A User Study of Relevance Judgments for E-Discovery

Jianqiang Wang
Department of Library and Information Studies
Graduate School of Education
University at Buffalo - SUNY
jw254@buffalo.edu

Dagobert Soergel
Department of Library and Information Studies
Graduate School of Education
University at Buffalo - SUNY
dsoergel@buffalo.edu

ABSTRACT

This paper presents a comparative user study that investigates the relevance judgments made by assessors with a law background and assessors without. Four law students and four library and information studies (LIS) students were recruited to judge independently the relevance of 100 documents for each of four requests for production of responsive documents for litigation purposes, two requests for the tobacco Master Settlement Agreement collection and two for the Enron corporate emails collection. Both quantitative and qualitative methods are used to analyze data collected through the relevance judgment task, an entry questionnaire, and an exit interview. Being given the same task guidelines, the LIS student assessors judged *relevant* documents just as accurately as the law student assessors, while they judged *nonrelevant* documents slightly less accurately than the law student assessors. In addition, participants achieved moderate to substantial agreement on their relevance judgments. Relevance judgment speed varied significantly among participants, although on average it was about the same for the two groups. Factors influencing the accuracy and the speed of participants' relevance judgments are discussed based on preliminary analysis of qualitative data collected through the exit interviews.

Keywords
e-discovery, relevance judgment

INTRODUCTION

Legal e-discovery is the task of searching electronically stored business records such as correspondence, memos, emails, and balance sheets for documents that are relevant or *responsive* to a lawsuit or a government investigation. As more and more business documents are created and stored in digital format, legal e-discovery has become an important business sector and an attractive research area. As of December 1, 2006, the U. S. Federal Rules of Civil Procedures (FRCP) include a new category of evidence, namely, "electronically stored information" (ESI) in "any medium," intended to stand on an equal footing with existing rules covering the production of documents.

Searching electronic business records for the purpose of litigation or government investigation, however, is a challenging task. Lawyers do not want to miss any responsive documents which could be the deciding factor in the outcome of their clients' lawsuits. This means that recall is of particular importance for e-discovery. On the other hand, with the sheer volume of electronic business documents, less accurate retrieval means more labor and time are needed for culling responsive documents from non-responsive ones. This suggests that economically retrieval precision is also very important. Making the task even more challenging is the fact that the concept of "relevance" in e-discovery seems to be quite different from that in many other fields familiar to information retrieval (IR) researchers and practitioners.

In this paper, we report a comparative user study that investigates relevance judgments made by people with a law background and those without. To the best of our knowledge, this is the first comparative study looking at the effect of legal expertise on the nature of relevance judgments for e-discovery. Ultimately our goal, through this study and follow-up studies, is to more accurately define the concept of relevance and model the relevance judgment process for e-discovery, so that better system technology and search skills can be suggested and developed.

Four law students and four Library and Information Studies (LIS) students were recruited to judge independently the relevance of 100 documents for each of four requests for production of responsive documents for litigation purposes, two requests for the tobacco Master Settlement Agreement collection and two for the Enron corporate emails collection Both quantitative and qualitative methods are used to analyze data collected through the relevance judgment task, an entry questionnaire, and an exit interview. Being given the same task guidelines, the LIS student assessors judged *relevant* documents just as accurately as the law student assessors, while they judged *nonrelevant* documents

ASIST 2010, October 22–27, 2010, Pittsburgh, PA, USA.
Copyright © Jianqiang Wang and Dagobert Soergel, Department of Library and Information Studies, University at Buffalo, the State University of New York

slightly less accurately than the law student assessors. In addition, participants seemed to achieve moderate to substantial agreement on their relevance judgments. Relevance judgment speed varied significantly among participants, although on average it was about the same for the two groups of participants. Factors influencing the accuracy and the speed of participants' relevance judgments are discussed based on preliminary analysis of qualitative data collected through the exit interview. Finally, our plan of ongoing research and future work on e-discovery is outlined.

The rest of the paper is organized as follows. First, it provides background, the research questions, and a summary of previous work related to the study. Next, it describes the research questions and the study design and then the data collection and analysis results. The paper concludes by highlighting the major findings, contributions, and limitations of the study and outlines our thoughts of future work.

BACKGROUND AND RESEARCH QUESTIONS

In 2006, a new Legal E-Discovery track (termed „Legal' track in this paper) was added to the Text Retrieval Conference (TREC), an annual IR evaluation activity mainly sponsored by the U.S. National Institute of Standards and Technology (NIST). The goal of the TREC Legal track is to create a forum for lawyers, e-discovery practitioners, and academic researchers to study the capabilities and limitations of automatic search technology to support e-discovery in the legal community (Baron et al, 2006). Over the last four years, two test collections have been developed for the TREC Legal track. One collection contains about 7 million documents that were released under the tobacco "Master Settlement Agreement." The other collection, known as the Enron Email collection, contains 570K unique messages and 280K attachments of former Enron corporate emails released by the Federal Energy Regulatory Commission (FERC). Search topics, known in legal terms as "requests for production of responsive documents", were also created based on hypothetical legal complaints as the second part of the test collections. "Ground truth" relevance judgments, the third component of the test collections, were produced by voluntary assessors (lawyers and law school students) reviewing documents sampled from search results of the retrieval "runs" submitted by participating teams of the TREC Legal track.

One of the tasks of the TREC Legal track that is closely related to the work reported in this paper is the Interactive task; the first author has been a member of a participating team. A unique characteristic of that task is that a senior lawyer is assigned to each search topic as the "Topic Authority" (TA). Each participating team could communicate with the TA through email or on the phone to clarify the topic, in other words, to define the scope of responsive documents and factors that should be considered in determining the relevance (responsiveness) of documents. Based on that knowledge, the participating team then develops its own search strategies, systems, and techniques and eventually produces and submits search results for official evaluation by NIST. Another characteristic of the TREC Interactive Legal task is the use of an „appeal and adjudication' process. After the initial official relevance judgment results are released by NIST, participating teams can appeal to the topic authorities to adjudicate the relevance judgments with which they disagree. The topic authorities then review the appealed judgments and provide the final relevance judgments.

The main motivation of the study reported here originates from our experience of working on the TREC Interactive Legal tasks as well as our observation of the challenges of relevance judgments, including those of the appeal and adjudication process, for the task. We found sometimes lawyers defined and perceived document relevance quite differently than we as IR experts did, while at other times it was hard to judge whether documents were relevant or not. The complexity of defining document relevance for e-discovery is also evidenced by the fact that the initial official relevance judgments of most documents appealed by participating teams for adjudication were indeed overturned by the topic authorities (Oard et al, 2008, Hedin et al, 2009). But what has caused the incorrect initial relevance judgments remains largely unknown.

Given the TREC experience, we are particularly interested in the following questions:

1. Do assessors with a law background and/or legal service experience judge documents more accurately than assessors without that background or experience?
2. Do assessors with similar legal domain expertise judge documents similarly?
3. How reliable are the relevance judgments made by assessors with or without legal expertise?
4. What document review techniques do assessors use and what relevance criteria do they consider when making relevance judgments?
5. What is the nature of the relationship between relevant documents and the legal case? (Huang & Soergel 2006)
6. What can be done to facilitate and improve assessors' relevance judgments?

This paper focuses on describing our effort to find answers to questions 1, 2, and 3. The other questions will be considered in future work.

RELATED WORK

Much research has been done, in LIS and other fields, to define the concept of relevance and to study the criteria and techniques users use in seeking information to satisfy their information needs. For example, Cooper (1971) defined logical relevance based on a strict logical deduction relationship between a statement and a sought-after answer; Wilson (1973) defined evidential relevance and situational relevance; Barry and Schamber (1998) studied the relevance criteria used by actual information users; Wang and Soergel (1998) constructed a cognitive model of decision-making in information users' selection of documents; Huang and Soergel (2006) focused on the evidentiary connection between a piece of information and a user's question, topic, or task; Huang (2009) provides a comprehensive analysis..

Another related body of work has been done by researchers of the development of test collections for IR evaluation, with the ground truth relevance judgments made (usually) by subject experts as a key component. Cleverdon (1970) is perhaps the first who discussed the value of IR test collections. Voorhees (1998) studied the variations of relevance judgments made by different types of assessors for the TREC test collections. Despite the marked variations in the relevance judgments, she concluded the relative effectiveness of different retrieval strategies is stable. Bailey et al (1998) compared the relevance judgments made by topic experts, task experts, and people without either types of expertise. While there was some low level agreement among the three groups, they concluded that "it appears that test collections are not completely robust to changes of judge when these judges vary widely in task and topic expertise."

While providing valuable insights into the concept of relevance, criteria specifically used by actual information users and the validity of expert-created relevance judgments for IR evaluation, none of these studies directly addressed the issues of e-discovery. This is not surprising because e-discovery did not attract much attention until quite recently. Our study, while methodologically resembling some of the previous studies of relevance, contributes uniquely to the understanding of the nature of relevance and the process of relevance judgments for legal e-discovery.

STUDY DESIGN

This section describes the study design, including the participants, their tasks, the system they used to complete the tasks, and the guidelines and instructions facilitating the completion of the tasks.

Participants

Two groups of participants were recruited for the study, four law students (LAW[1-4]) and four LIS students (LIS[1-4]), representing assessors with and without legal domain expertise. The number of participants was decided mainly based on the number of topics we wanted to use in the relevance judgment tasks (described below in the Subsection „Tasks'). Although LIS students may possess better search skills than students from most other disciplines, we did not think this actually mattered much in this study, as it focuses on the relevance judgment phase in a typical IR task. For future study investigating the whole process of e-discovery, however, LIS participants form an interesting group because interaction between (law participants') subject knowledge and (LIS participants') search skills can be conveniently studied.

For comparison purposes, we limited LIS participants to those without a law background (there is a law librarianship specialization in the LIS program that has students who already have a degree in law.). In addition we required participants to be native English speakers to eliminate language skills as an independent variable.[1]

Participant recruitment started in mid-April, 2010. Electronic copies of the recruitment notice were sent to the student listservs of the law school and the LIS department. Printed copies of the notice were posted inside the buildings of the two academic units. In three days, about 20 law students and 10 LIS students responded, from which four law students and four LIS students were selected. In selecting law participants, we gave priority to those in their later years in the law program who have richer legal service experience.

Tasks

The task was to judge the relevance (responsiveness) of business documents for intended search topics (requests for production of responsive documents). The documents and topics were selected from the two test collections used in the TREC Legal track. We wanted to study relevance judgments of two collections of documents rather than one since we want to see how the situation differs across collections. Also, within a collection, we wanted to include more than one topic so that within-collection comparison is possible. In addition, we planned to use one topic from each collection for training and practicing purposes.

The test collection of MSA tobacco documents used in the 2008 TREC Interactive track has only one hypothetical complaint, which contains three requests for production, identified by topic IDs T102, T103, and T104, respectively. T102 and T103 were used as two of the four formal study topics while T104 was used as a training topic. The 2009 TREC Enron Email collection contains seven search topics for a hypothetical complaint. We decided to use Topic T202 and T203 for our formal study task and Topic T204 for training purposes.[2]

In deciding topics to be used in the formal study and for training and practicing, we took into consideration the

[1] Participants' gender was not considered and not recorded as part of the consent in the study. Each participant is referred to with the male pronouns throughout the paper.

[2] By the time we started the study, the final official relevance judgments for the remaining four Enron Email topics had not been completed.

LAW1/LIS1	LAW2/LIS2	LAW3/LIS3	LAW4/LIS4
T102	T103	T202	T203
T103	T102	T203	T202
T202	T203	T102	T103
T203	T202	T103	T102

Table 1. Design of Task Sequences.

difficulty of each topic. Specifically, we wanted to use the more difficult topics for the formal study. Among the three tobacco collection topics, T104 seems relatively easy, hence it used as the training topic. Among the three Enron email topics, T202 requires more subject knowledge and our TREC experience showed T203 is quite difficult, so these two are used as the formal study topic and the other as the training topic. Table 1 shows the design of the topic sequence for each participant, with a consideration that every topic will get a chance to be judged first, second, third and fourth in each participant group.

The next decision is the number of documents to be judged for each topic. Several factors had to be considered, including the availability of participants for the study, the expected time span, budget, and the minimum number of documents that would make comparison and analysis reasonably meaningful. On average, it took one hour for the TREC Legal voluntary assessors to judge about 20 documents[3], which is quite similar to the speed we estimated through a pilot study run with two graduate teaching assistants. Therefore, we decided to use 100 documents for each topic, expecting it would take approximately 20 hours on average for a participant to complete his relevance judgment tasks in the study.

In order to make our study results directly comparable with the TREC official relevance judgment results, we limited the documents to be used in the study to only those that had already been officially judged by the TREC Legal track. From these we selected 50 relevant documents and 50 nonrelevant documents for each topic. Furthermore, for each of these two categories, we wanted to select 25 documents whose initial judgments were overturned and 25 documents whose initial judgments were either not appealed or not overturned. It turned out that T102 has only five documents with overturned relevance judgments (three relevant and two nonrelevant). Also, we did not include any overturned documents for T202 due to a last-minute switch from T204 to T202[4]. In these two cases, we simply drew the remaining documents from the un-appealed pool. To eliminate any bias due to document order in the official relevance judgment result lists (sorted in order of document IDs), we selected documents at roughly equal intervals in the lists.

For the two Enron email topics, about 40 of the 200 selected documents are attachments. We limited the formats of selected attachments to MS Word, Excel, or PowerPoint so that we did not need to deal with outdated file formats such as WP5. Attachments were judged independently of parent emails, same as how relevance judgments were made in the TREC Legal track.

A binary relevance judgment scale was used in the study to simplify the relevance judgment decision. In addition to judging the relevance of each document, participants were asked to describe briefly the rationale and rate the difficulty level of each relevance decision. Also, the total amount of time spent on judging each document was automatically recorded by the system (explained below).

Relevance Judgment System

We developed a password-protected Web-based system implemented in PHP. It consists of a login page, a topic selection page, a task status page, and a relevance judgment page. Following the login page, the topic selection page shows the IDs of the two practice topics and the four formal task topics. Clicking on a topic ID will bring up the task status page for that topic. On the top of that page is the topic description; in the middle and lower section is the list of the IDs of documents that have not been clicked, documents that have been clicked but not judged yet (to be judged later), documents that have been judged as relevant, and documents that have been judged as nonrelevant. This page is also where a participant will return whenever he submits a relevance judgment.

Upon clicking on a document ID in one of the four lists described above, the participant will see the relevance judgment page. Again, the topic description is displayed on the top of that page, followed by the selected document's ID and a hyperlink, which will open the document to be read in a separate browse window when clicked. Next is a list of three radio buttons for selecting a relevance judgment decision: Relevant, Nonrelevant, and To Be Judged Later (which is used for temporarily holding a document that he wishes to judge later). Following the relevance selection buttons is a textbox for entering the relevance judgment rationale. The last item on the page is a collection of three radio buttons of perceived difficulty level of a relevance judgment decision: Difficult, Average, and Easy.

Links of tobacco documents on the relevance judgment page point directly to document pages in the Legacy Tobacco Document Library (LTDL) hosted by the University of California at San Francisco.[5] Clicking a

[3] This is learned through personal communication with Bruce Hedin, the TREC Legal Interactive task coordinator.

[4] The data of overturned judgments had to be requested from the TREC Legal coordinators.

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

document link directly opens the document on the LTDL web site. Documents of Enron emails are hosted on our own server. Email text is directly opened in the web browser when clicked, whereas email attachments are opened using the appropriate MS Office tool.

Instructions and Protocols

Participants in the study were provided with the two legal complaints and the general guidelines and topic-specific guidelines used by the TREC Legal track assessors. The general guidelines explain in general terms the relevance judgment task and the meaning of relevance in e-discovery; the topic-specific guidelines define the scope of responsive documents for each topic and tips that helped the TREC Legal track participating teams to clarify the topic. The topic-specific guidelines were an aggregation of communications (mostly in the form of questions and answers) between the topic authority and each participating team. Some of these topic-specific guidelines contain specific terms; the appearance of any one of these terms would make a document likely relevant. In addition, a step-by-step tutorial document was created to help participants get familiar with the task and the system.

Each participant was instructed specifically to complete the relevance judgment task as if he were contracted by a company to review documents for litigation purposes.

An entry questionnaire was used to solicit information of each participant's law background, legal service experience, and knowledge of e-discovery. In addition, a post-task exit interview was conducted with each participant to further collect information regarding his/her experience of working on the relevance judgment task. All interviews were recorded using a digital recorder.

Procedure

In each participant's first session, the researcher started with a brief explanation of the purpose of the study, the task, and the types of data and information being collected. The participant then signed an informed consent form. Next, the researcher walked briefly through the general guidelines, the two complaints, the topic-specific guidelines for the two practice topics, and the step-by-step system tutorials. The participant was then ushered to an office, where he would read carefully the guidelines and instructions to gain a good understanding of the study and the tasks. After that, the participant began to practice the relevance judgment task with the two training topics. When the participant felt ready for the task, the topic-specific guidelines document of his first topic were handed to her. In the following sessions, when a participant had judged all documents for one topic, the guidelines document for a new topic was handed to him.

It was estimated each participant would need six to ten three-hour long sessions to complete the task. Accordingly, we scheduled all tasks to be done in about 2.5 weeks. Each participant signed in and signed out for each session, so that the amount of time spent on the study can be kept for calculating compensation ($10 per hour). The amount of time spent on judging each document, however, was recorded automatically by the system.

	LAW	**LIS**
Current degree program?	All in the 2nd or 3rd year of their J.D. program	3 in the 1st or 2nd year of their MLS program; 1 graduated
E-discovery knowledge?	Average to Above average	None to Quite limited
E-discovery experience?	Average	None
Other legal service experience?	Multiple jobs or internships in law firms or attorney's offices	None

Table 2. Summary of Entry Questionnaire Data.

Each participant was assigned an individual office so that participants did not interfere or communicate with each other. They were instructed not to communicate with each other about the task on any other occasions. Also, they were required not to logon into the system outside of the scheduled time slots. While with the web-accessible system participants could virtually work on the task in other places, we felt it was important to require them to review documents in the same environmental setting.

DATA COLLECTION AND RESULTS

Table 2 summarizes data from the entry questionnaire regarding participants' current degree programs, knowledge and experience of e-discovery, and other law-related job/internship experience. As can be seen, the two groups were quite different in terms of their legal knowledge and experience; these are quite contrastive groups as planned.

Relevance Judgment Accuracy

One of the most interesting things we wanted to investigate is how accurately our participants judged the relevance of documents, as compared to the TREC official relevance judgments. Relevance judgment accuracy is measured in this study using *recall* and *discrimination*. Recall is the proportion of the TREC-relevant documents that were judged by a participant as relevant; discrimination is the proportion of the TREC-nonrelevant documents that were judged by a participant as nonrelevant. Discrimination was chosen over precision as it *directly* measures participants' relevance judgment accuracy for nonrelevant documents.

Table 3 shows the values of recall and discrimination achieved by each participant for each topic. A bolded cell in that table highlights the highest recall or discrimination achieved for the corresponding topic (same below). Overall, the average score of discrimination is higher than that of

	T102	T103	T202	T203
LAW1	0.5	0.72	**0.88**	0.72
LAW2	0.64	0.76	0.82	0.36
LAW3	0.44	0.6	0.7	0.26
LAW4	**0.76**	0.64	0.86	**0.8**
LIS1	0.6	0.84	0.76	0.52
LIS2	0.66	**0.86**	0.78	0.56
LIS3	0.3	0.42	0.7	0.42
LIS4	0.66	**0.86**	0.84	0.56

Table 3a. Relevance Judgment Accuracy: Recall.

	T102	T103	T202	T203
LAW1	0.94	0.84	0.88	0.86
LAW2	0.9	0.76	**0.98**	0.9
LAW3	**1**	**0.88**	**0.98**	**1**
LAW4	0.7	0.78	0.86	0.84
LIS1	0.88	0.64	0.94	0.9
LIS2	0.92	0.64	0.72	0.8
LIS3	**1**	**0.88**	**0.98**	0.84
LIS4	0.9	0.56	0.78	0.7

Table 3b. Relevance Judgment Accuracy: Discrimination.

recall, showing participants were more accurate in judging nonrelevant documents than judging relevant documents..

Comparison by Groups

We further computed the average scores of recall and the average scores of discrimination based on participant groups. There is no difference between the two groups in terms of the average recall – both achieved 0.65. The average discrimination of the LAW group (0.88) is slightly higher than that of the LIS group (0.81).

Comparison by Topics

Table 4 compares the average recall and the average discrimination by topics. There are noticeable variations among topics for both recall and discrimination. Taking into consideration both measures, we find on average participants judged documents of T202 most accurately and T203 least accurately. It does not seem that the document collection type was a factor influencing participants' relevance judgments. Rather, the subject matter of individual topics caused some major differences.

The fact that participants did relatively well on topic T202 surprised us because we thought that topic, which seeks documents about "…transactions that the Company characterized as compliant with FAS 140 (or its predecessor FAS 125)," would call for the most subject knowledge among the four topics. Participants later said in the exit interview that, while our speculation was true, the topic-specific guidelines for T202 were more helpful than those for other topics because the guidelines provide a list of specific terms of Special Purpose Entity (SPE) transactions. Once any of the terms appears in a document, that document could be judged as relevant right away.

	Recall	Discrimination
T102	0.57	0.91
T103	0.71	0.75
T202	0.79	0.89
T203	0.53	0.86

Table 4. Average Recall and Discrimination by Topics.

Comparison between Overturned and Nonoverturned or Nonappealed Judgments

Table 5 shows the comparison of recall and discrimination between documents whose initial relevance judgments were overturned and documents whose initial relevance judgments were either not appealed (nonappealed) or not overturned (nonoverturned). T102 and T202 were not included in the figure due to either having very few overturned relevance judgments (T102) or having no overturned judgments (T202).

For both topics, all eight participants consistently judged more accurately the documents whose initial relevance judgments were either not appealed or not overturned than those whose initial relevance judgments were overturned. This indicates that documents with overturned relevance judgments are indeed more difficult to judge. In addition, participants were also more accurate in judging nonrelevant documents than relevant ones, regardless whether the initial official relevance judgments were overturned or not.

	T103		T203	
Overturned?	**Yes**	**No**	**Yes**	**No**
LAW1	0.68	0.76	**0.72**	0.72
LAW2	0.68	0.84	0.28	0.44
LAW3	0.52	0.68	0.12	0.4
LAW4	0.48	0.8	0.68	**0.92**
LIS1	0.76	**0.92**	0.44	0.6
LIS2	**0.84**	0.88	0.4	0.72
LIS3	0.36	0.48	0.28	0.56
LIS4	0.8	**0.92**	0.52	0.6

Table 5a. Relevance Judgment Recall: Overturned vs. Nonappealed or Nonoverturned.

	T103		T203	
Overturned?	**Yes**	**No**	**Yes**	**No**
LAW1	0.76	**0.92**	0.76	0.96
LAW2	0.68	0.84	0.84	0.96
LAW3	**0.88**	0.88	**1**	**1**
LAW4	0.72	0.84	0.68	**1**
LIS1	0.44	0.84	0.84	0.96
LIS2	0.52	0.76	0.64	0.96
LIS3	0.84	**0.92**	0.76	0.92
LIS4	0.44	0.68	0.6	0.8

Table 5b. Relevance Judgment Discrimination: Overturned vs. Nonappealed or Nonoverturned.

Inter-rater Agreement

The inter-rate agreement measures the degree of agreement or disagreement between two assessors. It tells how reliable the assessment results are (although it may not say much about the validity). The most commonly used measure for inter-rater agreement is Cohen's Kappa (Cohen, 1960). It is regarded as a better measure than simple percentage of agreement counts because it also takes into consideration how much the agreement is due to chance.

Cohen's Kappa

Values of Kappa can range between 1 and any negative numbers, although only values between 0 and 1 make sense. A kappa of 1 means perfect agreement between two assessors, whereas a Kappa of 0 means no agreement or any observed agreement is due to chance. Landis and Koch (1977) proposed the following way of interpreting Kappa statistics:

- 0.01–0.20: Slight agreement;
- 0.21– 0.40: Fair agreement;
- 0.41–0.60: Moderate agreement;
- 0.61–0.80: Substantial agreement;
- 0.81–0.99: Almost perfect agreement.

We computed pair-wise Kappa between all eight participants. Table 6 summarizes the value range and the mean Kappa computed for each pair of participants within each group and across the two groups. Generally speaking, participants within each group agree with each other from fairly to substantially, with the majority agreeing moderately. That is also true for the agreement between all pairs of participants across the two groups on most topics. Kappa values for T203 are noticeably smaller than those for the other three topics. This is not surprising, however, because participants felt it was the most difficult topic (as learned from the exit interviews). Naturally, when a decision is difficult to make, people tend to disagree with each other more than when a decision is easy to make.

	Within LAW		Within LIS		LAW vs. LIS	
	Range	**Mean**	**Range**	**Mean**	**Range**	**Mean**
T102	0.28 – 0.68	0.44	0.44 – 0.68	0.48	0.16 – 0.71	0.52
T103	0.34 – 0.57	0.42	0.21 – 0.69	0.46	0.28 – 0.56	0.47
T202	0.56 – 0.76	0.69	0.39 – 0.73	0.54	0.35 – 0.79	0.61
T203	0.24 – 0.60	0.38	0.24 – 0.38	0.30	0.28 – 0.56	0.45

Table 6. Value Range and Mean of Cohen's Kappa between Participants within the Same Group or across the Two Groups.

Consensus of Agreement (with TREC)

We also examined for each relevant or nonrelevant document, how many assessors agreed with its TREC-relevance. For lack of a more appropriate term for this measure, we call it "Consensus of Agreement" (COA). Figure 1 shows the COA statistics for T203. The figure should be read in this way. Among the 50 relevant documents of this topic, there are only two documents that all eight participants judged as relevant; there are 11 documents that at least seven participants judged as relevant (including the two judged relevant by all 8); there are 47 documents that at least one participant judged as relevant; there are three documents that none of the participants judged as relevant. The agreement data for the 50 nonrelevant documents are shown in the darker bars.

Figure 1 also confirms participants tended to agree more often on documents being nonrelevant than being relevant. For example, for 80% of the nonrelevant documents, there were at least six participants agreeing to their officially judged relevance, whereas for 80% of the relevant documents, that number of participants decreased to three.

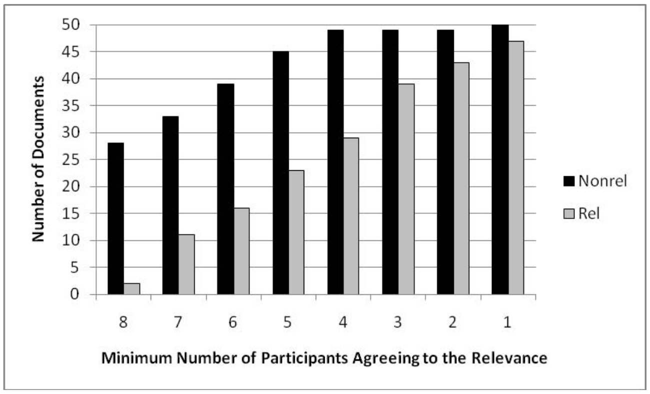

Figure 1. Consensus of Agreement for T203.

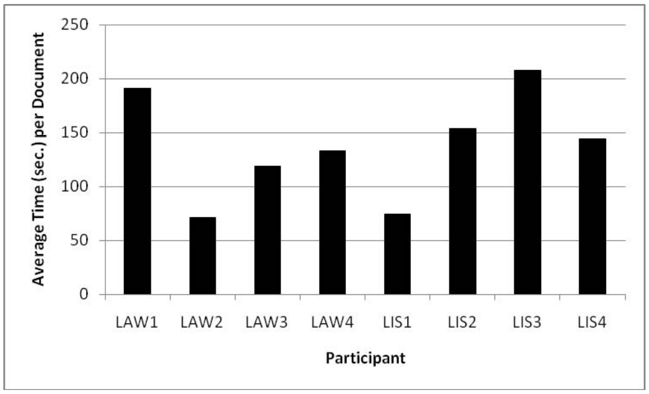

Figure 2. Average Relevance Judgment Speed

Also, very rarely did all participants agree to a relevance judgment or all disagree to a relevance judgment. These same patterns are observed for all the other three topics, although the actual numbers of documents in each category may vary.

Relevance Judgment Speed

We also computed the maximum, minimum, and average amount of time for each participant to judge relevance of documents for each topic. Figure 2 shows the comparison of relevance judgment speed by participants based on the average number of seconds per judgment. Three of the four law participants were among the fastest assessors, but the other participant (LAW1) in the group was one of the two slowest. We also noticed relevance judgment speed varied significantly among participants on the same topics and among different topics for the same participant. Overall, we do not find noticeable correlation between the relevance judgment accuracy and the relevance judgment speed.

Analysis of the Exit Interview Data

The qualitative data collected through the exit interview can help to confirm, clarify, complement, and extend the quantitative data collected through the relevance judgment system. While in-depth analysis of interview data and relevance judgment rationale data is still ongoing, we present some preliminary findings.

Topic difficulty

Three of the four law students thought the Enron Email topics were more difficult than the tobacco documents and two of these three participants specifically said T203 was the most difficult topic, mainly because these topics required more subject knowledge and/or emails were not so easy to read. The other law participant (LAW4) felt the tobacco topics were more difficult because generally the documents are longer. Interestingly, later in the interview we found LAW4 is also the only one who had studied the Enron case extensively in his law school courses and had even read some of the Enron emails used in this study. The participant acknowledged that experience helped his judgments of Enron emails significantly.

LIS1 and LIS3 also thought the two Enron Email topics were more difficult. LIS2 felt all topics had about the same, not too difficult and not too easy. LIS4 thought, prior to the formal task, Enron Email topics would be more difficult. After completing all the tasks, however, he felt that the tobacco topics were also challenging.

Overall, participants' perception of the topic difficulty, especially for T203, is consistent with their relevance judgment accuracy. That is, they tended to judge documents of more difficult topics less accurately.

Usefulness of Subject Knowledge

All four LAW participants acknowledged the usefulness of their law background and legal service experience for the relevance judgment task in the study, although LAW2 felt his reading skills may have helped more. LIS participants, on the other hand, did not feel as strongly that such subject expertise would have helped much. Indeed LIS1 did not think it would help at all with the tobacco topics while LIS3 felt it would only speed up the process occasionally but not necessarily improve the accuracy. Most of them felt the guidelines were sufficient to compensate for their lack of legal expertise.

We looked at some of the documents which most law participants judged correctly while most LIS participants judged incorrectly. This is something we will study carefully in our qualitative data analysis.

Document-, Topic-, and Collection-Level Learning Effect

All eight participants thought reading more and more documents for a topic made it easier and easier to make relevance judgments. Most participants also agreed judging documents of the first topic helped judging the second topic of the same complaint, although the help was largely for making them aware in advance what kind of topic and documents they would see next. No participant thought judging documents in one collection first benefited judging documents in the second collection as documents in the two collections and topics are quite different.

Usefulness of Additional Materials/Aides

All eight participants thought the guidelines, particularly the topic-specific guidelines, were quite useful and sufficient for them to make relevance judgments. Several mentioned more examples of relevant and nonrelevant documents or more question-answer pairs would certainly be helpful. LIS3 felt having a teammate or a group of teammates would make the task easier, especially with documents whose relevance can go either way. LAW3 wished they would be allowed to ask someone like a topic authority questions about specific documents. LIS4 said the complaints were not as helpful as he thought even if he spent quite some time reading the complaints.

Factors Deciding Relevance Judgment Difficulty

All participants felt a judgment was "easy" if it could be made either right after finishing reading the document or even before finishing reading it when some keywords were easily spotted, as in the case of T202. A judgment was "difficult" if the document had to be read a few times or had to be held for later judgment. Although generally long documents tended to be more difficult to judge, several participants said it was not always the case. Also, all participants felt some short documents actually were quite difficult to judge because they were not clear about the subject matter of these documents.

Skills of Judging Long Documents

Some documents in the two collections are very long, especially those in the tobacco collection. This is also true for some of the documents included in our study, e.g., at least two documents for T102 have each more than 300 pages. During the interview, participants were asked specifically how they review such documents. Skimming seemed to be the most commonly used approach. In addition, several participants mentioned they first read the abstract and conclusion section. One participant mentioned the use of an index term list at the end of a lengthy document. Two participants said they also considered the type of these long documents. For example, if a long document was comprised of summaries of scientific journal articles of lung cancer research, they felt it less likely relevant to the topic that sought documents about restrictions on advertising of tobacco products.

Causes of Incorrect Relevance Judgments

Participants were also asked to look at the relevance judgments, rationale, and perceived difficulty for five selected documents for each topic and explain them. Those documents were selected because either most participants rated them difficult to judge, the two groups judged them differently, most participants judged them incorrectly, or they are extremely long. While analysis of this part of the data is still ongoing, we find at least three contributing factors of relevance judgment mistakes:

- Not reading the whole document, in particular very long ones. Mistakes made due to this factor were mostly relevant documents being judged as nonrelevant.
- Over-relying on document types. For example, one law assessor and one LIS assessor thought scientific articles would not mention MSA restrictions on advertising. This of course is a risky assumption.
- Misunderstanding or lack of sufficient or accurate understanding of the concept of relevance for some particular topics. For example, there is a document for T102 concerning the Virginia public's opinions on federal restrictions on general tobacco advertisement, including advertising targeting teenagers. All four law assessors correctly judged the document as relevant, while only two LIS assessors thought it relevant. The other two LIS assessors thought relevant documents for the topic should talk about restrictions on *specific* advertising activities although that was clearly not required, hence an incorrect understanding of the topic.
- Lack of subject knowledge. None of the eight participants judged correctly a relevant document about a financial swap between Enron and Blackbird for T202. It should be noted, however, that the subject knowledge here is of finance, not law.

We believe we will gain more insights into the factors influencing participants' relevance judgment accuracy through detailed analysis of the rationale they provided.

CONCLUSION AND FUTURE WORK

This paper presented mainly our quantitative analysis of data collected from a study in which four law students and four LIS students judged the relevance of 100 business documents for each of four search topics. When provided with the same relevance guidelines, both groups judged 65% of the relevant documents correctly and more than 80% of the nonrelevant documents correctly, using the relevance judgments made by the assessors of the TREC Legal track as the ground truth. The relevance judgments made by the law participants on nonrelevant documents were just slightly more accurate than those made by the LIS group. The relevance judgment speed varied noticeable among the eight participants and we did not find significant differences between the two groups in that regard. This partly answered research questions 1 and 3. Based on these findings, we argue that people without a law background can review documents for e-discovery if given good guidelines on how to judge document relevance.

These findings are still preliminary and may be affected by several limitations of the study. Although we feel the number of documents used in the study was sufficient, the number of topics was quite limited, thus making it difficult to generalize these findings to other situations where quite different topics are used. Similarly, the number of participants was also limited. For quantitative analysis, four subjects per population are not enough, especially when significant variations of relevance accuracy and speed were observed in the study. Because of these limitations, we do not feel our second research question can be answered here.

Our preliminary analysis of the interview data did reveal several factors that could influence the accuracy of people's relevance judgments. All law participants felt their legal knowledge was useful for the task. Interestingly, LIS participants did not share that opinion, although all of them did acknowledge the usefulness of relevance guidelines, which actually provide lots of legal information. Lack of accurate understandings of topics seemed a major factor adversely affecting the relevance judgment accuracy, and part of that sometimes was due to participants' inaccurate assumptions.

Interview data also suggests some of the techniques used by participants for reviewing documents, namely, skimming abstracts and index term lists for keywords, considering document types, and cross-referencing documents and/or topics.

Future work will focus on the detailed analysis of the relevance judgment rationales that participants of the study provided with a view towards analyzing the different relevance relationships, how a document can support a legal argument in the case. We can then analyze how different participants used these relevance relationships and whether there is a difference between law students and LIS students. This should lead to insights into assessor techniques and processes of relevance judgments for e-discovery.

We also plan another study in which assessors will work in groups when reviewing documents. The study will tell us how complementary knowledge and skills can be used and whether that kind of practice should be preferred to document review done by individual assessors.

ACKNOWLEDGMENTS

The authors would like to thank the eight students for participating in the study, Bruce Hedin for sharing ideas and providing the TREC Legal track relevance judgment data, Doug Oard, Valerie Nesset, and Lorna Peterson for providing valuable comments on an earlier version of the paper, and Ying Sun for helping design the questionnaire and the interview protocol used in the study.

REFERENCES

Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A. P., Yilmaz, E. (2008). Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st Annual international ACM SIGIR Conference on Research and Development in information Retrieval,* 667-674.

Baron, J. R., Lewis, D. D., Oard, D. W. (2006). TREC-2006 legal track overview. In *Proceedings the Fifteenth Text REtrieval Conference (TREC 2006)*. Retrieved May 24, 2010 from http://trec.nist.gov.

Barry, C. L., Schamber, L. (1998). Users' criteria for relevance evaluation: a cross-situational comparison, *Information Processing and Management*, 34(2-3), 219-236.

Cleverdon, C. W. (1970). The effect of variations in relevance assessments in comparative experimental tests of index languages. *Technical Report ASLIB Part 2*, Cranfield Institute of Technology.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval*, **7**(1), 19-37.

Hedin, B., Tomlinson, S., Baron, J. B., Oard, D. W. (2009). Overview of the 2009 TREC legal track. In *Proceedings the Eighteenth Text REtrieval Conference (TREC 2009).* Retrieved May 24, 2009 from http://trec.nist.gov.

Huang, X., Soergel, D.. (2006). An evidence perspective on topical relevance types and its implications for exploratory and task-based retrieval. *Information Research*, 12(1). Retrieved May 26, 2010 from http://informationr.net/ir/12-1/paper281.html.

Huang, X. (2009) Topical Relevance, Rhetoric, and Argumentation: A Cross-Disciplinary Inquiry into Patterns of Thinking and Information Structuring PhD Dissertation, University of Maryland. Retrieved May 26, 2010 from http://terpconnect.umd.edu/~xiaoli/XH_DisserationAbstract.pdf

Landis, J. R., Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* 1977, 33, 159-74.

Oard, D. W., Hedin, B., Tomlinson S., Baron J. B. (2008). Overview of the TREC 2008 legal track. In *Proceedings of the Seventeenth Text REtrieval Conference*. Retrieved May 24, 2010 from: http://trec.nist.gov.

Voorhees, E. M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 315-323.

Wang, J., Sun, Y., Mukhtar, O, Srihari, R. (2008). TREC 2008 at the university at buffalo: legal and blog track. In *Proceedings the Seventeenth Text REtrieval Conference (TREC 2008)*. Retrieved May 24, 2009 from http://trec.nist.gov.

Wang, J., Sun, Y., Thompson, P. (2009). TREC 2009 at the university at buffalo: interactive legal e-discovery with Enron emails. In *Proceedings the Eighteenth Text REtrieval Conference (TREC 2009)* Retrieved May 24, 2009 from http://trec.nist.gov.

Wang, P., Soergel, D. (1998). A cognitive model of document use during a research project. Study I. Document selection. *Journal of the American Society for Information Science and Technology*, 49(2), 115-133

Wilson, P. (1973). Situational relevance. *Information Storage and Retrieval*, 9(8), 457-471.