Accuracy, Agreement, Speed, and Perceived Difficulty of Users' Relevance Judgments for E-Discovery

Jianqiang Wang
Department of Library and Information Studies
Graduate School of Education
University at Buffalo, the State University of New York
Buffalo, NY 14260
jw254@buffalo.edu

ABSTRACT

This paper presents a study in which four law students and four Library and Information Science (LIS) students judged independently the relevance of documents selected from the e-discovery test collections of the Text REtrieval Conference. The results were compared with the official relevance ground truth and among participants. Given the same task guidelines and minimal training, on average the law assessors achieved the same accuracy on judging relevant documents as the LIS assessors but slightly higher accuracy on judging nonrelevant ones than the latter. Assessors showed moderate to substantial agreements on their relevance judgments. Relevance judgment speed varied markedly among assessors, with a small number of documents costing much more time to review than others. While relevance judgment difficulty is large subjective, all assessors perceived significant distinction between 'difficult' judgments and 'average' or 'easy' ones. Strong correlations were observed between relevance judgment speed and perceived difficulty and between perceived difficulty and relevance judgment accuracy, but not between relevance judgment accuracy and speed. Document subjects, length, and legibility, assessor's subject knowledge and reading skills, relevance guidelines, and learning effects are identified as factors influencing relevance judgment quality. Implications for building test collections and practicing e-discovery are suggested.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human Information Processing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Human Factors, Measurement

Keywords

Relevance, E-discovery

1. 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. Searching electronic business records for the purpose of litigation or government investigation, however, is a challenging task. On the one hand, lawyers do not want to miss any responsive documents that could affect the outcome of their clients' lawsuits; this means 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 precision is also economically important. In order to build search systems that can achieve both high recall and high precision for e-discovery, we need to first understand what kind of knowledge lawyers use and how they use it to assess the *relevance*, or *responsiveness*, of documents to requests for production. Once we gain a good understanding of the subject expertise, search expertise, and other factors that influence relevance judgments for e-discovery, we can then model it with more effective and efficient information retrieval (IR) systems and technology.

The concept of relevance in e-discovery does not seem to have been well studied and understood in the general IR community. This is partly evidenced by the results and findings produced through the Legal E-Discovery track of the Text REtrieval Conference (TREC), an annual IR research and evaluation workshop organized by the U.S. National Institute of Standards and Technology (NIST). As a participating team of that task, we reviewed many documents and found in some cases the relevance of a document was indeed quite difficult to decide. The Interactive TREC Legal E-Discovery task uses a unique mechanism of "appeal and adjudication," through which teams can appeal the first-pass relevance judgments made officially by professional review firms, voluntary law school students, and/or practicing lawyers. A topic authority (usually a senior lawyer) will then review the appealed judgments and provide the final relevance judgments. It is worthy noting that about 79% and 91% of the appealed relevance judgments were reversed after adjudication for the Interactive TREC Legal E-Discovery

task in 2008 and 2009, respectively [13, 10]. While readers should be careful in interpreting these numbers (e.g., there are still many more judgments that participating teams and the voluntary reviewers agreed on), both cases indicate the difficulty of accurately defining the concept of document relevance in e-discovery seems to be a pervasive problem.

Partly due to this observed difficulty of making relevance judgments for e-discovery and our experience of participating in the Interactive TREC Legal E-Discovery task [17, 18], we conducted a comparative user study on relevance judgments made independently by four law school students and four Library and Information Science (LIS) students. Preliminary analysis of the results was reported in the 73rd American Society of Information Science and Technology (ASIS&T) Annual Conference Proceedings [16]. In that paper, we reported and compared the relevance judgment accuracy and agreement of these two groups of document reviewers. We also briefly reported the average relevance judgment speed of each assessor (over 400 documents of four topics) and identified several factors influencing participants' relevance judgments based on our analysis of data collected through an exit interview with each participant.

This paper, while reporting the same study as our ASIS&T paper, presents several additional contributions. Specifically, we look at participants' relevance judgment speed in much more detail through examining how much time was spent on each relevance judgment, identifying which documents cost more time to review than others, and trying to explain why; we examine the difficulty level of each relevance judgment as perceived by the participant; we then look into if there is any correlation between users' relevance judgment accuracy, speed, and perceived difficulty. Of course, to make our presentation more coherent we highlight in several places some of the results and findings (noticeably the relevance judgment accuracy, relevance judgment agreement, and influencing factors) reported in our previous paper.

The rest of the paper is organized as follows. Section 2 reviews related work on defining relevance and comparing relevance judgments for e-discovery. Section 3 describes our study design. Section 4 presents the results and our analysis. Section 5 concludes the paper by highlighting the findings and implications of the study and outlining future work.

2. RELATED WORK

Much research has been conducted, in LIS and other fields, to define the concept of relevance and to study the criteria and techniques users rely on in seeking information to satisfy their information needs. For example, Cooper defined logical relevance based on a strict logical deduction relationship between a statement and a sought-after answer [7]; Wilson defined evidential relevance and situational relevance [20]; Barry and Schamber studied the relevance criteria used by actual information users [4]; Wang and Soergel constructed a cognitive model of decision-making in information users' selection of documents [19]; Huang and Soergel focused on the evidentiary connection between a piece of information and a user's question, topic, or task [11].

Another related body of research has been conducted on the development of test collections for IR evaluation, with the ground truth relevance judgments made (usually) by subject experts as a key component. Cleverdon is perhaps the first who discussed the value of IR test collections [6]. Voorhees studied the variations of relevance judgments made by different types of assessors for the TREC test collections [15]. Despite the marked variations in the relevance judgments, she concluded the relative effectiveness of different retrieval strategies is stable. Bailey et al compared the relevance judgments made by topic experts, task experts, and people without either types of expertise [1]. 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 judges when these judges vary widely in task and topic expertise." While providing valuable insights into the concept of relevance, the criteria specifically used by actual information users, and the validity of expert-created relevance judgments for IR evaluation, none of these studies addressed directly the issues of e-discovery.

The TREC Legal E-Discovery track has produced several interesting studies on relevance judgment accuracy for e-discovery and agreement rates between assessors. Oard et al reported a pilot study that investigated the agreement between different assessors on judging the relevance of TREC Legal E-Discovery documents [13]. In their study, (up to) 10 documents judged relevant and 10 documents judged nonrelevant for each topic in 2006 or 2007 were assigned to the 2008 assessors for reassessment. Among the total 116 previously judged relevant documents for 12 topics, only 66 (or 58%) were judged as highly relevant or relevant by the 2008 assessors; among the 120 previously judged nonrelevant documents, 98 (or 82%) were rejudged as nonrelevant. Overall, the 2008 assessors agreed with the 2006/2007 assessors about 69% of the time. A similar study was also conducted for the 2006 TREC Legal E-Discovery track, in which 25 relevant documents and 25 nonrelevant documents judged for each topic by some assessors were assigned to some other assessors for rejudgment. The researchers found on average an agreement rate of 63% on relevant judgments, 81% on nonrelevant judgments, and 77% when both types of relevance were considered [3].

Several other studies of the TREC Legal E-Discovery track focused qualitatively on relevance factors or characteristics of the relevance judgment process. Efthimiadis and Hotchkiss studied how LIS graduate students and law students formulate queries, perform search, and judge documents for the Interactive task of the TREC 2007 Legal E-Discovery track [8]. They concluded that "legal training may not be all that critical to developing effective search queries" while "legal training is critical however to assessing relevance for use in a legal proceeding." For the same TREC E-Discovery task, Chu recruited three groups of LIS Ph.D. students to judge the relevance of documents [5]. 11 relevance factors were identified by at least five of the nine participants in her study. The researcher found that specificity of the search requests, ease of use of the choice scale (i.e., regarding the relevance of a document), and topicality were among the most important relevance factors. Yue et al investigated what they called "collaborative information behaviors" (CIB) through an experiment in which two LIS students were asked to judge collaboratively the relevance of documents for the TREC 2008 Legal track's interactive task

with another student (with e-discovery experience) available as a topic authority [21]. Frequent communications, division of labor, and teammate awareness were identified by the researchers as the three major characteristics of CIB in e-discovery. However, the researchers could not claim whether assessors could achieve higher relevance judgment accuracy collaboratively than individually.

Recently using the TREC 2009 Legal E-discovery track's experiment data, Grossman and Cormack studied the effectiveness and the efficiency of technology-assisted review of documents for e-discovery as compared to exhaustive human review [9]. The basic idea is humans review a small fraction of documents and along the way the computer retrieve potentially relevant documents using some machine learning techniques based on the manual review results. Such technology-assisted relevance judgment results were compared to the TREC official relevance judgments in terms of recall, precision, and a balanced f measure. The researchers concluded that "technology-assisted review can (and does) yield more accurate results than exhaustive manual review, with much lower effort." Interestingly, in their study document review was completed at a rate of three seconds per document (with the assistance of a computer system).

At least two other studies compared relevance judgments of documents for actual litigations or government investigations. Roitblat et al reported a study in which two teams of professional legal reviewers were asked to rejudge 5,000 (sampled) documents that were originally reviewed for the U.S. Department of Justice's investigation of the acquisition of MCI by Verizon [14]. The researchers found that the two teams agreed with each other on the relevance of 70% documents, while they agreed with the original reviewers on the relevance of 76% and 72% documents, respectively. Barnett et al found the responsiveness rate varied between 49% – 58% among five groups of reviewers who independently judged the relevance of 10,000 documents dealing with service industry subject matter [2].

Our study reported in this paper, while resembling some of the above (especially [3, 13, 2, 14]) in comparing users' relevance judgment accuracy and agreement, presents several unique contributions. The most obvious one is perhaps our comparison of relevance judgments made by people with a law background or legal service experience with those without that background or experience. In addition, we also investigated other factors such as relevance judgment speed and user perceived relevance judgment difficulty and more importantly whether they are correlated to each other as well as with relevance judgment accuracy.

3. STUDY DESIGN

In this section we briefly describe the participants, the topics and documents, the relevance judgment guidelines and instructions, and the relevance judgment system used in our study. A detailed description of the study design can be found in our ASIS&T paper [16].

3.1 Participants

We recruited four law students (designated as LAW[1-4]) and four LIS students (designated as LIS[1-4]) from a pool of about 30 responders. Table 1 summarizes the demographic data collected through an entry questionnaire regarding these participants' 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.

	Law participants	**LIS participants**
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 1: Summary of demographic data collected through the entry questionnaire.

LAW group
LAW1: T102 → T103 → T202 → T203
LAW2: T103 → T102 → T203 → T202
LAW3: T202 → T203 → T102 → T103
LAW4: T203 → T202 → T103 → T102
LIS group
LIS1: T102 → T103 → T202 → T103
LIS2: T103 → T102 → T203 → T202
LIS3: T202 → T203 → T102 → T103
LIS4: T203 → T202 → T103 → T102

Table 2: Relevance judgment task sequences. Each participant completed his tasks independently of other participants.

3.2 Dataset and User Tasks

The topics and documents used in our study were selected from the two test collections developed for the TREC Legal E-Discovery track. Here are the four formal study topics:

- *T102: Find 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.*
- *T103: Find 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.*
- *T202: Find all documents or communications that describe, discuss, refer to, report on, or relate to the Companys engagement in transactions that the Company characterized as compliant with FAS 140 (or its predecessor FAS 125).*

Test collection	Tobacco Docs		Enron Emails	
Search topic	T102	T103	T202	T203
Rel docs (1st-pass TREC judgments not appealed/overturned)	47	25	50	25
Rel docs (1st-pass TREC judgments appealed and overturned)	3	25	0	25
Nonrel docs (1st-pass TREC judgments not appealed/overturned)	48	25	50	25
Nonrel docs (1st-pass TREC judgments appealed and overturned)	2	25	0	25

Table 3: Documents statistics.

- *Find all documents or communications that describe, discuss, refer to, report on, or relate to whether the Company had met, or could, would, or might meet its financial forecasts, models, projections, or plans at any time after January 1, 1999.*

The first two topics are from one hypothetical complaint created for the Master Settlement Agreement (MSA) tobacco document collection for the 2008 TREC Legal Interactive task; the other two topics are from another hypothetical complaint created for the 2009 TREC Enron Email collection. Therefore, each pair of topics are somehow related. In addition, we also used another one topic (T104 and T204, respectively) from each collection for training purpose. The topics were permuted within each group of participants so that each would receive an equal chance of being judged the first, second, third, and last (see Table 2).

We used 100 documents for each topic in the study – 50 relevant documents and 50 nonrelevant documents, which were drawn from the final official TREC relevance judgment pool for each topic. Furthermore, for each group of documents (relevant or nonrelevant), we intended to include 25 documents whose first-pass TREC relevance judgments were appealed and overturned and 25 documents whose first-pass TREC relevance judgments were either not appealed or appealed but not overturned. It turned out only five documents of T102 whose first-pass TREC relevance judgments were appealed. Also, a last-minute decision prevented us from including any appealed and adjudicated TREC relevance judgments for T202. Nonetheless, T103 and T203 did provide us the kind of data we wanted. Table 3 shows the number of documents in each category. Documents in each category were randomly drawn from the much larger TREC relevance judgment pool and then mixed in no particular order into a 100-document set for each topic.

A password-protected Web-based system was implemented for participants to complete the relevance judgment tasks. A binary relevance judgment scale was used to simplify the relevance judgment decision. In addition to judging the relevance of each document, each participant was asked to describe briefly the rationale of each judgment in a designated textbox on the relevance judgment interface and rate the difficulty level of each relevance decision. The total amount of time spent on judging each document was automatically recorded by the system. At any point of working on a topic, the participant could choose any of the 100 document to review, be it already judged or not. Information of when a document was reviewed and how many time it was reviewed were also automatically logged.

3.3 Guidelines, Instructions, and Procedure

Materials given to each participant in the study include: (1) *two hypothetical legal complaints*, from which the six topics used in the study were developed; (2) *general relevance guidelines*, which were the same as the ones used by TREC assessors; (3) *step-by-step instructions of using the relevance judgment interface*; and (4) *topic-specific guidelines*, which were the same as the ones given to TREC assessors.

Items 1–3 and the topic-specific guidelines for T104 and T204 (i.e., the two training topics) were given to each participant at the beginning of his first session. Participants were instructed to read these materials carefully and practice using the relevance judgment system as long as needed. They were also encouraged to ask the researcher any questions they had regarding the instructions and guidelines, their tasks, and the system. After a participant reported to the researcher that he was ready to start with the formal study, the topic-specific guidelines of the first topic were distributed to him. When a participant finished judging documents for a topic, the guidelines for the next topic were distributed. After a participant completed all four topics, an exit interview was conducted to further collect information regarding his experience of working on the relevance judgment task. All interviews were recorded using a digital audio recorder.

While working on the formal study topics, participants were instructed not to ask questions unless they were technical issues such as failing to logon into the system or not being able to open a document (interestingly no such problems were reported by any participant). They were also instructed not to communicate with each other about the task, and they were required not to logon into the system outside of the scheduled time slots.

4. RESULTS

In this section, we first briefly summarize our initial findings of participants' relevance judgment accuracy and agreement that were reported in [16]. We then move on to present our new analysis of participants' relevance judgment speed and relevance judgment difficulty perceived by participants as well as the correlation between relevance judgment accuracy, speed, and perceived difficulty.

4.1 Summary of Relevance Judgment Accuracy

Relevance judgment accuracy in this study is measured by recall, discrimination, and overall accuracy. *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; *overall accuracy* is the proportion of documents that were judged correctly (i.e., regardless

of their official relevance).

Overall, we found relevance judgment recall varied markedly between assessors on each topic, ranging between 0.3 (by LIS2 on T102) and 0.88 (by LAW1 on T202). However, there was no difference between the two groups in terms of the average recall – both groups achieved 0.65. In addition, all assessors achieved the highest or the second to the highest recall on T202 among four topics, indicating T202 is perhaps the easiest topic. On the other hand, both groups achieved the lowest average recall on T203, which is from the same collection as T202. Therefore, it does not seem the types of documents (tobacco documents vs. Enron emails) affected participants' relevance judgment accuracy in this study.

Similarly, relevance judgment discrimination varied noticeably among the eight assessors. Comparing between the two groups, we found that the law group judged nonrelevant documents slightly more accurately than the LIS group (i.e, 88% vs. 82%). A two-tailed Student t-test comparing 32 pairs of recall and discrimination values indicates assessors judged nonrelevant documents significantly more accurately than relevant ones (with $p = 0.0001$).

Looking at the overall accuracy, we found on average the LAW assessors judged about 77 documents (out of 100) for a topic correctly while the LIS group judged the relevance of about 73 documents (out of 100) for a topic correctly. In other words, our assessors agreed with TREC assessors more than 70% of the time. This finding is quite consistent with those obtained from previous studies that were surveyed in Section 2. Also, the finding indicates the law participants judged slightly more accurately than the LIS participants, although the difference does not seem as striking as one would speculate, and more importantly, the difference is largely due to that of judging nonrelevant documents.

Comparing the recall, discrimination, and overall accuracy of relevance judgments of documents whose first-pass official TREC relevance judgments were overturned and those not appealed or not overturned,[1] we found all eight participants consistently judged more accurately documents in the latter category than those in the former category. This indicates 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 for either overturned relevance or nonappealed/nonoverturned relevance.

In summary, the relevance judgment accuracy results obtained in our study are consistent with and within the range of those reported in [3, 13, 2, 14]. As the assessors in all these other studies have a law background and/or e-discovery experience while the LIS assessors in our study do not, our results seem to suggest LIS people (without legal experience) can fulfill the task of reviewing documents for e-discovery.

4.2 Summary of Annotation Agreement

We computed Cohen's Kappa between each pair of assessors in order to see the degree of agreement between them. We relied on the following way of interpreting Kappa statistics,

[1]T102 and T202 were not included in this comparison.

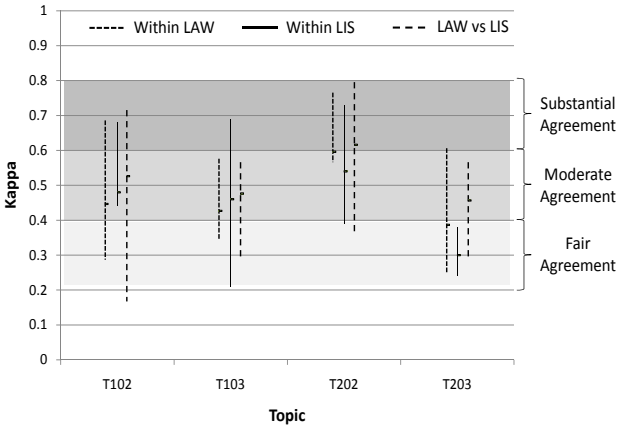

Figure 1: Value range and mean of Cohen's Kappa between participants within the same group or across the two groups.

as suggested by Landis and Koch [12]: (1) 0.01-0.20: Slight agreement, (2) 0.21- 0.40: Fair agreement, (3) 0.41-0.60: Moderate agreement, (4) 0.61-0.80: Substantial agreement, and (5) 0.81-0.99: Almost perfect agreement.

Figure 1 shows the value range and the mean Kappa computed for each pair of participants within each group and across the two groups when both relevant documents and nonrelevant documents are considered. 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, it is more likely for people to disagree on a difficult decision than on an easy one.

we also computed pair-wise Kappa for all eight assessors when only their relevance judgments of documents whose first-pass TREC relevance was appealed and overturned were considered. We focused on T103 and T203 as we included 50 such documents for each of these two topics in the study. Not surprisingly, Kappa values varied between 0.16 and 0.29 for T103 and 0.17 and 0.28 for T203 (i.e., with only fair agreement), showing assessors disagreed more on the relevance of such documents than on the relevance of those whose first-pass TREC relevance was not appealed or not overturned for adjudication. Therefore, it is safe to assume that the errors made by TREC assessors with judging the relevance of the appealed documents was at least partly due to the fact that those documents are more difficult to judge than others.

4.3 Relevance Judgment Speed

Investigation of relevance judgment speed by previous studies was largely limited to *average* values, e.g., the average number of documents reviewed per hour or the average amount of reviewing time per document. This is useful for

	T102	T103	T202	T203	Avg.
LAW1	**12.7**	20.4	29.1	18.9	18.6
LAW2	26.3	41.0	**83.1**	50.4	42.4
LAW3	39.0	51.6	48.4	30.2	40.5
LAW4	37.3	32.4	46.5	27.0	34.4
LAW Avg	23.6	32.3	45.2	28.0	**30.5**
LIS1	24.1	35.9	67.4	48.8	38.2
LIS2	18.7	16.4	29.0	23.4	20.9
LIS3	29.8	36.3	37.7	17.3	27.5
LIS4	27.3	16.8	49.3	25.1	25.6
LIS Avg	24.2	22.7	41.6	24.9	**26.8**
All Avg	24.0	26.7	43.3	26.3	28.5

Table 4: Average number of documents reviewed per hour.

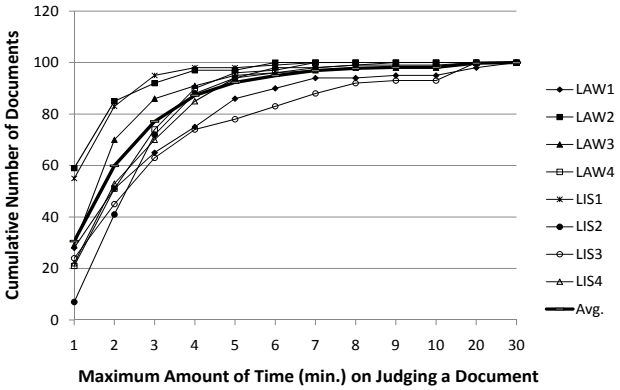

Figure 2: Document distribution based on relevance judgment speed for T203. Each data point (x, y) in the figure should be read like this: there are a total of y documents each of which was reviewed with no more than x minutes.

estimating the total amount of time and manpower needed for reviewing documents for law suits or for building test collections for e-discovery research. However, such studies revealed little about the amount of time that an assessor spent on judging *individual* documents and more importantly why an assessor would spend more time on some documents than on others and how relevance judgment speed and relevance judgment accuracy are related to each other. In this section, we attempt to look at these issues in some detail.

Table 4 shows the average number of documents that were reviewed per hour by each assessor for each topic, as well as the macro-average over assessors, groups, and topics. Several things stand out from that table. First, relevance judgment speed varied markedly among assessors, ranging from 13 documents per hour to 83 documents per hour with an average of 29 documents per hour. This macro-average speed is a bit higher than what's reported in [3, 13], but we suspect the difference is largely due to the fact that we only take into consideration the amount of time that our assessors spent on reading documents whereas the computation in these two previous studies had perhaps also counted the time assessors spent on consulting topic authorities. On the other hand, our assessors' relevance judgment speed is significantly lower than what's reported in [9]. Averaging over all topics, the LAW group judged about four more documents per hour than the LIS group, with three of the four law participants being the fastest among eight assessors, but the other participant (LAW1) being one of the slowest. Second, relevance judgment speed also varied noticeably across topics. Seven assessors judged documents of T202 faster than those of the other three topics. During the exit interview, all assessors responded that T202 was the easiest topic due to the inclusion of a list of key terms in the relevance judgment guidelines.

When we look at relevance judgment speed based on the type of relevance (relevant/nonrelevant) as given by assessors, however, the story seems different. We computed the average number of documents marked per hour as relevant and that of documents marked per hour as nonrelevant by each assessor for each topic, which gives us 32 pairs of speed values. A two-tailed Student t-test shows that the difference between the average speed values is statistically indistinguishable (with $p = 0.06$), thus indicating assessors reviewed nonrelevant documents about as fast as relevant documents. This finding is somehow against the traditional wisdom that an assessor can mark quickly a document as relevant as soon as he sees a relevant piece of text in the document (hence no need to read the rest of the document) whereas the assessor perhaps needs to read the full document before he can decide it is nonrelevant. That basically says, other factors being equal, on average it tends to take more time to make a nonrelevant judgment than a relevant one. At least for the topics used in our study, however, this argument is not true. In fact, on average (over all assessors and all topics) assessors judged four more nonrelevant documents per hour than relevant ones.

The average speed values reported so far do not tell the time allocation among individual documents. Therefore, we further looked at the amount of time on judging individual documents. Figure 2 shows the speed of each participant on judging the relevance of documents for T203 (which is regarded as the most difficult topic by most assessors). As we can see, the distribution of documents is highly skewed toward the high speed end (and the pattern also holds for other three topics). Specifically, all eight assessors reviewed more than 60 documents (out of 100) at a speed of no more than three minutes per document (or on average 45 documents per hour) and seven assessors finished judging 90 documents at a speed of no more than six minutes per documents (or on average 34 documents per hour). Clearly, only a small number of documents slowed down assessors.

We found there were 17, 10, 2, and 13 documents for T102, T103, T202, and T203 respectively, on each of which at least one assessor spent more than 10 minutes. Interestingly, there are only a few overlaps of these documents among assessors. Also, most of these documents were found in the relevance judgment data of LAW1, LIS3, and LIS4. We examined each of these documents and the assessor's relevance judgment notes. Most of these documents are much longer as compared to documents not included in the list, indicating document length is indeed a factor that influences assessors' speed. Assessors also marked most of them as

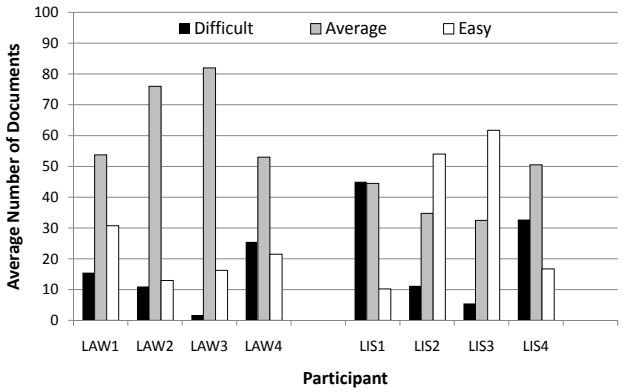

Figure 3: Document distribution based on the level of difficulty perceived by assessors.

'difficult' to judge. Two assessors spent much time on the first document their judged for two topics even if these documents are short. This seems to indicate the assessors were trying to get some basic understanding of the collection and the topic in the beginning of the task. There are another five instances in which the document is not long and the assessor did not feel it was difficult to judge but it still took more than 10 minutes to judge. We are not sure what exactly caused the assessors to slow down in these cases.

4.4 Relevance Judgment Difficulty

Assessors in our study were asked to mark whether it was 'difficult,' 'easy,' or somewhere in between ('average') to judge each document. We chose not to define what each of these three levels of difficulty means, leaving it to each assessor's own perception. Our goal is two-fold: to gain a better understanding of the difficulty that people encounter when reviewing documents for e-discovery and to see whether there is any correlation between assessors' perceived relevance judgment difficulty and their relevance judgment accuracy or speed.

Figure 3 shows the number of documents falling into each of the three categories of difficulty when averaged over four topics. Assessors showed great variations in their perception of relevance judgment difficulty. Interestingly, LAW assessors marked 26 more documents per topic as 'average' than LIS assessors, while the latter marked 10 more documents as 'difficult' and 15 more documents per topic 'easy' than the former. In other words, it seems the LAW assessors tended to have more *averaged* perception of relevance judgment difficulty whereas the LIS assessors tended to have more *polarized* perception of relevance judgment difficulty. Comparison of the average number of documents under each difficulty category between topics revealed similar patterns.

Furthermore, we compared the average number of documents falling into each difficulty category between documents whose first-pass TREC relevance judgments were overturned and those whose first-pass TREC relevance judgments were either not appealed or not overturned, for T103 and T203. In the cases of LAW assessors, the average number of 'difficult' and 'average' documents increased by 1 and 5, respectively, while the average number of 'easy' documents decreased by 6 (out of 50 documents per topic, comparing overturned to non-appealed/non-overturned); in the cases of LIS assessors, the corresponding numbers are 4, 2 and 6, respectively. In order words, there were relatively more documents perceived as 'easy' to judge in the overturned category than in the other category. This finding is consistent with what we revealed through our comparison of relevance judgment accuracy between these two categories of documents.

4.5 Correlation Analysis

In this section, we investigate whether there is any correlation between participants' relevance judgment accuracy, speed, and perceived difficulty.

4.5.1 Accuracy vs. Perceived Difficulty

In order to find out whether relevance judgment accuracy is correlated to perceived relevance judgment difficulty, we computed the overall relevance judgment accuracy of judgments marked as 'difficult,' 'average,' and 'easy' for each topic by each assessor, respectively. This gave us 32 relevance judgment accuracy data points under each difficulty category (except the 'difficult' category, which contains 29 elements since in the other three cases no document was marked as 'difficult' to judge.). Two-tailed Student t-tests on the data show statistically significant difference of the average accuracy between 'difficult' judgments and 'average' judgments ($p = 4 * 10^{-7}$) and between 'difficult' judgments and 'easy' judgments ($p = 10^{-5}$) but not between the other two ($p = 0.58$). The results clearly demonstrate that documents marked as 'difficult' to judge were indeed difficult and hence participants achieved significantly low relevance judgment accuracy with them.

Another interesting finding in this regard is that assessors perceived far less distinction of difficulty between 'average' judgments and 'easy' judgments than between either of them and 'difficult' judgments. In fact, solely based on the relevance judgment accuracy, we do not see any detectable difference between 'average' judgments and 'easy' judgments. One important message emerging from this analysis is that if re-assessment of relevance judgments is to be conducted, those perceived as 'difficult' should demand more attention; cost-benefit analysis between the other two levels of relevance judgment difficulty is perhaps not worthwhile.

4.5.2 Speed vs. Perceived Difficulty

Similarly, in order to find out whether relevance judgment speed (measure as the average number of documents judged per hour) is correlated to perceived relevance judgment difficulty, we computed the relevance judgment speed of documents marked as 'difficult,' 'average,' and 'easy' to judge by each assessor for each topic, respectively. That resulted in three groups of relevance judgment speed data with each containing 32 values (again, the 'difficult' group contains only 29 elements). All three pair-wise two-tail t-tests show statistically significant differences (with all three p values smaller than 10^{-5}), indicating assessors' relevance judgment speed and their perceived difficulty were strongly correlated with each other. Specifically, the easier the assessor feels a document to judge, the faster he can judge it.

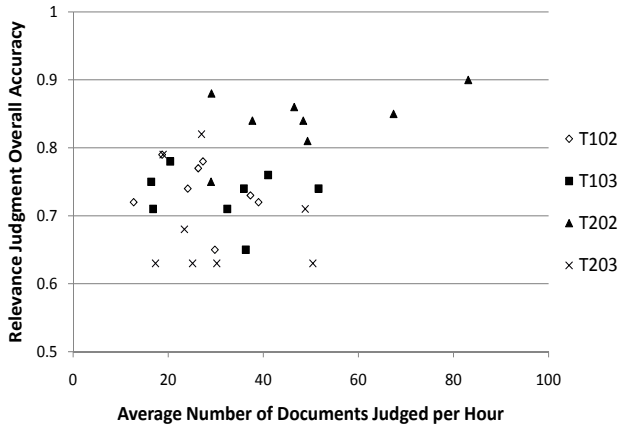

Figure 4: Lack of correlation between relevance judgment accuracy and relevance judgment speed.

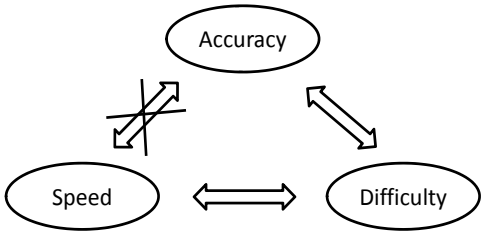

Figure 5: Correlation/Noncorrelation between relevance judgment accuracy, speed, and perceived difficulty.

4.5.3 Accuracy vs. Speed

The scatter plot in Figure 4 shows the relationship (or lack thereof) between the overall relevance judgment accuracy and the relevance judgment speed. The figure is plotted based on the data presented in Section 4.1 and Section 4.3, each of the 32 data points representing the accuracy and the speed of a participant's relevance judgments for a topic. Pearson's correlation coefficient on the accuracy-speed data turned out to be 0.39, which seems to indicate a very weak positive correlation between these two variables. However, examining the four topics individually, we found the accuracy and the speed were negatively correlated for T102 (with a correlation coefficient of -0.50), positively correlated for T202 (with a correlation coefficient of 0.80), and not correlated for the other two topics (with correlation coefficients of approximately 0). Putting these facts together, we do not feel it is reliable to claim a detectable correlation between the relevance judgment accuracy and the relevance judgment speed.

Figure 5 graphically illustrates the correlation between relevance judgment accuracy, speed, and perceived difficulty. In brief, relevance judgment difficulty perceived by participants is strongly correlated to their relevance judgment accuracy and to their relevance judgment speed. Given the data we have, however, there does not seem to exist a detectable correlation between relevance judgment accuracy and relevance judgment speed.

4.6 Influencing Factors

Based on our quantitative analysis presented above and our qualitative analysis of the interview data, we identified the following main factors that influenced assessors' relevance judgments.

Subject matters and subject knowledge are perhaps the most important factor influencing assessors' relevance judgments. Three of the four LAW participants thought the Enron Email topics were more difficult than the Tobacco topics since it demanded more specialized subject knowledge to understand the documents for the former. However, the other participant (LAW4) felt differently mainly because he had studied the Enron case extensively through his law school course work. Two LIS participants also agreed the Enron Email topics were more difficult for the same reason while the other two felt all four topics had about the same difficulty.

All four LAW participants acknowledged the usefulness of their law background and legal service experience for the relevance judgment task in the study. LIS participants, on the other hand, did not feel as strongly that such subject expertise would have helped much. However, all of them felt the guidelines were sufficient to compensate for their lack of legal expertise (in that sense, they were actually acknowledging the usefulness of subject knowledge).

Guidelines and Instructions, particularly the topic-specific guidelines, were viewed as a very useful source of information for participants to make relevance judgments. While acknowledging the difficulty of the subject of T202, all participants felt the topic-specific guidelines compensated greatly their lack of knowledge in that topic area (i.e., finance) for *deciding* the relevance of many documents, but not for *understanding* them. That is, even though assessors judged many documents correctly, they did not feel that they fully understood what these documents are about. In addition, half of the participants mentioned more examples of relevant and nonrelevant documents or more question-answer pairs (as a result of communications between TREC assessors or participating teams and topic authorities) would certainly be helpful. One participant specifically suggested to have a group of teammates work on the task so that more accurate judgments could be made when the relevance can go either way. Another participant wished there would be a topic authority available for questions about specific documents.

Document physical characteristics such as length and legibility and closely related to them, participants' *reading skills*, were also identified as factors influencing their judgments Long documents required special reviewing skills such as first skimming the abstract and the conclusion section or using the term index at the end of such documents if any. Participants reported that some scanned pdf documents had very poor legibility and some unformatted email messages were very hard to read due to things like HTML tags being mixed with the text. For these reasons, participants were concerned that they might have missed some important parts of the text and hence they might have judged these documents incorrectly.

Learning Effects could also have some influence. All assessors 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 and topics in the two collections are quite different.

5. CONCLUSION AND FUTURE WORK

This paper presented our 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. Given the same relevance guidelines and a minimal amount of training, 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. Assessors achieved moderate to substantial agreements on their relevance judgments. Taking into consideration the results from several previous studies that involved only assessors from professional e-discovery firms and/or practicing lawyers, our findings suggest that it is perhaps a viable idea to recruit people without necessarily much legal domain expertise to review documents for e-discovery, especially when relevance guidelines and topic authorities are available for assistance. One argument supporting this idea is that the assessor does not necessarily need to understand the subject matter of a document in order to judge it correctly, as evidenced by our participants' relevance judgments for T202 in the study.

As expected, relevance judgment speed varied markedly among assessors in our study. Our most important finding in that regard, however, is that most often it is only a small number of documents (in a relatively large pool) that slowed down the relevance judgment process. Some of these documents are just very long, so it calls for research on efficient (most likely) machine-assisted methods of reviewing long documents. It would be interesting to look into how the assessors in Grossman and Cormack's study reviewed such documents [9]. Menawhile, it is unclear why assessors spent much time on some short documents, something future researchers should definitely pay attention to. In addition, we did find a detectable correlation between relevance judgment accuracy and speed, hence we recommend relevance judgment speed not be used as an indicator of relevance judgment quality.

Our analysis showed that relevance judgment difficulty is largely a subjective matter. Our most important findings from this line of analysis is that assessors perceived far more distinction of difficulty between 'difficult' judgments and 'average' or 'easy' judgments than between the latter two; 'difficult' judgments were far less accurate while 'average' judgments and 'easy' judgments had comparable accuracy. When time and manpower are a concern (which often should be), deliberation and re-review of judged documents should then better be focused on relevance judgments that assessors perceived obviously difficult to make. For that reason, it is perhaps worthwhile to ask assessors to mark the difficulty of each of their relevance judgments, be it for building e-discovery test collections or conducting e-discovery for a practical purpose.

In our study, documents whose first-pass TREC relevance was overturned turned out more difficult to judge, as evidenced by the lower relevance judgment accuracy, lower agreement rates, lower reviewing speed, and higher perceived level of difficulty by our assessors when judging these documents. These findings suggest that, even without an appeal mechanism (as used by TREC Legal E-Discovery track), it is possible to improve the quality of document review for e-discovery, e.g., by sampling more documents that reviewers spent more time on or marked as more difficult to assess.

Our preliminary analysis of the interview data did reveal several factors that could influence the accuracy of relevance judgments. Subject knowledge is important, but relevance guidelines can sometimes compensate for lack of it to certain extent. Physical characteristics of documents such as length and legibility did have an effect, as suggested by assessors and confirmed by our quantitative analysis. In addition, participants also acknowledged the positive effects that they gained from judging documents for an earlier topic on a later topic if they are from the same legal complaint.

Our study has several limitations. Although we feel the number of documents used in the study was sufficient, the number of topics and the number of assessors are quite limited. For that reason, interpretation and generalization of some of the findings in this study – in particular, those based on comparisons between the law group and the LIS group – should be made with a great deal of caution. Also, since there were no topic authorities available for our study, the relevance judgment accuracy achieved by our participants may not reflect the actual performance of document reviewers for e-discovery in a practical setting.

Our immediate next step is to look closely at documents that assessors spent more time on or marked as difficult to judge. When coupled with the analysis of assessors' relevance judgment notes and interview data, such analysis should provide us more insights into the causes of incorrect relevance judgments as well as help us to identify different relevance relationships and relevance criteria, such as those described in Chu's study [5]. Also, further examination of assessors' relevance judgment accuracy, speed, and perceived difficulty at different time intervals of the relevance judgment process will perhaps provide us with more quantitative evidence for the claimed learning effect.

Another study is underway in which two assessors work as a group to make relevance judgments for e-discovery. We hope that 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, hence complementing previous studies like Yue et al's [21]. Ultimately, findings from these studies will guide us to model more accurately the concept of relevance in e-discovery and to build better search systems and technology supporting information access to business documents.

6. ACKNOWLEDGEMENTS

The author would like to thank Bruce Hedin for providing the official TREC relevance judgment data, Dagobert Soergel for his suggestions for the study design, Ying Sun for her help with the design of the questionnaire and the interview protocol, and the anonymous reviewers and Doug Oard for their valuable comments. This work has been supported in part by the Department of Library and Information Studies, University at Buffalo, the State University of New York through an internal research grant.

7. REFERENCES

[1] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. 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*, pages 667–674, 2008.

[2] Thomas Barnett, Svetlana Godjevac, Jean-Michel Renders, Caroline Privault, John Schneider, and Robert Wickstrom. Machine learning classification for document review. In *The 3rd DESI Workshop at the 12th International Conference on Artificial Intelligence and Law*, 2009.

[3] Jason R. Baron, David D. Lewis, and Douglas W. Oard. TREC-2006 legal track overview. In *Proceedings the Fifteenth Text REtrieval Conference (TREC 2006)*, 2006. http://trec.nist.gov.

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

[5] Heting Chu. Factors affecting relevance judgment: A report from TREC Legal track. *Journal of Documentation*, 67(2):264–278, 2011.

[6] Cyril W. Cleverdon. The effect of variations in relevance assessments in comparative experimental tests of index languages. *Technical Report ASLIB Part 2*, 1970.

[7] William S. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7(1):19–37, 1971.

[8] Efthimis N. Efthimiadis and Mary A. Hotchkiss. University of Washington (UW) at Legal TREC interactive 2007. In *Proceedings the Sixteenth Text REtrieval Conference (TREC 2007)*, 2007. http://trec.nist.gov.

[9] Maura R. Grossman and Gordon V. Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology*, 17(3), 2011. http://jolt.richmond.edu/v17i3/article11.pdf.

[10] Bruce Hedin, Steven Tomlinson, Jason R. Baron, and Douglas W. Oard. Overview of the 2009 TREC Legal track. In *Proceedings the Eighteenth Text REtrieval Conference (TREC 2009)*, 2009. http://trec.nist.gov.

[11] Xiaoli Huang and Dagobert Soergel. An evidence perspective on topical relevance types and its implications for exploratory and task-based retrieval. *Information Research*, 1(12), 2006. http://informationr.net/ir/12–1/paper281.html.

[12] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–74, 1977.

[13] Douglas W. Oard, Bruce Hedin, Steven Tomlinson, and Jason R. Baron. Overview of the TREC 2008 legal track. In *Proceedings the Eighteenth Text REtrieval Conference (TREC 2008)*, 2008. http://trec.nist.gov.

[14] Herbert L. Roitblat, Anne Kershaw, and Patrick Oot. Document categorization in legal electronic discovery: Computer classification vs. manual review. *Journal of the American Society for Information Science and Technology*, 61(1):70–80, 2010.

[15] Ellen M. Voorhees. 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*, pages 315–323, 1998.

[16] Jianqiang Wang and Dagobert Soergel. A user study of relevance judgments for e-discovery. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem*, pages 74:1–74:10, 2010.

[17] Jianqiang Wang, Ying Sun, Omar Mukhtar, and Rohini Srihari. TREC 2008 at the University at Buffalo: legal and blog track. In *Proceedings the Seventeenth Text REtrieval Conference (TREC 2008)*, 2008. http://trec.nist.gov.

[18] Jianqiang Wang, Ying Sun, and Paul Thompson. TREC 2009 at the University at Buffalo: interactive legal e-discovery with Enron emails. In *Proceedings the Eighteenth Text REtrieval Conference (TREC 2009)*, 2009. http://trec.nist.gov.

[19] Peiling Wang and Dagobert Soergel. 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, 1998.

[20] Patrick Wilson. Situational relevance. *Information Storage and Retrieval*, 9(8):457–471, 1973.

[21] Zhen Yue, Jon Walker, Yi-Ling Lin, and Daqing He. Pitt@TREC08: An initial study of collaborative information behavior in ediscovery. In *Proceedings the Seventeenth Text REtrieval Conference (TREC 2008)*, 2008. http://trec.nist.gov.