inspired the other participant to produce more relevant results. However, in spite of all these advantages, we still need to pay attention to the extra cognitive load and collaboration cost compare to individual information seeking. As in a collaborative environment, the tools are more complex and information flow is not just between system and users, but also between users.

Adaption to the tools. Many researchers have attempted to address the issue of cognitive load induced by a system (Fidel etc, 2004). Compare to individual information seeking where user only need to interactive with the search tool, in collaborative inforamtion seeking, they need to interact with the collaboration tools as well. In our experiment, the collaboration tool we selected are Skype and Wiki which our participants were familiar with already. However, knowing how to use a tool does

not necessarly means there is no extra cost adopting the tool in the task. Our participants pointed out a disadvantage of the tools that there were not enough screen space for all the search tools and collaboration tools. Sometime they focus on one screen and totally neglected what was going on with the other tools. A very interesting phenomenon in our experiment is that, for the first 2 sessions, the two participants use Skype to send and receive messages a lot. However, in the following sessions, the usage of Skype text exchanging became less and less. This means our participants didn't adopt Skype in a long run as they discovered that Skype actually slow down their communication pace.

Collaboration between searcher the topic authority. This is a type of collaboration which is unique to an e-discovery scenario. In a typical individual information seeking screnario, the information need is often initiaed by the searcher himself/herself. In a e-discovery task, the inforamtion need is not proposed by the searcher but the topic authority. Therefore, there will be an extra cost of "transferring" the information need. This is a very important step for the whole task as it decides the success of the following search steps. In our experiment, because the topic authority could only be remotely accessed by phone calls, sometime the searchers felt that there were not enough in time support form the topic authority. At the initial stage, the discussion with topic authority is very helpful for them to understand the task requirement. However, in the actual search process, they felt that they could make better use of the support from the topic authority for relevancy judgment if the support can be accessed immediately.

Collaboration between searchers. In individual inforamtion seeking, one only need to focus on his/her own activity. However, in collaborative inforamtion seeking, there will be several sources of extra cognitive load: (1) contributing to each other (sending or posting important messages that they think their partner should know), (2) keeping aware of what the other people is doing (3) pay attention to the group history in addition to the individual search history. While these activities may be the key benefitial of collaboration for the completing of task, they are still extra cost for users. Therefore, a balance between the collaboration benefitial and cost is crucial to the success of collaboration. Also, there is an issue of get familiar with each other. As I mentioned before, at the beginning, our two participants using text to chat a lot, vebal communication is very rare. Later, the verbal communication occurs more and more frequently and evently almost replaced the texting. This means there was a process for our participants to getting familiar with each other and gainning trust. Although in our experiment, the unfamiliarness does not cause untrust because they believe each other's expertise in their own field, this may not be the case for other collaboration scenarios.

4 CIB model

Based on the subtask analysis, we proposed the following model in table 2 for collaborative information behavior in an e-discovery scenario.

We found that there are stages in the collaborative information seeking. The users often start the task with intensive collabortion. As in this stage, they need to explore the task, make sure they have the same understanding of the inforamtion need and set up search strategies. Once they are clear with the executing process of the task, a division of labor is common for the next stage as they would like to allocate subtasks and pursue paths independently. Although in the our experiment, the exploring stage and division of labor stage only occur once, in a general case, the transition of stages is not

necessarily to be one-direction. Circles may occur from the division of labor back to exploring. When users feel that they have reached certain point of completion on searching, there will be a final stage for results wrapping up when they combine their results together. During the whole process, the cognitive load of collaborators is not only reflected as adaption to the system, including searching tools and collaborative tools, but also reflected as getting familiar with other collaborators, including gaining trust and making contribution to each other.

Table 2 CIB model

Stages	Initial & Exploring ------> Division of labor ------> .... --------->Final result
Cognitive Load	--> adaption to the system ----------------------> getting familiar with other collaborators
Collaboration Levels	Collaborative ------------------> Individual ------------------> Collaborative Coordination Communication Cooperation Contribution
Collaboration Type	Synchronous --------------------> Asynchronous ----------------------> Synchronous
Subtask complexity	Easy --------------------------------> difficult -------------------------------> easy
Types of supports needed	Support for frequent communication and keeping awareness	Implicit collaborative support (recommendation) support for relevancy judgment (topic authority)	shared space for common results

Shah (2008) proposed a five layers model of collaboration: (1) communication (2) contribution, (3) coordination (4) cooperation and (5) collaboration. Each of the higher layer requires more cost than the lower layer. In our experiment, we found that different levels of collaboration exists in different stages. At the initial stage when the main goal is to explore the dataset and set up search strategies. It involves more coordination and cooperation. In this situation, sychronous collaboration support is needed as the participants need to keep awareness in time. However, when the partipants set up search strategies and decided to pursue different paths, communication and contribution played more important roles. Under such circmstance, asychorous collaboration support is needed as the participant need to focus on their own task.

In our experiment, the subtasks in initial and exploring stage were recognized as relatively easier than that in division of labor stage. This means in synchronous collaboration when collaborators are

making decisions together, the subtasks are easier to complete as long as the communication and awareness are well supported. For asynchronous collaboration, when user need to make judgment call by himself/herself more often, the subtasks are harder to complete and they could really make better use of extra implicit collaboration support like recommendations and support for relevancy judgment. In the final stage for result wrapping up, the subtasks are relatively easy as the collaborators only need to combine the results and solve controversial problem if any. In this stage, a shared space for common results is very helpful for supporting the collaboration.

5 Conclusion

In this report, we present a study of CIB in the context of TREC Legal Track. Through participating in the interactive tasks, we designed a collaborative information seeking experiment to investigate the manifestation of collaboration in e-discovery. Based on the analysis of experiment results, we proposed a model for understanding CIB in e-discovery. The model incorporates collaboration type, collaboration level and subtask complexity into different stages. It reveals the essence of collaboration in information seeking process. Also, this model could be used to recognize different types of supports needed for different stages in a collaboration information seeking process.

Reference

Fidel, R., Pejtersen, A. M., Cleal, B. & Bruce, H. (2004) A multidimensional approach to the study of human-information interaction: A case study of collaborative information retrieval. Journal of the American Society for Information Science and Technology, 55(11):939-953.

Shah, C. (2008) Toward Collaborative Information Seeking (CIS). In: 1st International Collaborative Search Workshop, (pp., Pittsburgh, PA, USA.

Surowiecki, J. (2004) Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. Doubleday Publishing.