Effect of Prevalence on Relevance Assessing Behaviour

by

Chandra Prakash Jethani

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2011

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Relevance assessing is an important part of information retrieval (IR) evaluation in addition to being something that all users of IR systems must do as part of their search for relevant documents. In this thesis, we present a user study conducted to understand the relevance judging behaviour of assessors when the prevalence of relevant documents in a set of documents to be judged is varied. In our user study, we collected judgements of participants on document sets of three different prevalence levels. The prevalence levels that we used were low (0.1), balanced (0.5) and high (0.9). We found that participants who judged documents at the 0.9 level made the most mistakes, and participants who judged documents at the 0.5 level made the least mistakes. We did not find a statistically significant difference in judging quality between 0.1 and 0.5 prevalence levels.

Acknowledgements

I would like to thank my supervisor, Dr. Mark D. Smucker for his patient guidance, insightful advices, and constant encouragement throughout the course of my Masters. I would also like to thank Dr. Gordon V. Cormack and Dr. Charles L. A. Clarke for being my thesis readers and providing worthy comments and suggestions that helped improve this thesis.

I thank all my Information Retrieval and Programming Languages group members for the great time I have had with them. My friends in Canada, especially Karanbir Singh Chahal and Gauravdeep Singh Kamboj have made my stay memorable. A big thanks to them.

I would also like to thank anonymous participants who took part in my study and helped me achieve this goal. Finally, I thank Garima Bajwa and Jaspreet Singh Notey for proof reading this thesis.

To my parents.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Information retrieval (IR) has become an important part of our day to day life. We all go to our favourite information retrieval system (commonly known as a search engine) whenever we want to find something on the internet. When we go there, we type in a query for our information need and the system gives us a list of ranked documents. The development of a IR system involves using various retrieval and ranking algorithms to retrieve documents. Once the system is built, it is very important to evaluate it for the quality of documents it returns to make sure that the documents returned by the system satisfy the user's information need. To evaluate the IR system, we need to know whether a document is relevant or non-relevant for a user query. The judgements on documents about their relevance are called relevance judgements (or relevance assessments). Relevance assessors are hired to collect relevance judgements on documents for many search queries and these judgements are used to evaluate new IR systems or re-evaluate existing IR systems after they have gone through any changes.

Traditionally Cranfield methodology is used to evaluate IR systems (Cleverdon, 1967).

In this methodology tuples of query (or search topic), document and relevance judgement are required to evaluate a IR system. Once we have these tuples, for every document returned by the IR system for a query, we check its relevance using the tuples we have. At the end of this process, the number of relevant documents returned for our information need (in this case query) is computed which is then used to calculate the various evaluation measures for the IR system.

Due to the importance of relevance judgements in evaluating IR systems, often relevance assessors are hired to make relevance judgements. These assessors work as if they are working on a IR system, looking for certain information like the normal users of IR systems and judge the relevance of the documents. One of the most well known instances of this process is what is done at National Institute of Standards and Technology (NIST), USA. NIST is a government body that organises a conference called the Text REtrieval Conference (TREC) (Voorhees and Harman, 2005b). The main aim of this conference is to support information retrieval research by providing test collections and advancing research on the evaluation of IR systems. TREC consists of a set of tracks. Each of these tracks focuses on a particular information retrieval task. Participants develop their own retrieval systems and try to solve these tasks using the document collection provided by NIST and return to NIST, a list of the retrieved top documents by their IR system. NIST pools these individual results to create a list of documents to assess the relevance of these documents. NIST hires assessors to judge these documents and produce relevance assessments which are known as qrels. Using these qrels, NIST evaluates the ranked list of documents submitted by participants at a track. A product of a TREC track is usually a new test collection. A test collection consists of a document collection, a set of search topics, and a set of qrels. Over the years, these test collections have been provided to the IR research community to use in their research projects.

Various experiments have been conducted to understand this whole relevance judging process and thereby relevance judgements. Cuadra and Katter Cuadra and Katter (June, 1967) identified various variables which might affect the relevance judgements. Some example of these variables are experience of judges, specification of the task, judging attitude, kind of response required, and so on. Since there are so many variables which can affect relevance judgements, these effects might introduce some errors in relevance judgements. If these erroneous judgements are used to calculate the various evaluation measures then the values will not be accurate. So, the need for calculating correct values of the evaluation measures calls for careful study of the behaviour of assessors and minimization of the effects of variables on relevance judgements.

Another reason that underscores scrupulous study of relevance assessing behaviour is its ability to predict the search performance of users on IR systems. If we have developed an IR system and want to predict how users would perceive the system, we should have prior understanding of the behaviour of users on IR systems.

Another advantage of studying the behaviour of participants is that we can apply this learning in creating better systems for assessors as well as users of IR systems. Systems can be built for assessors which will minimise the effect of the various variables on their behaviour and produce results with less errors. Better IR systems can also be built for users so that they find more and more relevant documents and satisfy their information need.

Similar to Cuadra and Katter's work, many researchers have tried to study the relevance assessing behaviour. In an experiment conducted by Smucker and Jethani (2010a), they studied how precision of a rank list affects the human search performance; that is the number of relevant documents found in a given amount of time.

Precision is the fraction of retrieved documents that are relevant. Their study was divided in two phases and in both those phases, they investigated human performance at two different precisions - low (0.3) and high (0.6). In phase 1, participants alternated between judging summaries and documents. In phase 2, participants were presented with a modern web search like interface to click on summaries and judge documents. Even though they found out that at high precision ranking, participants find more relevant documents in a given amount of time, the number of relevant documents found for high precision system was not twice of the same found for low precision system. The study proposed that a possible cause of the difference between the performance predicted by precision and actual performance is that participants exhibit changes in behaviour depending on the precision of the result list. When judging a high precision list, the true positive rate of participants was found to be lower than when they worked on low precision list. True positive rate is defined as fraction of relevant documents marked as relevant. So while judging the high precision lists, participants were less likely to judge relevant documents as relevant, compared to when judging the low precision lists. This effect was stronger for phase 2 than for phase 1. There was no effect visible on the false positive rate of the participants. False positive rate is defined as the fraction of non-relevant documents marked as relevant.

Research studies have been conducted to study the behaviour of assessors in clinical psychology and visual search tasks. Classical examples of tasks in these fields are cancer detection and airport luggage screening. In these tasks, user typically sees a lot of images with some of them having targets (tumour in cancer detection or sensitive material in airport screening) in them. The task is to find targets in these images. The fraction of images with targets present in them is defined as prevalence. Various research studies have shown that this prevalence of targets affects the behaviour of participants and detection of targets. This effect on the behaviour of participants due to the prevalence is referred to

as *prevalence effect*.

In one of the experiments conducted by Wolfe et al. (2005), prevalence effect in artificial baggage-screen task was studied. The participants in the experiment looked for tools among various objects by looking at the scanned images. The prevalence of the tools was varied and the behaviour of participants was studied at each prevalence. They reported that miss errors (saying "no" when tools are present) increased as the prevalence of the tools decreased. In other words the rate at which participants missed the tools was more at low prevalence than at high prevalence.

Wolfe and Van Wert (2010) conducted two experiments using simulated baggage search to study the effect of changing prevalence on the user behaviour. In the first experiment participants were asked to look for targets in balanced (50%) and high (98%) prevalence tasks. They found that false error rate of the participants increased as prevalence of targets increased. In second experiment, target prevalence was varied sinusoidally from high to low and back to high. Results of this experiment showed that varying prevalence affected participants' decision criterion. Decision criteria is defined as how liberal or conservative an assessor is while judging the relevance of the documents. If a participant is liberal, he is willing to commit false positive mistakes but he does not want to miss any targets. A conservative participant is willing to miss out on targets but does not want to commit any false positive mistakes.

Task of finding relevant documents from a list or set of relevant and non-relevant documents is very similar to finding targets in visual search tasks. Precision which is ratio of relevant documents to relevant and non-relevant documents is also similar to prevalence. Findings from the study by Smucker and Jethani (2010a) and research done to study the effect of prevalence on assessor behaviour in other fields motivated us to study the relevance assessing behaviour of users due to varying precision in information retrieval.

To find out the effect of varying precision on user assessing behaviour, we conducted a controlled user study. Our study was conducted in three phases. First phase referred to as tutorial phase was to train participants on document judging process. In the second phase which we refer to as qualification phase tested participants on their judging ability. Participants who cleared this phase were asked to judge documents in the final task phase. We used 8 topics from AQUAINT collection of 1,033,461 newswire documents used in TREC Robust Retrieval Track (Voorhees, 2005a) for this study. 2 topics were used for tutorial and qualification phases and 6 topics for task phase.

Our study started with participants filling their demographic information and reading various instructions about study. Then they judged 10 documents in tutorial phase for two topics at precision 0.5. While judging the documents, topic was visible to them all the time. After every judgement they made they were told the reason of relevance of the document so that they learn what to look for in a document to judge its relevance. After they finish the tutorial phase, they were asked to judge 20 document for the same two topics at precision 0.5. This time they were not told whether their judgement about the document's relevance was correct or incorrect. Participants had to judge the relevance of 14 out of 20 (70%) documents correctly in 30 minutes to qualify for the task phase. In our study total 55 participants took part in tutorial and qualification phases. All of them qualified for the task phase, but 1 participant did not continue for the task phase.

In task phase, each of the 54 participants was asked to judge documents for a precision. The documents were judged for 3 different precisions. Precisions we used for this study were low (0.1), balanced (0.5) and high (0.9). After the qualification phase, each of the 54 participants was assigned randomly to one of three groups of precisions. Each participant judged 40 documents per topic for 2 topics for a precision. Participants were presented with a pre-task questionnaire before they started judging documents for a topic. This question-

naire asked questions about their familiarity and knowledge about the topic. After they finished judging the documents for the topic, they were presented with a post-task questionnaire in which they were asked questions about their judging experience. We captured their judgements about the relevance of the documents, time spent on each document and their responses to pre-task and post-task questionnaires. There were 18 participants per precision group. Behaviour of participants was studied by calculating various measures like true positive rate, false positive rate, accuracy, ability to discriminate and criterion in all three phases of the study and conclusions were drawn from these measures.

Based on our analysis in this thesis, we make the following contributions:

- We show that prevalence affects the quality of participants' judgements. Participants working on precision 0.5 have the highest true positive rate and lowest false positive rate. Increasing the precision to 0.9 hurts the quality of the judgements produced. Both true positive rate and false positive rate are worse at precision 0.9 as compared to precision 0.5. Even though we did not find statistically significant difference between the judging quality at precision 0.1 and precision 0.5, both true positive rate and false positive rate were worse at precision 0.1 than the same at precision 0.5.

- Prevalence also has an effect on amount of time spent on the documents while judging. If we consider NIST judgements as the source of truth, participants spend more time on relevant documents than on non-relevant documents. On the other hand if we consider participant's judgement as the source of truth, this behaviour is reversed for participants working on 0.9 precision. They spend on more time on non-relevant documents. As the prevalence increases, the time spent on NIST relevant documents decreases.

- We also show that participants speed up as they get used to the task of judging

documents. Initially they take some time to get used to the task. Once they learn about the task, they start working at their own pace and this pace increases to a certain extent.

- In this experiment, we saw little to no correlation between judging speed and quality of the judgements. It appears that participants work at a pace needed to produce an expected quality.

The rest of the thesis is organized as follows: we discuss related work done in the area of relevance judging behaviour, then the design of our user study is presented followed by results and discussions. Finally conclusions drawn from this study are laid out.

Chapter 2

Related Work

There has been a considerable amount of work done to understand the behaviour of assessors who judge documents. In order to understand relevance judging behaviour of assessors, it is important to define what we mean by relevance. Notion of relevance has been in existence since the first library when one had to find relevant information. Over the years, modern researchers have tried to define relevance in various ways. Researchers have also identified criteria which affect relevance and various modes to express relevance.

Commonly relevance can be thought of as a relation between portions of *stored information* and *information need* (Cooper, 1971). We classify part of the stored information as "relevant" and rest as "non-relevant" depending upon which part satisfies our information need. Starting from this simple idea, (Cooper, 1971) defined relevance in a logical form. According to him, a statement or a sentence is relevant to a question asked if the statement belongs to a set of non-redundant statements, from which an answer to that question follows logically. Wilson (August 1973) inspired the idea that relevance arises from the situation or the task and termed this as situational relevance. He said that concept of rel-

evance is related to the actual users, so user's stock of knowledge, situation and personal concerns of users should be taken into account. He defines situational relevance as the relation between an information object and information recipient's individual and personal view of the world and his situation in it. An information object is situational relevant for a user if it brings about a change in the user's view of his situation. A information retrieval system user has to go through many potentially relevant documents and determine the relevance of a specific document in the context of their current information needs. This process is termed as *document triage*. Researchers have tried to understand this process by conducting various user studies. A recent study have shown that users create first impression about the relevance of the document in a very short span of time (Buchanan and Loizides, 2007) and if they spend further time on document it is intended to confirm this initial impression rather than test it systematically. While judging the relevance of documents, users rely mostly on the document features like title, abstract and heading text (Cool et al., 1993), (Saracevic, 1969), and pay relatively low importance to main content of the document.

There are various parameters which affect the relevance. Cuadra and Katter (June, 1967) identified 38 variables which might affect relevance of documents. They categorized these variables into five groups. First group had variables related to the document. This included subject matter, diversity of content, difficulty level, amount of information, textual attributes, and so on. Second group had variables which define the information requirement statement, namely subject matter, diversity of content, difficulty level, specificity, textual attributes, and so on. Third group had variables related to the person judging the documents. This included assessor's knowledge and experience, cognitive style, biases, judging experience, judging attitude, error preference, etc. Variables related to judgement conditions were placed in the fourth group which included amount of time permitted

for judging, order of presentation, size of document set, specification of the task, and so on. And the fifth group had variables related to available modes of expression to report judgements. These were type of scale, number of rating categories, ease of use of these expressions and so on. After careful analysis, they selected a list of variables to study from this comprehensive list and conducted fifteen user studies. They reported on the effects of these variables on assessor behaviour. Order of presentation of documents which deals with how the flow of relevant and non-relevant documents might affect assessor's relevance judging behaviour, was part of the list of variables they identified, but its effect on assessors' relevance judging behaviour was not studied. Our work in this thesis very closely relates to the effect of order of presentation of documents on assessors' behaviour.

Effect of relevance judgements collected from different assessors (or judges) on IR system evaluation has been studied by various researchers. Bailey et al. (2008) ran experiments to find out the effect of different type of judges on relevance judgements. They recruited participants from three different populations to simulate three types of judges - participants in "gold standard" were the originators of the topics as well as experts in the information seeking task, participants in "silver standard" were the task experts but did not create the topics and the participants in "bronze standard" were neither the originator of the topics nor the experts in the tasks. They found that there is low level of agreement among different types of judges and agreement is even less between "gold" and "bronze" standard than "gold" and "silver". They also found that "bronze" standard judges may not be a reliable substitute for "gold" standard. Judgements of "gold standard" population are closer to that of "silver standard" than they are to the judgements of "bronze standard". Lesk and Salton (1968) conducted a study to verify if the differences in relevance assessments affect retrieval systems' rankings. In their experiment, the sources of different relevance assessments were different judges. They collected relevance judgements from query authors and

query non-authors on document abstracts. From these two sets of judgements, they produced two sets of judgements more by taking the intersection and union of these two sets. In all, 4 sets of judgements were available for 48 queries on 1260 documents to compare rankings of systems. They found that even though there was only 30% agreement among the 4 relevance judgement sets, each of these sets of relevant judgements produced the same relative ranking of different processing methods. Similar results were observed by Burgin (1992) in a separate experiment. In experiments of studying differences between nineteen different indexing methods, Cleverdon (October, 1970) used four independent sets of judgements. He found that even though the judgements used were from different sources, they did not alter the ranking of indexing methods. Similarly, Voorhees (1998) ran experiments to study the effect of inter-assessors disagreements on judgements on TREC collections. She used the judgements for TREC-4 datasets provided by NIST and judgements for part of TREC-6 dataset provided by NIST and University of Waterloo judges. What she found was that even though there were disagreements among the judgements, these disagreements did not affect the relative rankings of the systems evaluated using these judgements. She also pointed out that the disagreements might affect the relative rankings of the systems if the systems (runs) compared used significant amount of relevance feedback or the number of relevant documents for topics is low. Harter (1996) did an extensive survey of the past literature of empirical studies on how variations in relevance assessments affect measures of retrieval effectiveness. In his own words, "All find significant variations in relevance assessments among judges. And all conclude that these variations have no appreciable effect on measures of retrieval effectiveness, that is, in the comparative ranking of different systems."

As mentioned in the introduction, Smucker and Jethani (2010a) found that when judging a high precision list, true positive rate of participants was found to be lower than when

they worked on low precision list and there was no effect visible on the false positive rate of the participants. Findings from this work prompted us to look into the effect of precision on assessors' behaviour in greater detail and the work presented in this thesis is a step into that direction. In addition to these findings, as part of their study, Smucker and Jethani (2010a) collected data to capture the mood of the participants, their perceived difficulty of the task and various other parameters (Smucker and Jethani, 2010b). By analysing this data they were able to conclude that participant's perceived the task less difficult when they were working on high precision rank list than when they were working on low precision rank list. Higher precision increased participants' enjoyment and it also influenced their ability to concentrate. As a by-product of their study (Smucker and Jethani, 2010a), Smucker and Jethani were able to compare behaviour of NIST assessors with that of participants in the study (Smucker and Jethani, 2011). They selected the documents for which there were 10 or more judgements and classified them judged as relevant by the participants if number of relevant judgements on them were more than non-relevant judgements. They considered these judgements as a gold standard to compare NIST assessors and study participants. They found that the participants' true positive rate was as good as NIST assessors' true positive rates, but NIST assessors had better false positive rates. True positive rate is defined as the fraction of relevant documents judged as relevant and false positive rate is defined as fraction of non-relevant documents judged as relevant.

The behaviour of assessors has been studied in other forms of information retrieval tasks. Grossman and Cormack (2011) studied if the difference in opinion on the relevance of legal documents is due to the ambiguity or inconsistency in applying the criterion for responsiveness to particular documents, or is it due to human error. Their experiment was based on TREC Legal Track 2009 corpus and judgements collected from the assessors (law students, attorneys and "topic authorities"). From the documents where "Topic Authori-

ties" and other assessors did not agree, they randomly selected a sample of documents and one of the authors (Cormack) re-assessed them for their responsiveness to their respective topics. For one of the topics, the re-assessment clearly disagreed with the Topic Authority (TA) in one case, and was "arguable" in nine other cases. One of the authors (Grossman) of the paper was the TA for this topic. The ten documents were presented to the TA for fresh reconsideration, in random order, with no indication as to how they had been previously judged. In 5 out of 10 cases, TA changed her opinion; in 3 of these 5 cases TA was found incorrect in her judgement first time and in rest of the 2 cases, position became "arguable". Similar analysis was done for another topic that showed similar results. In summary, they report that vast majority of cases of disagreement among assessors are a product of human error rather than the documents that fall in some "grey area" of judgements. In the work by Wang and Soergel (2010) and Wang (2011), authors asked 4 law students and 4 library and information studies (LIS) students to judge responsiveness of documents for litigation purposes. The documents were taken from two collections used in TREC Legal Track. One of them was the test collection of MSA tobacco documents and other collection was 2009 TREC Enron Email collection. After every judgement a participant made, he was asked to rate the judgement at three scales: 'difficult', 'average' and 'easy'. They found the LIS students judged relevant documents as accurately as the law students. LIS students judged non-relevant document slightly less accurately than law students. They also found that participants perception of difficulty is a subjective matter. But participants could distinguish more accurately between 'difficult' judgements and 'average' or 'easy' judgements than between the latter two. This was evident from the fact that 'difficult' judgements were less accurate while 'average' and 'easy' judgements had comparable accuracy.

In the last few years, due to the growing popularity of Amazon Mechanical Turk [1] (AMT) and CrowdFlower [2], a new concept called *crowdsourcing* [3] has also become a way of collecting judgements quickly and at a very low cost. Researchers have started using crowdsourcing paradigm to conduct their user studies and they are trying to compare if this concept can replace the traditional in-lab user studies in terms of quality of the assessor judgements (Alonso et al., 2008).They employ different techniques to control the quality. Kittur et al. (2008) discusses that care must be taken while designing the task for crowdsourcing, especially for tasks which require qualitative or subjective judgements. They show that properly designed task can produce very high quality judgements. Similar results were achieved by Alonso and Mizzaro (2009) in their experiment. To devise the mechanism to control the quality of judgements, Le et al. (2010) ran a crowdsourcing experiment on Amazon Mechanical Turk. The tasks were set up using CrowdFlower and the aim of the tasks was to categorize the documents in one of the categories given. In their experiment, participants were trained on document judging process before they started actual tasks and then were tested sporadically while judging the documents. The distribution of various category labels across all the training labels was kept uniform. It was found that uniform distribution of category labels across training data labels produced judgements with the most optimal precision. Authors pointed out that this was because of the lack of formation of bias towards one category displayed by the participants when the labels were equally distributed across the training labels. Periodic inclusion of training data while participants judged the actual documents, helped participants to learn as they judged the documents. In an another study, Snow et al. (2008) used AMT to collect annotations on natural language tasks. They report high agreement between annotations

[1] www.mturk.com/

[2] www.crowdflower.com/

[3] http://www.wired.com/wired/archive/14.06/crowds.html

collected from AMT non-experts and existing gold standard annotations and also stress on the importance of carefully designed tasks.

Studies to understand judging behaviour of assessors have also been conducted in clinical psychology, current biology, visual search tasks and other streams of science. In these fields, assessors are involved in looking at a lot of images and flagging images with targets in them. Some of the examples of such targets are tumours in cancer detection tasks, sensitive material in airport screening tasks, etc. Fraction of images with targets in them is referred to as prevalence and the effect of this prevalence on assessors' behaviour is referred to as *prevalence effect*. Gur et al. (2003) studied the prevalence effect on assessors using radiology images under laboratory conditions. In their experiment, prevalence of targets was varied from 2% to 28%. There were 8 radiologists, 4 fellows and 4 third year residents who took part in the study. Participants' responses were collected in the form of a check-list type responses which varied from 0 (no targets) to 100 (definite targets). According to this study, under laboratory condition, prevalence effect, even if it was present, it was likely to be small in magnitude; hence, it would not likely alter conclusions derived from such studies. They also said that their study result might not be generalized to general clinical environment or to any reading conditions that did not require a formatted check-list type response. Wolfe et al. (2005) studied prevalence effect in artificial baggage-screen tasks. The participants in their study looked for tools among various objects by looking at the scanned images. The prevalence of the tools was varied and the behaviour of participants was studied at each prevalence. They reported that miss errors (saying "no" when tools are present) increased as the prevalence of the tools decreased. In other words the rate at which participants missed the tools was more at low prevalence than at high prevalence. The reason they reported for such behaviour was due to the variation in reaction time of participants during the course of experiment. Participants require a threshold for quitting

when no target has been found. This threshold is constantly adjusted as the participants go through the screening tasks. At high prevalence, reaction times are longer when target is absent than when the target is present. At low prevalence, reaction times are lower when target is absent than when the target is present. This behaviour of participants brings quitting threshold down at low prevalence resulting in more miss errors. In an another study of visual search tasks, Wolfe et al. (2007) found that target miss error rates (failing to notice a target) increase and false error rate (saying "yes" when target is absent) decrease at low prevalence. Fleck and Mitroff (2007) conducted an experiment to figure out why observers miss targets at low prevalence. In their experiment, they gave half of the participants an option to correct the judgement they made on the previous trial (correction condition). Other half did not have this option (no-correction condition). They reported that the participants who were given the option of correcting their previous judgement could correct miss errors. They also say that participants did not miss the targets because of prevalence effect. Miss error rates are due to response-execution error. Observers know that the target was present, but they just respond too quickly. Giving an option to correct their response on previous trial can reduce the miss errors dramatically. In one of their recent studies, Wolfe and Van Wert (2010) conducted two experiments using simulated baggage search to study the effect of changing prevalence on the user behaviour. In the first experiment, participants were asked to look for targets in balanced (50%) and high (98%) prevalence tasks. They found that false error rate of the participants increased as prevalence of targets increased. In second experiment, target prevalence was varied sinusoidally from high to low and back to high. Results of this experiment showed that varying prevalence affected participants' decision criterion. Decision criteria is defined as how liberal or conservative an assessor is while judging the relevance of the documents. If a participant is liberal, he is willing to commit false positive mistakes but he does not

want to miss any targets. A conservative participant is willing to miss out on targets but does not want to commit any false positive mistakes.

Chapter 3

Methods and Materials

3.1 Study Design

We conducted a user study in order to understand the prevalence effect in relevance judging behaviour of assessors. This chapter explains the design of our study and reasons for the decisions made during the design phase.

We conducted a user study of 54 participants who judged documents for 3 different precision levels. Before participants were asked to judge the documents in our study, all were trained and tested on the document judging process. Participants started with signing consent forms, reading instructions, clearing a quiz on instructions and filling out demographic information. Then, they were trained on document judging process. We refer to this phase as tutorial phase in this thesis. In this phase, participants judged the relevance of 5 documents one at a time for each of the 2 search topics. The precision of this 10 document set was 0.5 and all the participants saw the same documents in the same order. All the documents for first topic were presented one after the other followed by the

documents for the second topic. After every judgement they made about the relevance of these documents, they were told why the document was relevant or non-relevant for a search topic. We recorded their judgements and time spent on each of the 10 documents. After the tutorial phase, we tested them on 10 documents for each of the same 2 topics. This phase is referred as qualification phase in this thesis. This phase was same as the tutorial phase except that in this phase, we did not inform participants whether the judgements they made were correct or not. Participants who achieved a desired level of accuracy in a given amount of time in qualification phase qualified for the full study.

In the full study, qualified participants were divided randomly into three groups of 18 participants each. Each of the groups worked on one of the three precision sets. The precisions used in this study were 0.1, 0.5 and 0.9. We refer to this phase as the task phase in this thesis. In this phase, participants judged sets of 40 documents for each of the 2 search topics. These two topics were different from the ones used in tutorial and qualification phases. Precision of both documents sets (one for each search topic) judged by a participant was same. We collected their judgements and time spent on each of the 80 documents. Participants also filled pre and post task questionnaire before and after judging the documents for a search topic.

In the next few pages, we give a description of the data set we used for the study and, then, of each of the phases, in detail.

3.1.1 Collection and Search Topics

For this study, we used 8 topics from the AQUAINT collection of 1,033,461 newswire documents used in TREC Robust Retrieval Track (Voorhees, 2005a). The documents for this collection are obtained from the New York Times, Associated Press and Xinhua News

TREC Topic 427

Title: UV damage, eyes

Description: Find documents that discuss the damage ultraviolet (UV) light from the sun can do to eyes.

Narrative: A relevant document will discuss diseases that result from exposure of the eyes to UV light, treatments for the damage, and/or education programs that help prevent damage. Documents discussing treatment methods for cataracts and ocular melanoma are relevant even when a specific cause is not mentioned. However, documents that discuss radiation damage from nuclear sources or lasers are not relevant.

Figure 3.1: Example of a TREC Topic.

Agency.

A TREC topic comprises of a title, description and a narrative. Description gives the general guidelines of what would make a document relevant to the topic. Narrative lists small details and/or any exceptions to what is written in the description section. An example of a sample TREC topic is shown in figure 3.1.

For this study, we chose a topic's description and narrative, combined them and presented along with topic's title to the participants as a criterion to judge if a document is relevant or non-relevant to the topic. Example of a topic used in the study is shown in figure 3.2. The complete list of topics we selected for this study is given in table 3.1. We

TREC Topic 427

Title: UV damage, eyes

Description: Relevant documents will discuss the damage ultraviolet (UV) light from the sun can do to eyes. A relevant document will discuss diseases that result from exposure of the eyes to UV light, treatments for the damage, and/or education programs that help prevent damage. Documents discussing treatment methods for cataracts and ocular melanoma are relevant even when a specific cause is not mentioned. However, documents that discuss radiation damage from nuclear sources or lasers are not relevant.

Figure 3.2: Example of a combined TREC Topic used in the study.

Number	**Topic Title**	**Relevant**	**Non-Relevant**
310	Radio Waves and Brain Cancer	65	709
336	Black Bear Attacks	42	553
362	Human Smuggling	175	471
367	Piracy	95	526
383	Mental Illness Drugs	137	408
426	Law Enforcement, Dogs	177	620
427	UV Damage, Eyes	58	425
436	Railway Accidents	356	343

Table 3.1: 8 topics used in the study and the number of NIST relevant and non-relevant documents for each of the topics.

used relevant judgements (qrels) provided by NIST for this study. NIST assessors judged documents to be non-relevant, relevant and highly relevant. For this study, we combined relevant and highly relevant documents and treated them as relevant documents.

All the 8 topics used in this study were also used in the study conducted by Smucker and Jethani (2010a), so we had prior knowledge about the nature of these topics which aided us in choosing the topics for our study.

3.1.2 Tutorial and Qualification Phases

Participants started with reading introduction of the study. The introduction included instructions on how to judge the documents and also information like number of documents to judge in tutorial phase, qualification phase and finally in the task phase. At the end of the introduction, participants were asked to go through a quiz which tested them on how well they understood the instructions. Participants could take as many attempts as they wanted to clear the quiz. There was no penalty for this as it was just meant to familiarize them with the study instructions. The instructions also stated that they were required to judge the documents as quickly as possible with high enough accuracy.

Once they cleared the quiz, they were presented with 5 documents for each of the 2 topics to judge. The topic and its description were visible to a participant at all times during the study. After every judgement they made, they were told whether they judged the document correctly or not and they were also explained the reasoning behind correct judgement. This way we expected them to learn to judge documents by the end of the tutorial. At the end of the tutorial, their performance was reported to them.

Once done with the tutorial phase, participants were presented with 10 documents for each of the 2 topics to judge for the qualification phase. These topics were same as topics

used in tutorial phase. This time they were not told whether their judgements were correct or not. At the end of the qualification task, participants who judged at least 14 out of 20 (70%) documents correctly in no more than 30 minutes were qualified for the task phase.

In both tutorial and qualification phases, all participants judged same documents in the same order. In other words, there was no difference between tutorial and qualification phases for any participant. This was to ensure that all the participants received same training.

Construction of Documents Sets for Tutorial and Qualification Phases

We planned that tutorial and qualification phases should take about 40 minutes to complete, should give participants a good amount of training and should also be able to test them before they start doing the actual search tasks. From the study by Smucker and Jethani (2010a), we know that an average user does not take more than 1 minute to judge the document. So, we decided to train them on 10 documents and test on 20 documents considering that the users will spend 10 minutes in going through the instructions and filling out demographic information.

The aim of these two phases was to train and test participants on whether they can distinguish between relevant and non-relevant documents. We chose documents for these two phases from the topics: *mental illness* (topic id 383)and *railway accidents* (topic id 436). We chose *mental illness* as one of the topics because it is easy to distinguish between relevant and non-relevant documents for this topic but it requires one to read deeply, as sometimes the information is hidden deep in the documents. For *railway accidents*, it is easy to distinguish between relevant and non-relevant documents as well and it does not require one to read deeply. Information in the documents is on the surface. There was one

more topic, *Piracy*, for which it was easy to distinguish between relevant and non-relevant documents and it required deep reading to judge. But, since it was similar in nature to *mental illness* and both of them required a lot of time to judge, we chose one of them for tutorial and qualification phases and other for task phase, to balance the time during different phases of the study.

During tutorial phase, each participant judged the relevance of 5 documents for each of the 2 topics, 10 in total. In qualification phase, each participant judged 10 documents for each of the 2 topics, 20 in total.

To generate these document sets for two topics, we randomly selected 8 relevant and 7 non-relevant documents for the first topic. We randomly selected 3 relevant documents from relevant list and 2 non-relevant documents from non-relevant list and shuffled them to get a list of 5 documents for this topic for tutorial phase. Remaining 10 documents were shuffled to generate a list for 10 documents for qualification phase for this topic. We repeated the same process for the second topic but started with 7 relevant documents and 8 non-relevant documents and selected 2 and 3 documents from relevant and non-relevant documents respectively for tutorial phase and rest for the qualification phase. In the end, number of relevant and non-relevant documents were balanced in each of the tutorial and qualification phases (a precision of 0.5).

3.1.3 Task phase

In the task phase, participants were divided in three groups with 18 participants in each group. The division was done randomly. All participants in a group judged 40 documents for each of the 2 topics for a given precision. Three precision levels used in this study were 0.1, 0.5 and 0.9. For a given participant, precision of the sets were same for both topics.

Participants started with answering a short pre-task questionnaire on the topic. This questionnaire asked questions about their familiarity with the topic, their perceived difficulty of the topic, their interest in learning about the topic, etc. Answering these questions before starting to judge the documents for the search topic may have forced the participants to read the topic and its description and this may have prepared them better for the judging task. Participants ended their task by answering a post-task questionnaire in which they were asked questions about their experience about judging the documents for that topic. All these questions were taken from Bailey et al. (2009) except for one question in post-task questionnaire where we provided a text-box to allow participants to report any issues they may have encountered during the course of study. These questionnaires are displayed in section B.13 of Appendix B in the appendices chapter.

In the study, each participant judged 40 documents for each of the 2 topics. These 2 topics were different from the ones used in tutorial and qualification phases and were assigned from the list generated in the way explained in the next section. There was no time limit to judge these documents but it was instructed to the participants to work as quickly as possible while making as few mistakes as possible.

Construction of Document Sets for Task Phase

We chose 6 topics for this phase of the study. These were all the topics in 3.1 except *mental illness* and *railway accidents*. To create document sets for each of the 6 topics, we separated relevant documents and non-relevant documents for that topic using qrels available from NIST. To generate a document set for a given precision, we selected precision$\times$40 number of relevant documents and (1 - precision)$\times$40 number of non-relevant documents randomly. Then, we randomly shuffled these 40 documents to create a set. We repeated this process 10 times to create 10 document sets per topic per precision.

Balanced Design

We wanted to balance topics and precisions across two tasks in the task phase. This was achieved by devising the following strategy. We started with 6 alphabets (A, B, C, D, E and F) and arranged them in a 6×2 matrix such that no row has the same alphabet in both the columns and each column has all 6 alphabets. There can be only 5 such matrices. For each of these matrices, we randomly permuted the rows and columns and then randomly assigned 6 topics to these 6 alphabets. In this way, 5 different matrices of topics were obtained, having 2 topics in a row with no row having same topic in both columns and each column having all 6 topics. As explained further, first matrix out of these 5 matrices was used to design the first block of 18 participants.

Each row in the matrix represented two topics to be worked on by a participant in the order of columns. We assigned these 6 rows to 6 participants who did 0.1 precision sets. In the same way, this matrix was assigned to next 6 participants who did 0.5 precision sets and similarly for next 6 participants who did 0.9 precision sets. This completed a block of 18 participants. Thus, in a block of 18 participants, each topic was worked upon twice at each precision level. Two out of 10 different document sets (as described above), for a given precision and topic, were used in a block of 18.

Second and subsequent blocks of 18 participants were designed in the similar manner using rest of the matrices. Three blocks of 54 participants were completed for this study.

3.1.4 Participants

First step in recruiting participants for the study was to get the approval from university's office of research ethics. After getting the approval, we sent an email to a university wide

graduate student mailing list (see Appendix A). We selected participants on first come first serve basis. In all, 55 participants took part in the study and all of them qualified for the full study. One of the participants chose not to proceed with the study after clearing the qualification phase. Consequently, this participant's data was removed from the study. So, the analysis presented in this thesis is based on the remaining 54 participants' data. When participants started the user study, we collected data about their demographics. Participants were asked various questions about any IR training that they might have received and their English fluency. Please see section B.3 in Appendix B for the complete demographic questionnaire we presented to participants. The participants consisted of 30 males and 24 females. All the participants were students - 51 graduate students and 3 undergraduate students. 40 were science, technology, engineering, or mathematics students. The other 14 students identified themselves as affiliated to arts or other disciplines. The median age was 25, with the minimum as 21, and maximum as 41. Out of 54 participants, 36 of them felt they were fast readers and 18 were neutral. All participants considered themselves fluent speakers of English.

3.1.5 Study Interfaces

We designed various user interfaces to display instructions, topics and documents to participants. We will present two most important interfaces in this section. All other interfaces are shown in the appendices chapter.

Figure 3.3 shows the user interface used in the tutorial and qualification phase to display a topic.

Figure 3.4 shows the way a document was displayed to participants in all three phases. The interface instructed users to judge the relevance of the document to the topic. The

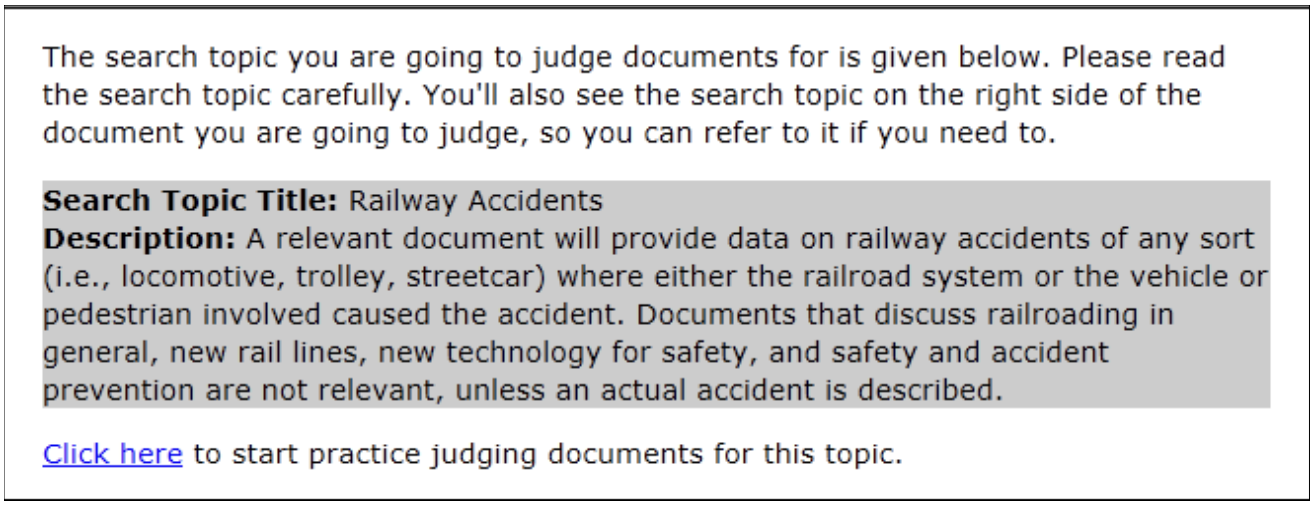


Figure 3.3: User interface for a topic displayed in tutorial and qualification phase.

topic was visible on the right side of the document for the entire task. Search topic's title terms were highlighted in the document text. To see the next document, participants had to make a relevance judgement. Once they had decided about the relevance of the document, participants could report their decision using the two buttons provided on the top of the document.

3.2 Analysis of Rates

The task of judging documents can be mapped to classic signal detection task with yes/no decisions. When the judging task is mapped to signal detection task, we get to use measures like true positive rate (TPR) and false positive rate (FPR) to study performance of participants. Even though accuracy is not a good measure to predict the performance when relevant and non-relevant documents are not balanced, we used it wherever possible.

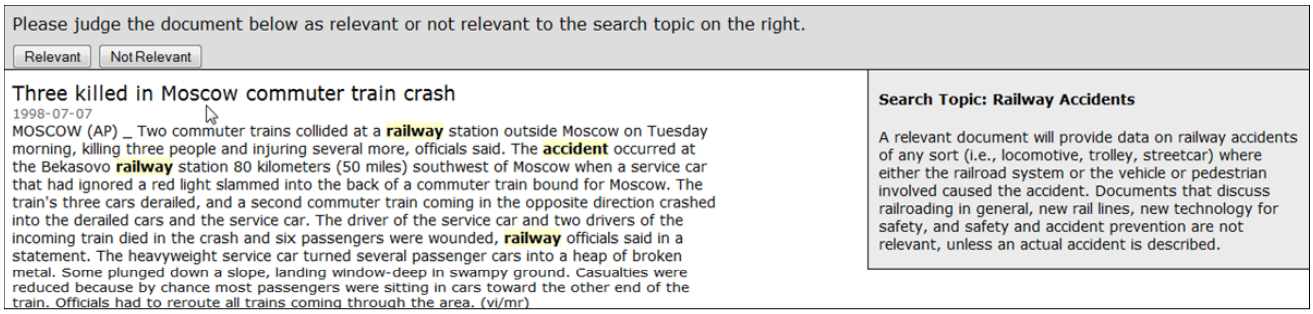


Figure 3.4: User interface for displaying a document.

The true positive rate is measured as:

$$TPR = \frac{|TP|}{|TP| + |FN|} \tag{3.1}$$

and the false positive rate as:

$$FPR = \frac{|FP|}{|FP| + |TN|} \tag{3.2}$$

and accuracy is:

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |FP| + |TN| + |FN|} \tag{3.3}$$

where TP, FP, FN, TN are from table 3.2.

For our experiment, the ideal way of computing true positive rate, false positive rate and accuracy for every precision is to calculate them for every document judged and and then average it for that particular precision. But, for precision 0.1 there are very few judgements for documents that are relevant and similarly for precision 0.9, there are very few judgements for the documents that are non-relevant. Since our data points are very

	NIST Judgement	
Participant Judgement	**Relevant (Positive)**	**Non-Relevant (Negative)**
Relevant	TP = True Positive	FP = False Positive
Non-Relevant	FN = False Negative	TN = True Negative

Table 3.2: Confusion Matrix.

low, we used pooling of judgements to calculate pooled true positive rate, pooled false positive rate and pooled accuracy (Macmillan and Creelman, 2005). Using this approach, to calculate pooled true positive rate for a given precision, we computed total number of relevant documents which were judged as relevant and divided it by total number of relevant documents used in the study for that precision. Similarly, to calculate pooled false positive rate for a precision, we computed total number of non-relevant documents which were judged as relevant and divided them by total number of non-relevant documents used in the study. Pooled accuracy was calculated in a similar fashion. The mathematical formulae remain the same as in equations 3.1, 3.2 and 3.3 and the TP, FP, FN, TN are calculated across all the documents for a given precision.

3.3 Analysis of Time

Similar to the rates, we calculated average time spent on a document at a given precision by dividing total time spent by total number of document judged.

In mathematical form,

$$Average\ time\ spent\ on\ a\ document = \frac{Total\ Time\ Spent\ on\ Judging\ Documents}{Total\ Number\ of\ Documents\ Judged} \quad (3.4)$$

3.4 Analysis of Assessor Criterion and Ability to Discriminate

Since we have modelled document judging task as a signal detection task, we used two other measures from signal detection theory to understand the behaviour of participants in our study. These are called assessor's ability to discriminate which is defined as:

$$d' = z(TPR) - z(FPR) \tag{3.5}$$

and assessor's criterion defined as:

$$c = -\frac{1}{2}(z(TPR) + z(FPR)) \tag{3.6}$$

where the function z is the inverse of the normal distribution function (Macmillan and Creelman, 2005).

Greater the value of d', greater is the ability to discriminate between relevant and non-relevant documents. d' value of 0(zero) indicates random behaviour.

A criterion represents how liberal or conservative an assessors is while judging documents. A negative criterion means that the user is liberal in judging behaviour, that is, he is willing to commit false positive mistakes to avoid missing relevant documents. A positive criterion for user indicates his conservative behaviour, where he tries to keep the false positive rate low at the expense of missing relevant documents.

True positive rate or false positive rate of 1 and 0 in the function z results in infinities. To better understand the rates and avoid infinities, we employ standard smoothing mech-

anism to our calculations of true positive rates and false positive rates. This smoothing is achieved by adding a pseudo-document to the count of documents judged.

Thus the estimated true positive rate (eTPR) is:

$$eTPR = \frac{|TP| + 0.5}{|TP| + |FN| + 1} \tag{3.7}$$

and the estimated false positive (eTPR) rate as:

$$eFPR = \frac{|FP| + 0.5}{|FP| + |TN| + 1} \tag{3.8}$$

3.5 Relevance Judgements

In this user study, we have used NIST judgements (qrels) on 8 topics from AQUAINT collection (2 topics in tutorial and qualification phases and 6 topics in task phase). NIST's relevance judgements are known to have inconsistencies (Harman, 2011, chap. 2). A document judged relevant by one assessor may be judged non-relevant by another assessor and vice versa. So there are errors in the relevance judgements provided by NIST and these judgements do not make a gold standard. These inconsistencies will affect various measures like true positive rates, false positive rates and accuracies calculated for different groups of precisions. But, since the assignment of documents to documents sets used in this study is random, this effect will equalize among three precision groups. Hence, even though the absolute numbers for true positive rates, false positive rates and accuracies for each group of precision may not be the ones reported in this thesis, the comparison of precision groups and their rankings will remain valid.

Chapter 4

Results and Discussion

4.1 Analysis of Rates

To study the behaviour of participants at varying precisions, we calculated true positive rates, false positive rates and accuracies (all of them are "pooled" measures) of the participants in all three phases of the study. Table 4.1 lists these results. Each row lists results for a precision group and the data in that row is calculated using the judgements of 18 participants who belong to that precision group in each phase. Last row displays the averages of the values in all the rows above. All 54 participants worked on document sets of precision 0.5 in tutorial and qualification phases. They were divided among three precision groups in task phase. For better presentation of results, we have placed them in three precision groups in all three phases of the study.

As we can see from the table 4.1, true positive rates, false positive rates and accuracies are almost same for three groups in tutorial phase and qualification phases. This is not surprising as all participants worked on same document sets in tutorial and qualification

Group	Precision = 0.5						Precision = Group		
	Tutorial			Qualification			Task		
	TPR	**FPR**	**Acc**	**TPR**	**FPR**	**Acc**	**TPR**	**FPR**	**Acc**
0.1	0.89	0.07	0.91	0.97	0.02	0.98	0.65	0.06	0.91
0.5	0.90	0.07	0.92	0.96	0.03	0.96	0.69	0.05	0.82
0.9	0.80	0.04	0.88	0.94	0.03	0.95	0.63	0.10	0.66
Avg.	0.86	0.06	0.90	0.96	0.03	0.96	0.66	0.07	0.80

Table 4.1: Comparison of true positive rates, false positive rates and accuracies in different phases of the study

Tutorial			Qualification			Task		
TPR	**FPR**	**Acc**	**TPR**	**FPR**	**Acc**	**TPR**	**FPR**	**Acc**
0.10	0.77	0.41	0.45	0.54	0.28	0.02	0.07	< 0.001

Table 4.2: p-values for different phases of the study.

phases. But when these participants were divided among three different groups based on precision in task phase, they exhibited different behaviour.

From table 4.1, we see that true positive rates and false positive rates are best for precision 0.5 and worst at precision 0.9.

We performed χ^2 test to understand the independence among different precision groups for each phase. p-values for χ^2 test are displayed in table 4.2.

We performed χ^2 tests to find out the independence among true positive rates and found p-value of 0.02. Similar analysis for false positive rates resulted in a p-value of 0.07. We also wanted to compare two precisions at a time, so we performed χ^2 tests between

Groups compared	**p-value (TPR)**	**p-value (FPR)**
All	0.02	0.07
0.1 and 0.5	0.38	0.21
0.5 and 0.9	0.01	0.04
0.1 and 0.9	0.72	0.17

Table 4.3: χ^2 test statistics for different precision levels for task phase.

each pair of precisions assuming that the experiment was conducted only for these two precisions. All these results are summarized in table 4.3.

By looking at the numbers in table 4.3, we see that increasing the precision from 0.5 to 0.9 adversely affects both true positive rate and false positive rate with p-value of 0.01 and 0.04 respectively. Even though it is not certain whether decreasing precision from 0.5 to 0.1 hurts the quality of judgements, both true positive rate and false positive rate were worse at precision 0.1 than at precision 0.5.

We compared true positive rates and false positive rates computed in the task phase of this study with same measures computed in the study by Smucker and Jethani (2010a) which used the same 8 topics used in this study. Table 4.4 shows the true positive rates and false positives rates from phase 1 of that study. We only chose phase 1 because it was the one most similar in nature to this study.

We see that false positive rates in the study by Smucker and Jethani (2010a) were much higher than the ones found in this study. There are four possible reasons for this difference. First of all, the manner in which the documents were selected in both the studies was different. In the current study, documents were selected at random whereas in the study by Smucker and Jethani (2010a), the documents were the top ranked relevant

Precision	True positive rate	False positive rate
0.3	0.78	0.24
0.6	0.73	0.23

Table 4.4: Data from phase 1 of the study conducted by Smucker and Jethani (2010a).

and non-relevant documents of ranking list generated by performing a reciprocal rank fusion (Cormack et al., 2009) on all of the runs submitted to the 2005 TREC Robust track except the 4 lowest performing runs. We think that non-relevant documents of this list may be difficult to judge, that is why very high false positive rate in that study was observed. Second, in this study participants were quizzed on how well they understood the instructions. Only if they answered all the quiz questions correctly, were they permitted to enter the tutorial. In phase 1 the study by Smucker and Jethani (2010a), they were not tested on this and it was assumed that the participants must have read the instructions and understood them. Third, in this study, participants were trained and tested on document judging process in tutorial and qualification phases. They were well aware of how to judge a document and they had displayed this by qualifying for the task phase. In the study by Smucker and Jethani (2010a), there was no such training and testing of the participants who took the tasks. Fourth, in both studies there was a difference in the nature of the tasks with respect to the time and number of documents. In the phase 1 of the study by Smucker and Jethani (2010a), participants were instructed to "try to find as many relevant documents as possible in the 10 minutes while still making as few mistakes in judging the documents' relevance as possible." This could have resulted in speeding up the judging process in order to judge more and more documents. In this study, participants had prior information that they had to judge 40 documents for a topic and that there was no time limit.

To understand the behaviour of each participant, we calculated each participant's true positive rate, false positive rate and accuracy in each phase of the study. This data is displayed in tables 4.5, 4.6 and 4.7.

Since the relevant and non-relevant documents are not balanced in precision 0.1 and precision 0.9 document sets, we cannot use accuracies to judge the performance of the participants. For example, participant id 35 who worked on 0.1 precision ranking list has an accuracy of 0.90, but his true positive rate is only 0.38, which is very low as compared to participant id 15 who worked on 0.5 ranking list and has an accuracy of 0.89 and true positive rate of 0.85. Due to this reason, true positive rates and false positive rates were used to compare the performance of participants in this thesis.

Participants in 0.5 precision group have low false positive rates in general. We think this could be because at precision 0.5, participants always see on an average one out of two documents as relevant or non-relevant; this may have kept them always attentive and resulted in committing very few mistakes.

Participants who worked on 0.9 precision sets seem to reflect both extremes of false positive rates. Some participants have very high false positive rates and some have low to moderate false positive rates. One possible reason of this behaviour could be that some participants in this group may have gotten carried away with the flood of relevant documents and ended up making a lot of mistakes. On the other hand, there were some participants who seemed to become very cautious as and when they saw non-relevant documents and they did not make any mistake in judging them.

The behaviour of 0.1 precision group is not exactly the reverse of the same exhibited by 0.9 participants. Participants at precision 0.1 do not judge a lot of relevant documents as non-relevant. Our explanation of this behaviour is as follows. We believe that to judge

Group	PID	Precision = 0.5						Precision = Group		
		Tutorial			Qualification			Task		
		TPR	FPR	Acc	TPR	FPR	Acc	TPR	FPR	Acc
0.1	6	0.80	0.20	0.80	1.00	0.10	0.95	0.62	0.06	0.91
0.1	11	0.80	0.20	0.80	0.70	0.00	0.85	0.00	0.12	0.79
0.1	3	1.00	0.00	1.00	1.00	0.10	0.95	1.00	0.00	1.00
0.1	12	1.00	0.00	1.00	1.00	0.10	0.95	0.75	0.03	0.95
0.1	14	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.07	0.94
0.1	25	0.60	0.00	0.80	0.90	0.00	0.95	0.25	0.06	0.88
0.1	18	0.80	0.20	0.80	1.00	0.00	1.00	0.75	0.15	0.84
0.1	31	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.11	0.90
0.1	32	0.80	0.00	0.90	1.00	0.00	1.00	0.75	0.10	0.89
0.1	35	0.80	0.00	0.90	1.00	0.00	1.00	0.38	0.04	0.90
0.1	36	0.80	0.20	0.80	1.00	0.00	1.00	0.88	0.15	0.85
0.1	30	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.06	0.95
0.1	39	1.00	0.00	1.00	1.00	0.00	1.00	0.00	0.00	0.90
0.1	40	0.60	0.20	0.70	0.90	0.00	0.95	0.62	0.07	0.90
0.1	43	1.00	0.00	1.00	1.00	0.00	1.00	0.38	0.03	0.91
0.1	42	1.00	0.00	1.00	0.90	0.00	0.95	1.00	0.04	0.96
0.1	49	1.00	0.00	1.00	1.00	0.00	1.00	0.50	0.03	0.93
0.1	50	1.00	0.20	0.90	1.00	0.00	1.00	0.88	0.03	0.96

Table 4.5: True positive rates, false positive rates and accuracies for 0.1 precision participants.

Group	PID	Precision = 0.5						Precision = Group		
		Tutorial			Qualification			Task		
		TPR	FPR	Acc	TPR	FPR	Acc	TPR	FPR	Acc
0.5	5	1.00	0.00	1.00	1.00	0.10	0.95	0.93	0.00	0.96
0.5	9	0.80	0.00	0.90	0.80	0.00	0.90	0.95	0.12	0.91
0.5	4	0.80	0.20	0.80	0.90	0.20	0.85	0.82	0.12	0.85
0.5	15	1.00	0.20	0.90	0.80	0.10	0.85	0.85	0.07	0.89
0.5	16	1.00	0.20	0.90	1.00	0.00	1.00	0.68	0.03	0.82
0.5	17	1.00	0.00	1.00	1.00	0.00	1.00	0.47	0.07	0.70
0.5	22	0.80	0.00	0.90	0.90	0.00	0.95	0.28	0.00	0.64
0.5	24	0.80	0.00	0.90	1.00	0.00	1.00	0.80	0.05	0.88
0.5	21	1.00	0.00	1.00	0.90	0.00	0.95	0.55	0.00	0.78
0.5	29	0.60	0.20	0.70	1.00	0.20	0.90	0.85	0.20	0.82
0.5	28	0.80	0.00	0.90	1.00	0.00	1.00	0.62	0.00	0.81
0.5	37	1.00	0.20	0.90	1.00	0.00	1.00	0.82	0.03	0.90
0.5	33	1.00	0.00	1.00	1.00	0.00	1.00	0.80	0.03	0.89
0.5	41	0.80	0.20	0.80	0.90	0.00	0.95	0.70	0.03	0.84
0.5	44	1.00	0.00	1.00	1.00	0.00	1.00	0.33	0.00	0.66
0.5	46	1.00	0.00	1.00	1.00	0.00	1.00	0.65	0.03	0.81
0.5	48	1.00	0.00	1.00	1.00	0.00	1.00	0.93	0.05	0.94
0.5	54	0.80	0.00	0.90	1.00	0.00	1.00	0.47	0.05	0.71

Table 4.6: True positive rates, false positive rates and accuracies for 0.5 precision participants.

Group	PID	Precision = 0.5						Precision = Group		
		Tutorial			Qualification			Task		
		TPR	FPR	Acc	TPR	FPR	Acc	TPR	FPR	Acc
0.9	1	0.80	0.00	0.90	1.00	0.00	1.00	0.75	0.25	0.75
0.9	2	0.60	0.00	0.80	0.90	0.10	0.90	0.61	0.00	0.65
0.9	8	0.80	0.00	0.90	1.00	0.00	1.00	0.78	0.00	0.80
0.9	7	0.60	0.00	0.80	0.80	0.00	0.90	0.40	0.12	0.45
0.9	10	1.00	0.00	1.00	0.90	0.00	0.95	0.86	0.00	0.88
0.9	19	0.80	0.00	0.90	0.90	0.00	0.95	0.72	0.00	0.75
0.9	13	1.00	0.00	1.00	1.00	0.00	1.00	0.35	0.00	0.41
0.9	27	1.00	0.00	1.00	1.00	0.10	0.95	0.47	0.25	0.50
0.9	26	1.00	0.00	1.00	1.00	0.20	0.90	0.81	0.12	0.81
0.9	20	0.60	0.00	0.80	0.90	0.00	0.95	0.62	0.12	0.65
0.9	23	0.80	0.20	0.80	1.00	0.00	1.00	0.86	0.12	0.86
0.9	34	0.80	0.60	0.60	0.90	0.20	0.85	0.79	0.75	0.74
0.9	38	0.80	0.00	0.90	0.90	0.00	0.95	0.79	0.00	0.81
0.9	45	0.80	0.00	0.90	0.90	0.00	0.95	0.40	0.00	0.46
0.9	47	0.60	0.00	0.80	1.00	0.00	1.00	0.85	0.00	0.86
0.9	51	0.80	0.00	0.90	1.00	0.00	1.00	0.39	0.00	0.45
0.9	53	0.80	0.00	0.90	0.90	0.00	0.95	0.53	0.00	0.57
0.9	52	0.80	0.00	0.90	0.90	0.00	0.95	0.42	0.00	0.47

Table 4.7: True positive rates, false positive rates and accuracies for 0.9 precision participants.

a document as non-relevant one has to read the whole document to make sure that there is nothing in the document which matches to the relevance criterion. Participants who worked on 0.1 precision sets see a lot of non-relevant documents and they get used to judging non-relevant documents, so they do not make a lot of false positive mistakes. They also get used to reading the documents carefully, so they do not miss out on relevant documents, as judging a document as relevant requires them to just find a few sentences which are relevant to the topic.

4.2 Analysis of Time

As part of the data collected in this study, we recorded time taken by participants on every document for all phases. We calculated the average amount of time taken to judge a relevant document, non-relevant document and also the average time to judge a document in each phase for each of the three groups of participants. All this data was computed considering NIST's judgements as the source the truth. We also considered participant's judgement as the source of truth and re-computed the time data. Table 4.8 displays both data sets.

By looking at the data in tables 4.8, we see, in both the cases participants took the longest to judge a document in tutorial phase, then in qualification phase and the least in the task phase. We can justify this by saying that they were learning to judge the documents in tutorial phase. In qualification phase, even though they had learnt to judge the documents, they wanted to qualify for the full study which made them work attentively but once they cleared the qualification phase, they started working on their own pace.

We see that when we consider NIST's judgement as the source of truth, participants spend more time on relevant documents. But when we consider participant's judgement

Average time spent on a document considering NIST as the source of truth.

Group	Precision = 0.5						Precision = Group		
	Tutorial			Qualification			Task		
	Rel	Non-rel	All	Rel	Non-rel	All	Rel	Non-rel	All
0.1	50	39	45	28	33	30	33	24	25
0.5	53	36	44	33	31	32	30	23	27
0.9	52	41	46	35	39	37	26	25	26
Avg.	52	39	45	32	34	33	30	24	26

Average time spent on a document considering user's judgement as the source of truth.

Group	Precision = 0.5						Precision = Group		
	Tutorial			Qualification			Task		
	Rel	Non-rel	All	Rel	Non-rel	All	Rel	Non-rel	All
0.1	36	52	45	26	34	30	28	24	25
0.5	43	46	44	32	32	32	29	25	27
0.9	36	54	46	34	39	37	23	29	26
Avg.	38	51	45	31	35	33	27	26	26

Table 4.8: Comparison of average time spent (in seconds) per document in different phases of the study. First table displays numbers considering NIST as the source of truth and second table shows the values if user's judgement is considered as the source of truth.

as the source of truth this behaviour is reversed for participants who judged 0.9 precision sets: they spend more time on non-relevant documents.

We performed two-sided paired t-tests with 95% confidence interval between average time spent on relevant and non-relevant documents for both cases (NIST as source of truth and participant's judgement as the source of truth) for different phases of the study. The results of the t-test are displayed in table 4.9.

Table 4.10 shows the confusion matrix for the average time spent on a document in task phase considering NIST as the source of truth.

By looking at the numbers in table 4.10, we observe that participants spend more time when they make mistakes while judging documents for 0.1 or 0.5 precision sets. While judging documents for 0.9 precision sets, participants take the most amount of time while committing the mistake of judging NIST relevant document as non-relevant. When making mistake of judging NIST non-relevant as relevant, they take the least amount of time. We believe that this is due to the flow of judging a lot of relevant documents that they end up judging a non-relevant document as relevant so quickly.

To observe the behaviour of each participant with respect to time spent on documents, we calculated average time spent by each participant on relevant documents, non-relevant documents and average time spent on a document in each phase of the study. This data is displayed in tables 4.11, 4.12 and 4.13.

t-test statistics considering NIST as the source of truth.

Group	**Precision = 0.5**		**Precision = Group**
	p-value (Tutorial)	**p-value (Qualification)**	**p-value (Task)**
0.1	0.15	0.08	0.13
0.5	0.02	0.58	0.01
0.9	0.01	0.15	0.67
All	< 0.001	0.10	0.01

t-test statistics considering user's judgement as the source of truth.

Group	**Precision = 0.5**		**Precision = Group**
	p-value (Tutorial)	**p-value (Qualification)**	**p-value (Task)**
0.1	0.23	0.01	0.20
0.5	0.93	0.78	0.07
0.9	0.03	0.02	0.01
All	0.06	0.01	0.40

Table 4.9: t-test statistics of average time spent (in seconds) by participants on relevant and non-relevant documents in different phases of the study. First table displays the numbers considering NIST as the source of truth and second table shows the results considering user's judgement as the source of truth.

Group	Participant Judgement	NIST Judgement	
		Relevant	**Non-Relevant**
0.1	Relevant	24	31
	Non-Relevant	50	23
0.5	Relevant	29	29
	Non-Relevant	34	22
0.9	Relevant	24	13
	Non-Relevant	30	26
Avg.	Relevant	26	25
	Non-Relevant	33	24

Table 4.10: Average time spent (in seconds) in task phase considering NIST as the source of truth.

Group	PID	Precision = 0.5						Precision = Group		
		Tutorial			Qualification			Task		
		Rel	Non-rel	All	Rel	Non-rel	All	Rel	Non-rel	All
0.1	6	11	41	26	17	43	30	18	33	31
0.1	11	76	62	69	45	31	38	21	14	15
0.1	3	20	25	23	28	32	30	41	16	19
0.1	12	7	9	8	6	14	10	58	78	76
0.1	14	34	23	29	29	13	21	31	11	13
0.1	25	196	113	154	62	67	64	70	23	28
0.1	18	59	64	62	37	56	47	11	18	18
0.1	31	22	41	32	33	50	42	30	34	34
0.1	32	51	30	40	33	27	30	24	30	29
0.1	35	67	46	56	28	49	39	23	22	22
0.1	36	65	50	57	37	36	37	15	14	14
0.1	30	6	14	10	11	13	12	9	13	13
0.1	39	140	62	101	41	58	50	103	22	30
0.1	40	48	31	39	32	24	28	25	18	19
0.1	43	43	35	39	21	27	24	48	19	22
0.1	42	16	31	23	14	23	18	27	30	30
0.1	49	11	11	11	8	10	9	23	15	16
0.1	50	27	23	25	14	16	15	18	19	19

Table 4.11: Comparison of average time spent (in seconds) per document by participants who judged 0.1 precision sets.

Group	PID	Precision = 0.5						Precision = Group		
		Tutorial			Qualification			Task		
		Rel	Non-rel	All	Rel	Non-rel	All	Rel	Non-rel	All
0.5	5	57	29	43	28	35	32	21	20	21
0.5	9	70	52	61	31	25	28	27	27	27
0.5	4	28	20	24	34	24	29	33	28	31
0.5	15	18	18	18	32	27	30	21	11	16
0.5	16	33	22	28	35	26	31	41	40	41
0.5	17	37	30	33	51	39	45	43	31	37
0.5	22	69	36	52	55	41	48	21	16	18
0.5	24	99	69	84	42	66	54	34	37	36
0.5	21	13	20	17	23	19	21	31	10	21
0.5	29	18	8	13	12	15	14	8	11	10
0.5	28	22	16	19	10	11	11	19	7	13
0.5	37	55	58	57	56	61	58	48	22	35
0.5	33	24	15	19	23	18	21	23	15	19
0.5	41	7	11	9	14	11	13	17	9	13
0.5	44	64	41	52	26	40	33	40	18	29
0.5	46	204	89	147	58	38	48	59	40	49
0.5	48	30	28	29	21	17	19	9	24	17
0.5	54	101	92	97	39	51	45	51	42	47

Table 4.12: Comparison of average time spent (in seconds) per document by participants who judged 0.5 precision sets.

Group	PID	Precision = 0.5						Precision = Group		
		Tutorial			Qualification			Task		
		Rel	Non-rel	All	Rel	Non-rel	All	Rel	Non-rel	All
0.9	1	40	18	29	20	12	16	17	28	18
0.9	2	48	26	37	31	19	25	16	20	16
0.9	8	48	43	45	32	42	37	29	19	28
0.9	7	47	37	42	10	17	14	18	12	18
0.9	10	46	44	45	44	39	42	23	37	25
0.9	19	55	55	55	63	65	64	45	53	45
0.9	13	11	22	16	9	14	11	26	27	26
0.9	27	110	92	101	67	90	78	19	7	18
0.9	26	55	49	52	56	72	64	32	36	32
0.9	20	60	45	52	41	35	38	27	31	28
0.9	23	72	53	62	58	61	59	32	27	31
0.9	34	14	7	10	5	6	6	3	3	3
0.9	38	7	27	17	13	17	15	20	25	21
0.9	45	43	30	36	40	40	40	31	18	30
0.9	47	140	90	115	40	75	57	35	52	36
0.9	51	54	32	43	25	27	26	37	16	35
0.9	53	34	30	32	47	42	44	31	25	31
0.9	52	46	39	43	20	21	21	25	11	23

Table 4.13: Comparison of average time spent (in seconds) per document by participants who judged 0.9 precision sets.

4.3 Analysis of Assessor Criterion and Ability to Discriminate

We calculated ability to discriminate (d') and criterion (c) values for participants of three groups for all phases. These values are displayed in table 4.14.

We see that in the task phase, participants in 0.9 precision group are the least conservative in their judging behaviour and the worst in their ability to discriminate between relevant and non-relevant documents. Participants in 0.5 precision group are the best in their ability to discriminate between relevant and non-relevant documents and are the most conservative in their judging behaviour.

Group	**Precision = 0.5**				**Precision = Group**	
	Tutorial		**Qualification**		**Task**	
	d'	**c**	d'	**c**	d'	**criterion**
0.1	2.72	0.14	3.96	0.15	1.92	0.57
0.5	2.78	0.11	3.54	0.07	2.17	0.58
0.9	2.54	0.43	3.38	0.14	1.64	0.48
Avg.	2.65	0.23	3.60	0.12	1.89	0.53

Table 4.14: Ability to Discriminate and Assessor Criterion.

4.4 Cheaters

In this study, we trained participants on how to judge a document for a given topic. After they were trained we tested them on the same and made sure that only qualified

participants took part in the task phase. Since the participants had an incentive of earning more remuneration by qualifying for the task phase, we thought it might have motivated them to do well in the qualification task. But in the task phase, there was no motivation to do well, rather, they just had to complete the study and earn remuneration. They were neither being tracked for the quality of judgements they produced, nor, were they being timed for the amount of time they spent on the task phase. We wanted to see if this situation lured the participants to just finish the tasks as quickly as possible and not care about the quality of work they produced. To discern such behaviour, we sorted the participants based on the average time spent a document in the task phase. Table 4.15 lists top 20 participants with fastest document judging time in task phase and various measures associated with them.

The fastest participant, participant id 34, does not have a good true positive rate. The false positive rate for this participant is very high, but since he saw only 8 documents which were non-relevant (0.9 precision set), we cannot say that this participant cheated. Moreover, despite his liberality in judging documents, his ability to discriminate is above 0.

Group	PID	Time	TPR	FPR	eTPR	eFPR	d'	Criterion
0.9	34	3	0.79	0.75	0.79	0.72	0.22	-0.69
0.5	29	10	0.85	0.2	0.84	0.21	1.8	-0.09
0.1	14	13	1	0.07	0.94	0.08	2.96	-0.07
0.5	28	13	0.62	0	0.62	0.01	2.63	1.01
0.1	30	13	1	0.06	0.94	0.06	3.11	0
0.5	41	13	0.7	0.03	0.7	0.04	2.28	0.61
0.1	36	14	0.88	0.15	0.83	0.16	1.95	0.02
0.1	11	15	0	0.12	0.06	0.13	-0.43	1.34
0.9	2	16	0.61	0	0.61	0.06	1.83	0.64
0.5	15	16	0.85	0.07	0.84	0.09	2.34	0.17
0.1	49	16	0.5	0.03	0.5	0.03	1.88	0.94
0.5	48	17	0.93	0.05	0.91	0.06	2.9	0.11
0.9	1	18	0.75	0.25	0.75	0.28	1.26	-0.05
0.9	7	18	0.4	0.12	0.4	0.17	0.7	0.6
0.1	18	18	0.75	0.15	0.72	0.16	1.58	0.21
0.5	22	18	0.28	0	0.28	0.01	1.74	1.45
0.9	27	18	0.47	0.25	0.47	0.28	0.51	0.33
0.1	3	19	1	0	0.94	0.01	3.88	0.39
0.5	33	19	0.8	0.03	0.79	0.04	2.56	0.47
0.1	40	19	0.62	0.07	0.61	0.08	1.68	0.56

Table 4.15: Top 20 participants with fastest document judging time (in seconds) during task phase.

Another way we tried to find out the participants who were cheating was to see the jump in their rank based on average time spent in qualification phase and average time spent in task phase. This data is displayed in 4.16.

The participant with the longest jump in the rank had a reasonable true positive rate, false positive rate, d' and criterion.

Graphically, we can see in figure 4.1 that participants who worked faster in qualification phase seemed to be working faster in task phase as well.

In summary, we can say that even though participants had the option of rushing through the task phase and not care about the quality of work they produced, they did not take advantage of this situation and worked honestly.

4.5 Time to judge as the phase proceeds

Figure 4.2 and figure 4.3 show the plots between document rank and the average time spent on the document by participants in each group and document rank and average time spent on the document across all groups respectively for tutorial phase. The vertical dotted line in the plots separates two topics judged.

We see that, in general, the amount of time spent on the document decreases as the participant proceeds for both the topics.

Similar plots for qualification phase are displayed in plots 4.4 and 4.5. Similar to the plots for tutorial phase, the vertical dotted line in the plots separates two topics judged in the qualification phase.

The plots do not clearly indicate reduction in the amount of time as the qualification phase proceeds. An explanation of this can be that, participants wanted to clear the

Group	**PID**	**Qual time**	**Task time**	**Rank jump**	**TPR**	**FPR**	**eTPR**	**eFPR**	d'	**c**
0.9	27	78	18	37	0.47	0.25	0.47	0.28	0.51	0.33
0.1	18	47	18	28	0.75	0.15	0.72	0.16	1.58	0.21
0.5	22	48	18	28	0.28	0	0.28	0.01	1.74	1.45
0.1	11	38	15	26	0	0.12	0.06	0.13	-0.43	1.34
0.1	36	37	14	26	0.88	0.15	0.83	0.16	1.95	0.02
0.1	25	64	28	19	0.25	0.06	0.28	0.06	0.97	1.07
0.5	15	30	16	17	0.85	0.07	0.84	0.09	2.34	0.17
0.1	14	21	13	12	1	0.07	0.94	0.08	2.96	-0.07
0.9	2	25	16	11	0.61	0	0.61	0.06	1.83	0.64
0.1	35	39	22	11	0.38	0.04	0.39	0.05	1.37	0.96
0.9	10	42	25	10	0.86	0	0.86	0.06	2.64	0.24
0.9	26	64	32	10	0.81	0.12	0.8	0.17	1.8	0.06
0.1	39	50	30	10	0	0	0.06	0.01	0.77	1.94
0.9	23	59	31	9	0.86	0.12	0.86	0.17	2.03	-0.06
0.5	5	32	21	8	0.93	0	0.91	0.01	3.67	0.49
0.1	3	30	19	7	1	0	0.94	0.01	3.88	0.39
0.5	29	14	10	7	0.85	0.2	0.84	0.21	1.8	-0.09
0.5	37	58	35	4	0.82	0.03	0.82	0.04	2.67	0.42
0.9	20	38	28	3	0.62	0.12	0.62	0.17	1.26	0.32
0.1	40	28	19	3	0.62	0.07	0.61	0.08	1.68	0.56

Table 4.16: Top 20 participants with biggest rank jump in judging time (in seconds) spent in qualification phase and in task phase. In this table "c" stands for "criterion".

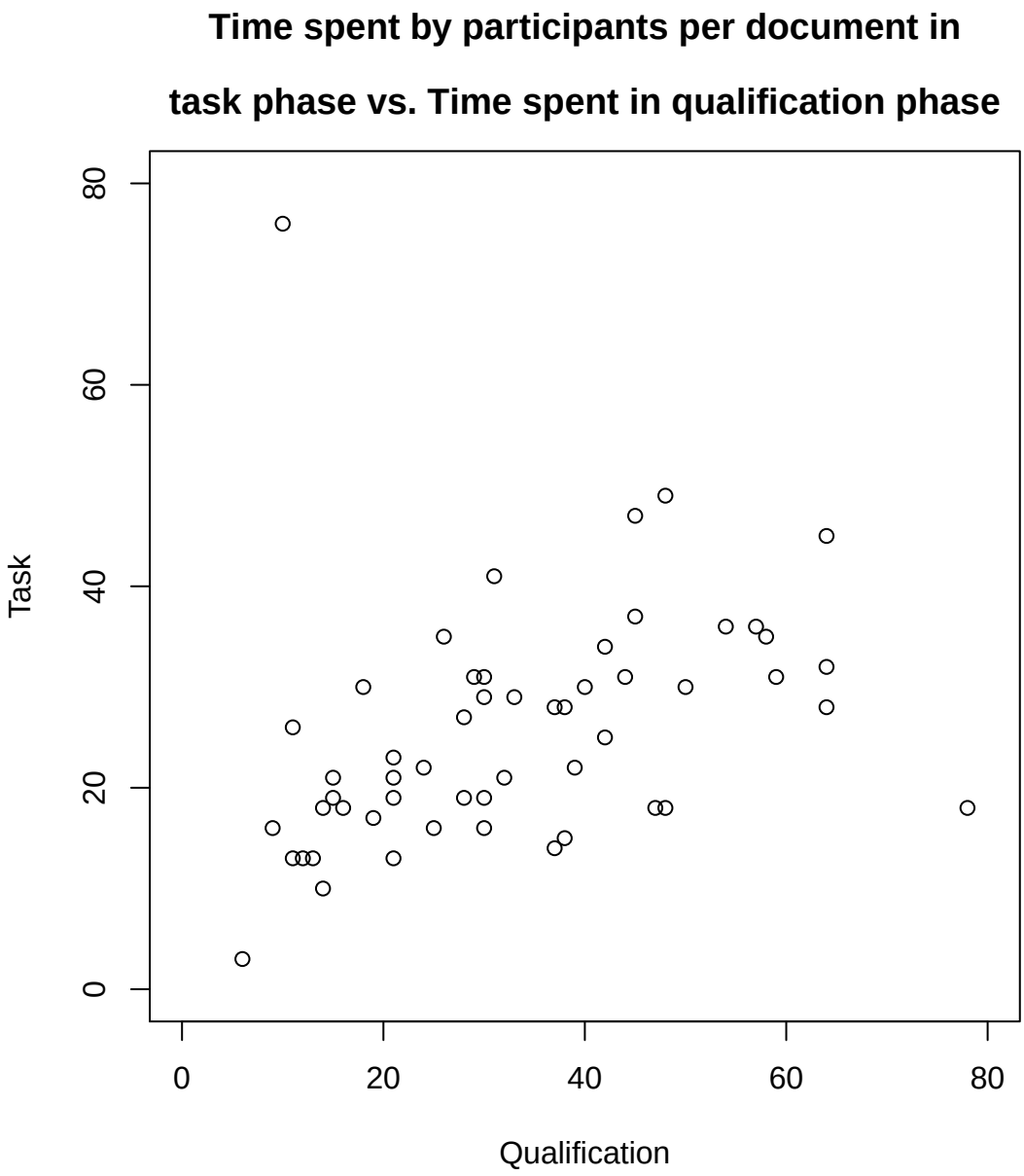


Figure 4.1: Average time spent by participants on a document in task phase versus the same in qualification phase.

qualification phase to qualify for the full study and hence, they took as much time as required to correctly judge the documents. In addition to this, the variance among the time spent by the participants on a document is more for first topic (*mental illness*) than for the second topic (*railway accidents*). As mentioned in the section 3.1.2, for "*mental illness*", even though it is easy to distinguish between a relevant and a non-relevant document, but one may have to read through the document carefully to judge it. Different participants have different abilities to read and comprehend documents and this could be the cause of variance in the time spent by participants on the documents for this topic. This is not the case for "*railway accidents*", as for this topic the information about relevance or non-relevance is on the surface.

Similar plots for task phase displayed in figures 4.6 and 4.7 show that the amount of time spent on a document decreases as the task phase proceeds.

The figure 4.8 shows the plot between doc rank and time spent on each document data for all the phases.

In general, we can say that the time spent on a document decreases as a task proceeds. Participants spend some amount of time learning about the task, but once they learn judging the documents for a topic, the amount of time spent on a document decreases.

4.6 Effect of Time Taken to Judge Documents on True Positive Rate and False Positive Rate

Figure 4.9 shows a plot between the average time spent on a document by participants in task phase and the same in qualification phase. In general it appears that participants faster during the qualification phase work faster in task phase.

Figures 4.10 and 4.11 display the plots between the true positive rate and average time spent on a document and the false positive rate and average time spent on a document by participants respectively in task phase.

We see that there is no correlation between the time spent on a document and the true positive rate and little to no correlation between the time spent on a document and false positive rate. Participants seem to be working at their own pace, i.e. they work at a pace they think is sufficient to judge the documents.

This observation is cemented by the similar plots shown in figures 4.12 and 4.13 for qualification phase.

In qualification phase, participants had an incentive to clear the phase to qualify for the task phase. All of these participants produced almost the same true positive rate and false positive rate but took different amounts of time. This means that participants work at the pace they think is sufficient to produce an expected quality of judgements.

We calculated the values of d' and criterion for each of the participants in our study and plotted them against the average time taken to judge a document. The plots are shown in figures 4.14 and 4.15.

We see that there is no correlation between the ability to discriminate or criterion and the time on a document.

In conclusion, we can say that there is little to no correlation between the amount of time taken to judge a document and the quality of judgements produced. Participants work at a pace they think is sufficient to judge a document.

4.7 Analysis of Pre and Post Task Questionnaires

Section B.13 in the appendices chapter displays pre-task and post-task questionnaires presented to participants before they started and after they finished a task respectively. We used 5 point Likert scale for each question asked to capture participants' response. The preliminary analysis of this data is presented in table 4.17. In our analysis, we have mapped this scale to the values 1 through 5 with the most negative response mapped to 1, e.g. "Very Difficult" and the most positive response mapped to 5, e.g. "Very Easy", and the neutral response to 3. Non-responses to any of the questions have been excluded while doing this analysis.

From the data in table 4.17 we observe that participants find it easiest to find relevant documents for precision 0.9. Participants' experience was most enjoyable and they felt most engaged for precision 0.5. It was easiest for participants to concentrate when they worked on precision 0.1.

In our future work, we will do detailed analysis of this data similar to what Smucker and Jethani (2010b) had done for their user study.

	Question	Precision Group			
		0.1	**0.5**	**0.9**	**All**
Pre-Task	Knowledge about Topic (Nothing Known - Details Known)	2.22	2.28	2.31	2.27
	Finding Relevant Docs (Very Difficult - Very Easy)	3.0	3.08	3.19	3.09
	Relevancy to Life (Not at all - Very Much)	2.5	2.14	2.11	2.25
	Interested to Learn (Not at all - Very Much)	3.25	3.44	3.5	3.40
Post-Task	Finding Relevant Docs (Very Difficult - Very Easy)	2.72	3.03	3.14	2.96
	Experience (Very Unenjoyable - Very Enjoyable)	2.72	3.08	2.94	2.92
	Mood (Very Bored - Very Engaged)	3.08	3.17	2.97	3.07
	Ability to Concentrate (Very Hard - Very Easy)	3.67	3.47	3.6	3.6

Table 4.17: Analysis of Pre and Post Task Questionnaires

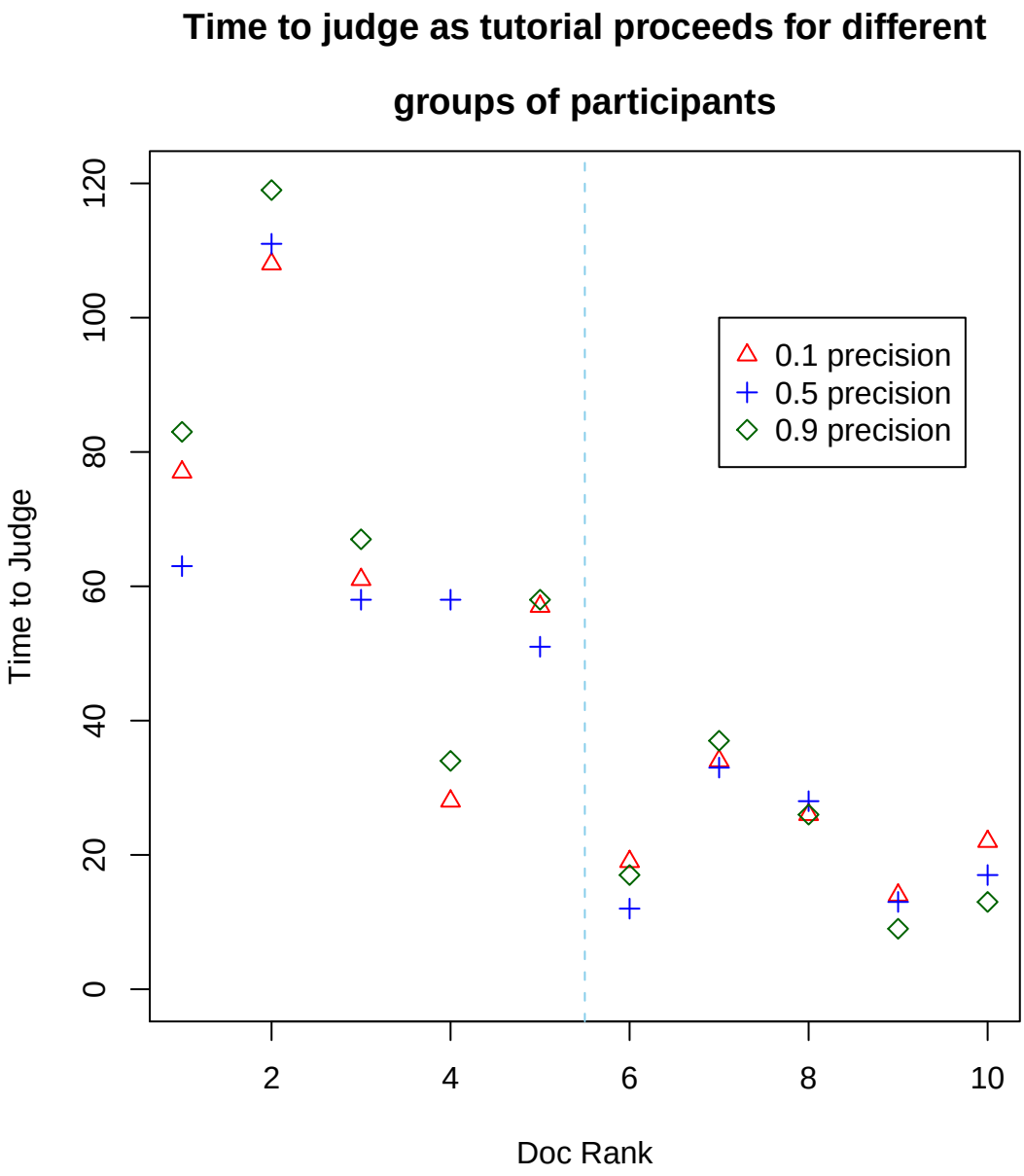


Figure 4.2: Time to judge as tutorial phase proceeds for different groups of participants. The vertical dotted line separates two topics judged in the tutorial phase.

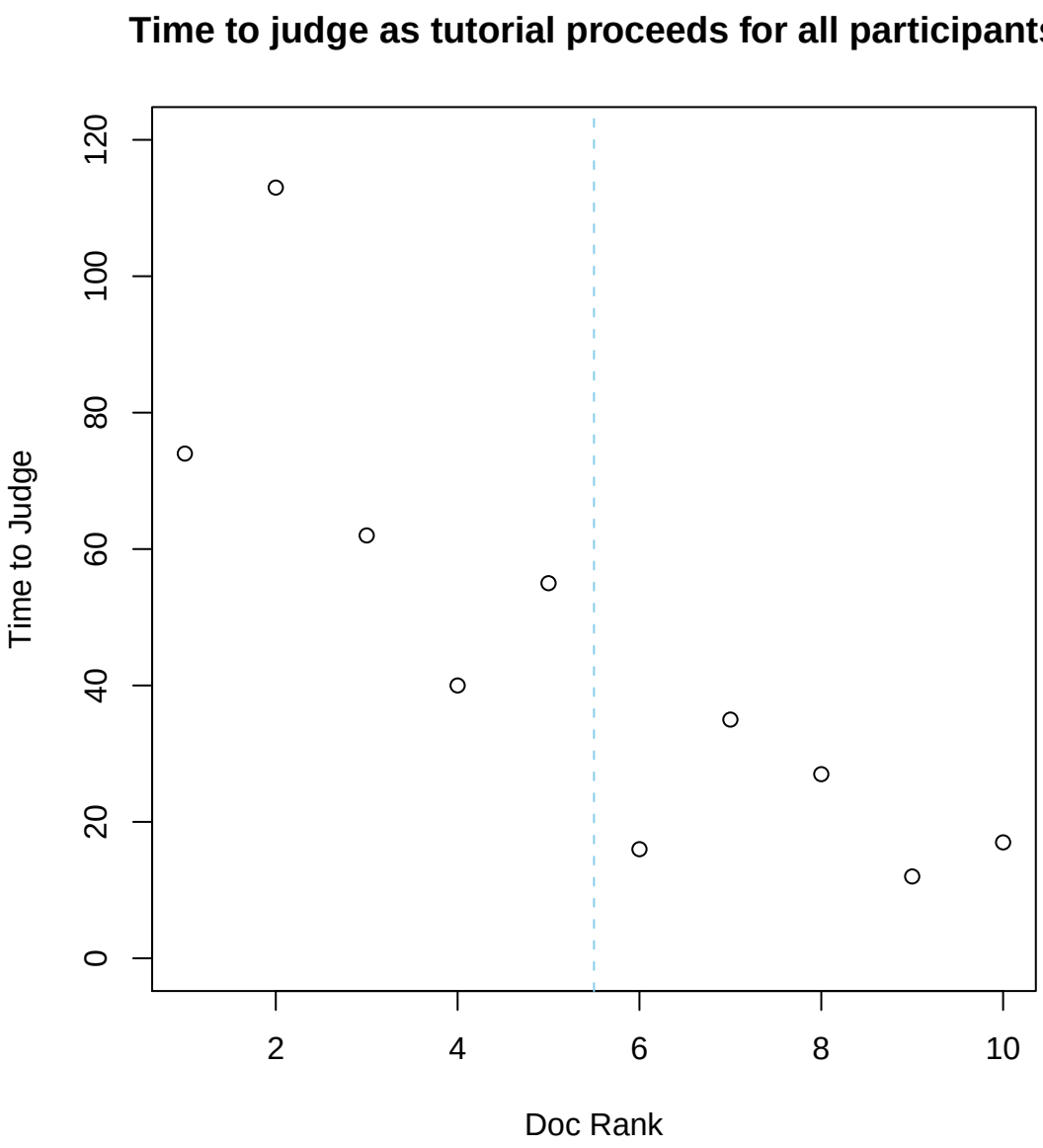


Figure 4.3: Time to judge as tutorial phase proceeds for all participants. The vertical dotted line separates two topics judged in the tutorial phase.

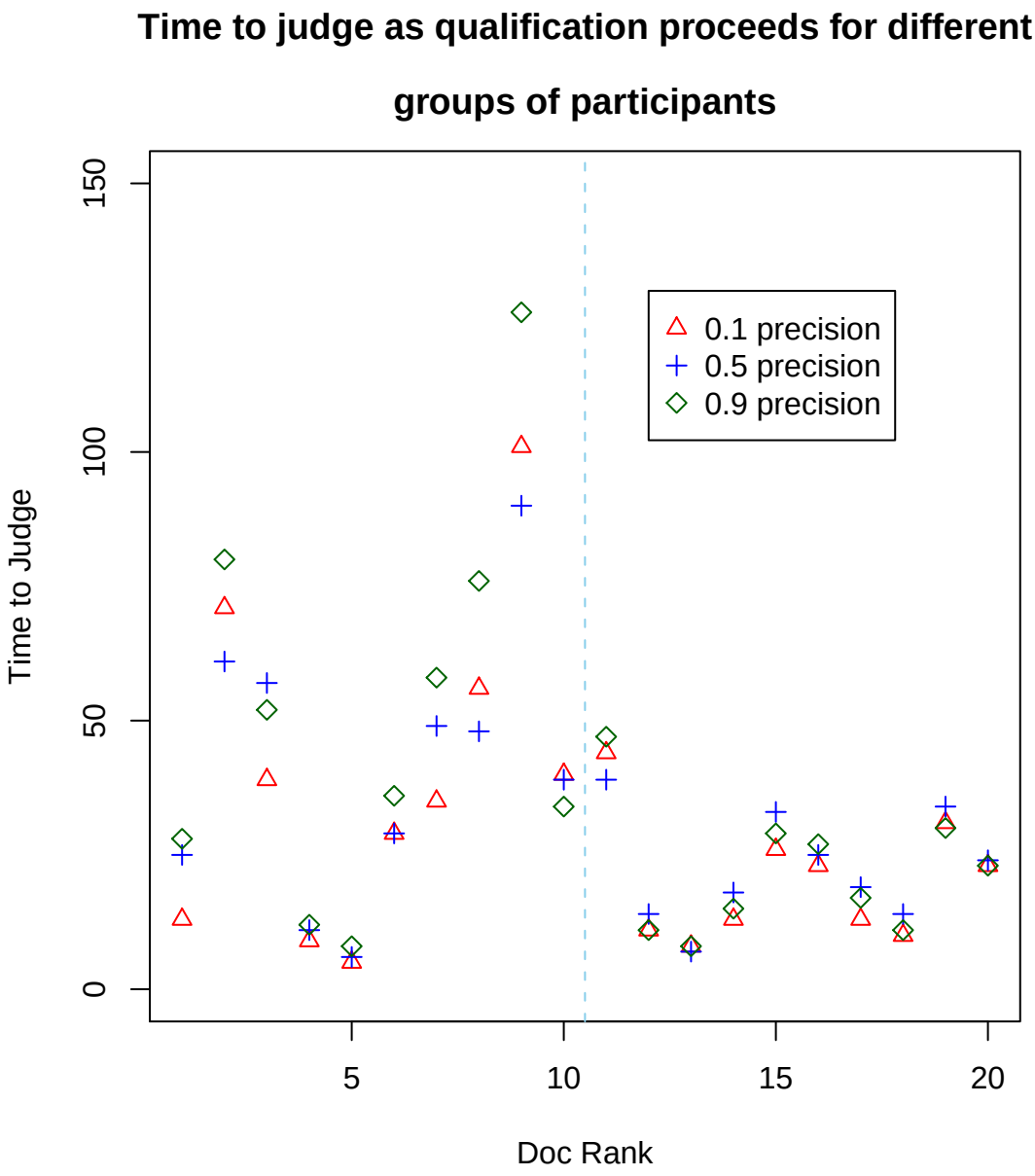


Figure 4.4: Time to judge as qualification phase proceeds for different groups of participants. The vertical dotted line separates two topics judged in the qualification phase.

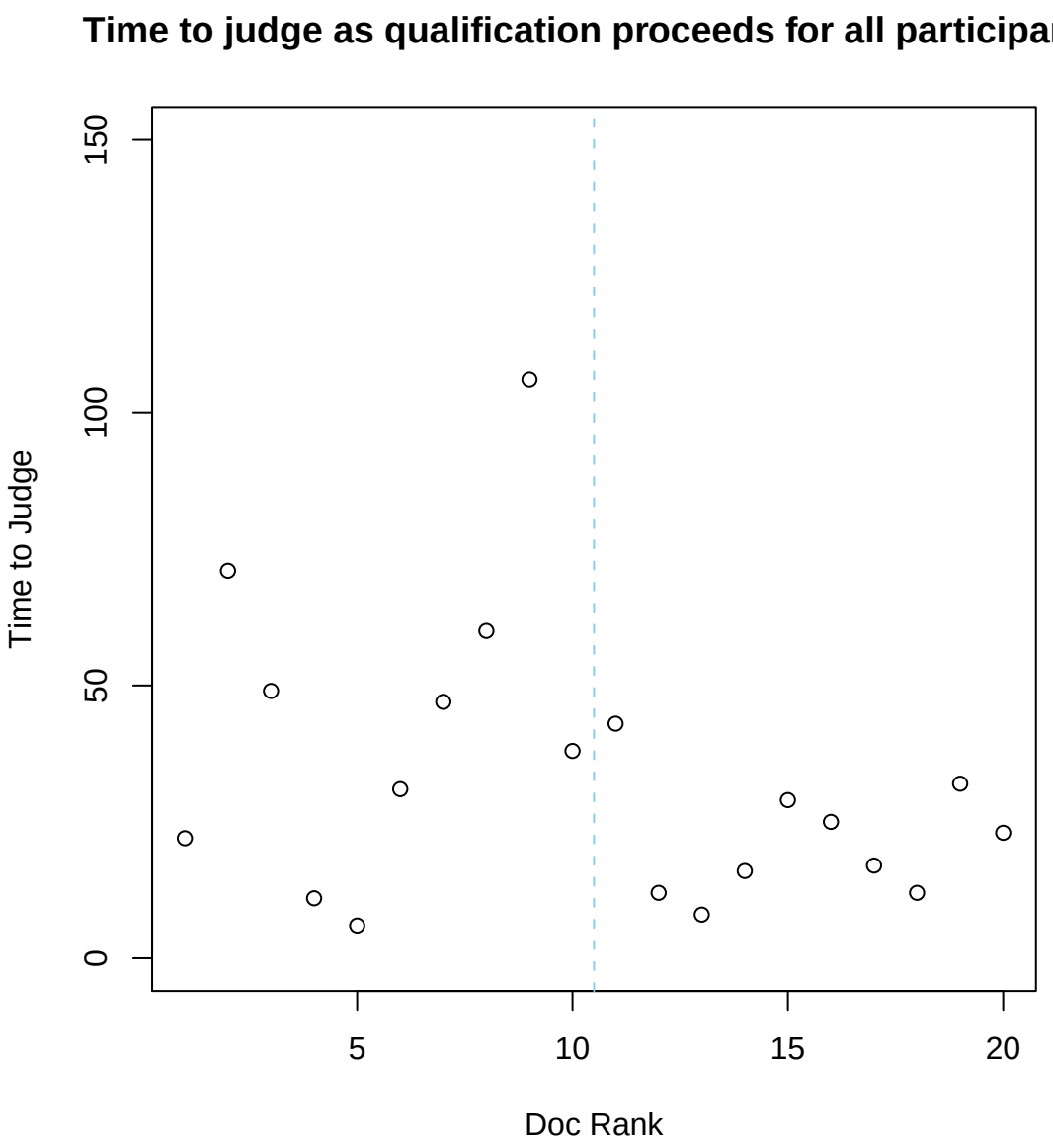


Figure 4.5: Time to judge as qualification phase proceeds for all participants. The vertical dotted line separates two topics judged in the qualification phase.

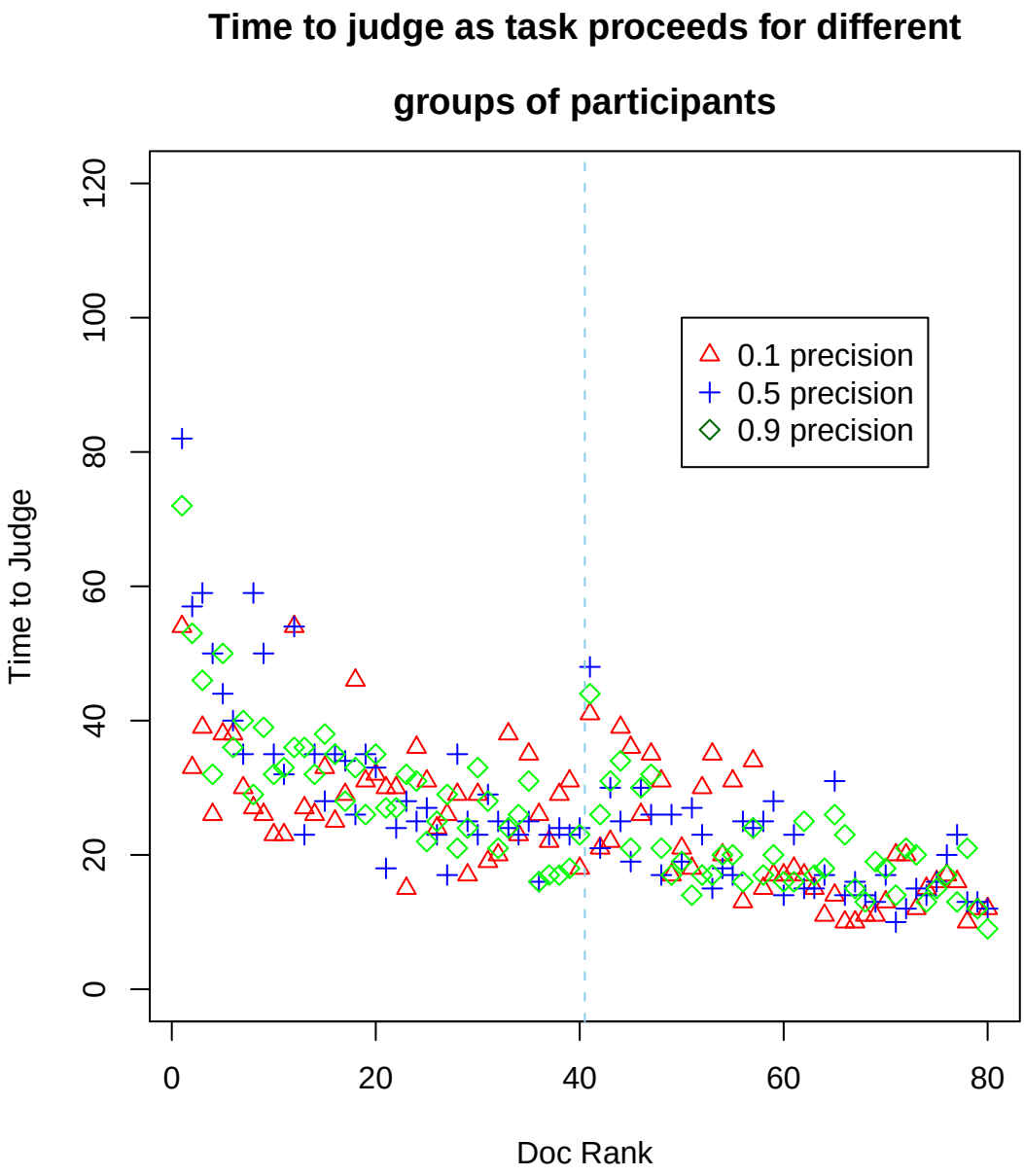


Figure 4.6: Time to judge as task phase proceeds for different groups of participants. The vertical dotted line separates two topics judged in the task phase.

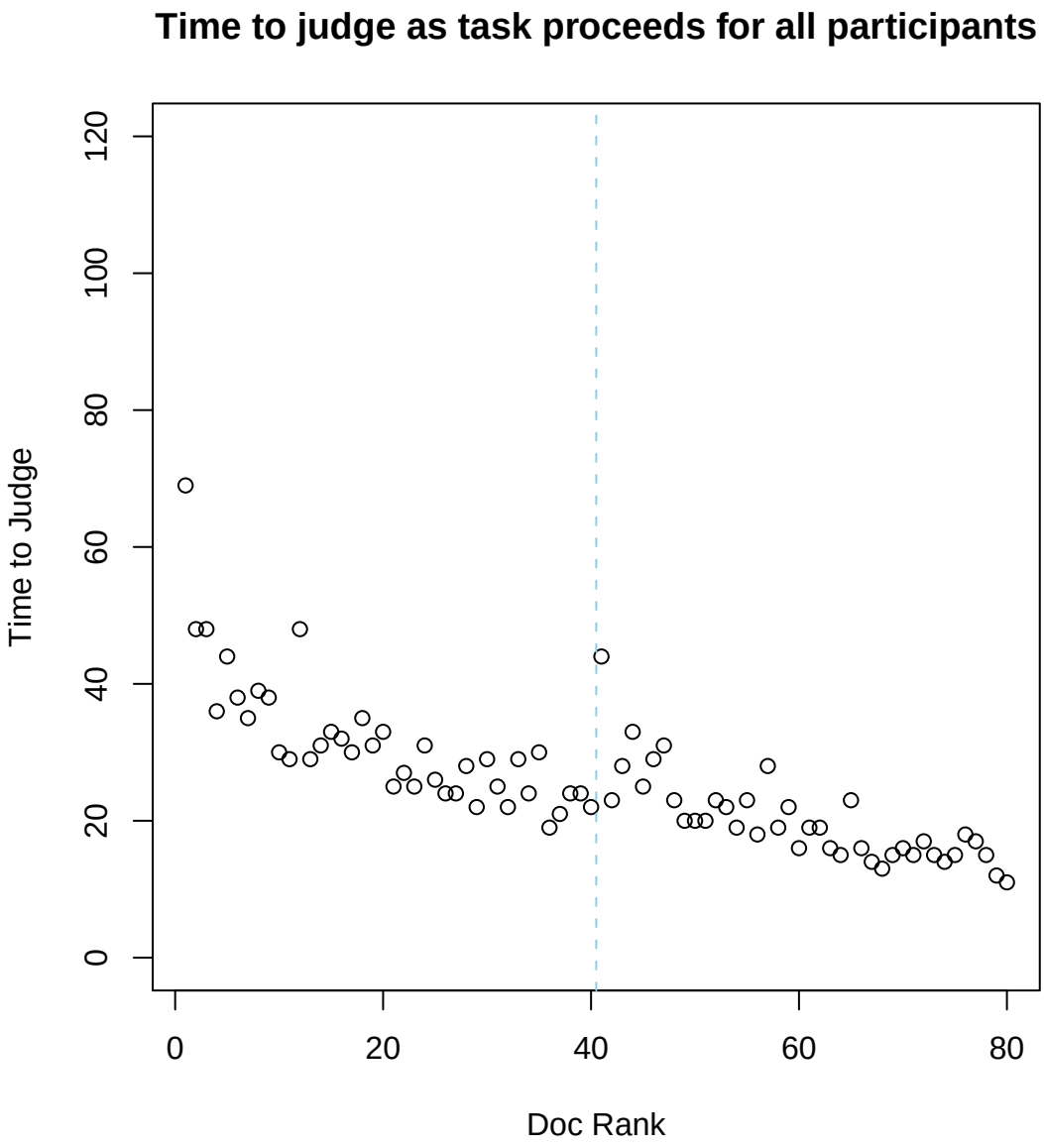


Figure 4.7: Time to judge as task phase proceeds for all participants. The vertical dotted line separates two topics judged in the task phase.

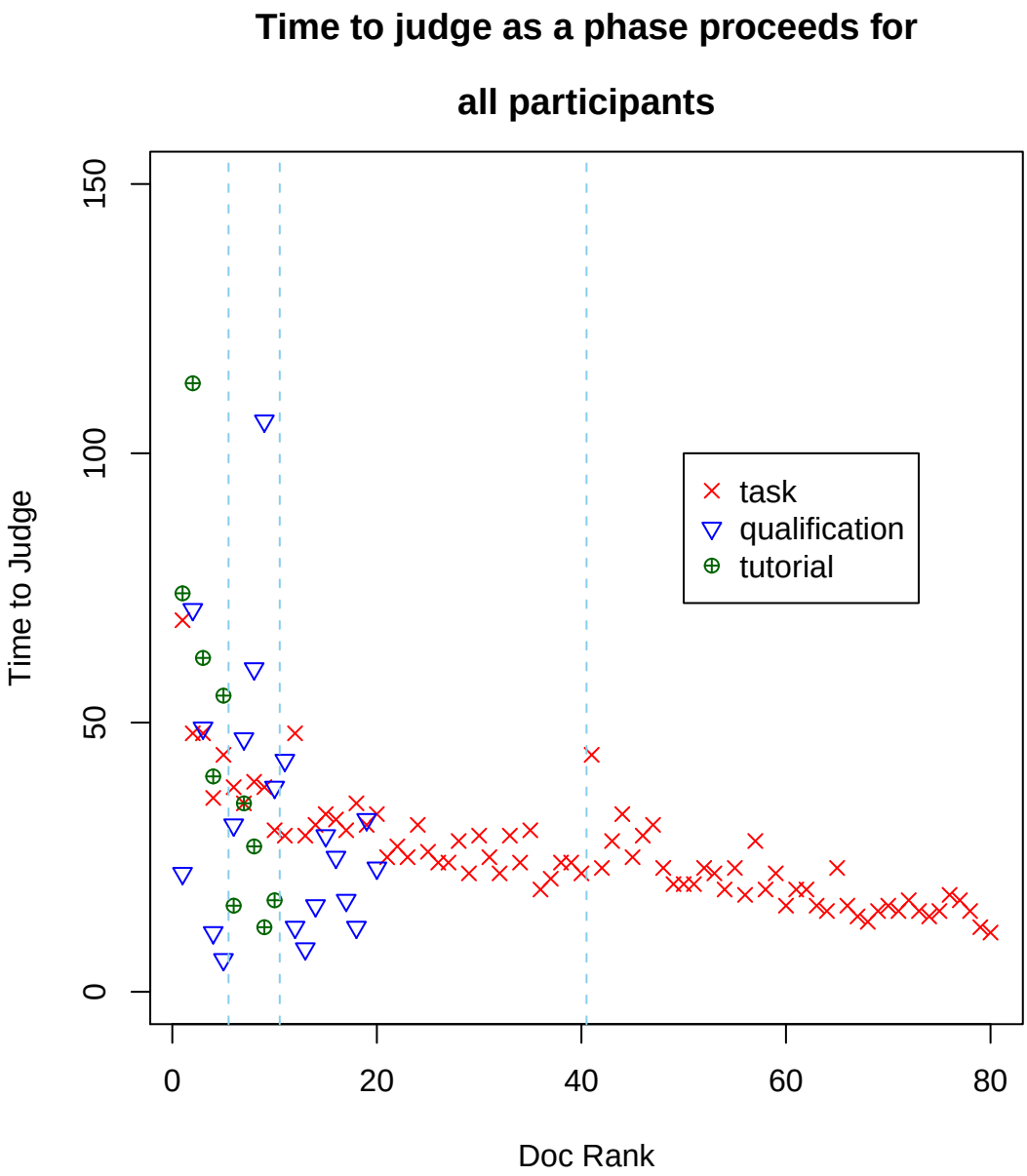


Figure 4.8: Time to judge as different phases proceed for all participants. Three vertical dotted lines from left to right separate the two topics judged in tutorial phase, qualification phase and task phase respectively.

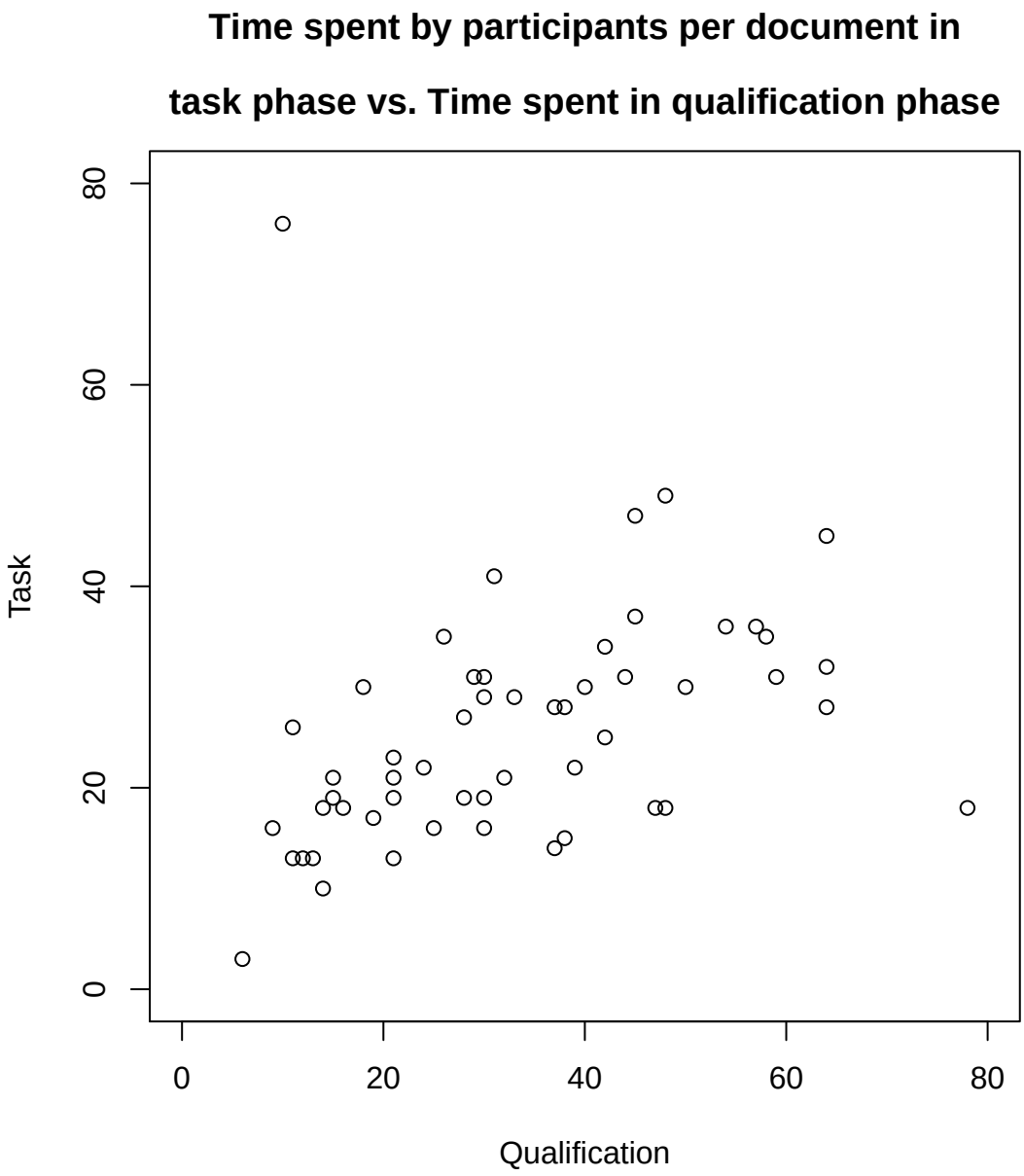


Figure 4.9: Average time spent by participants on a document in task phase vs. in qualification phase.

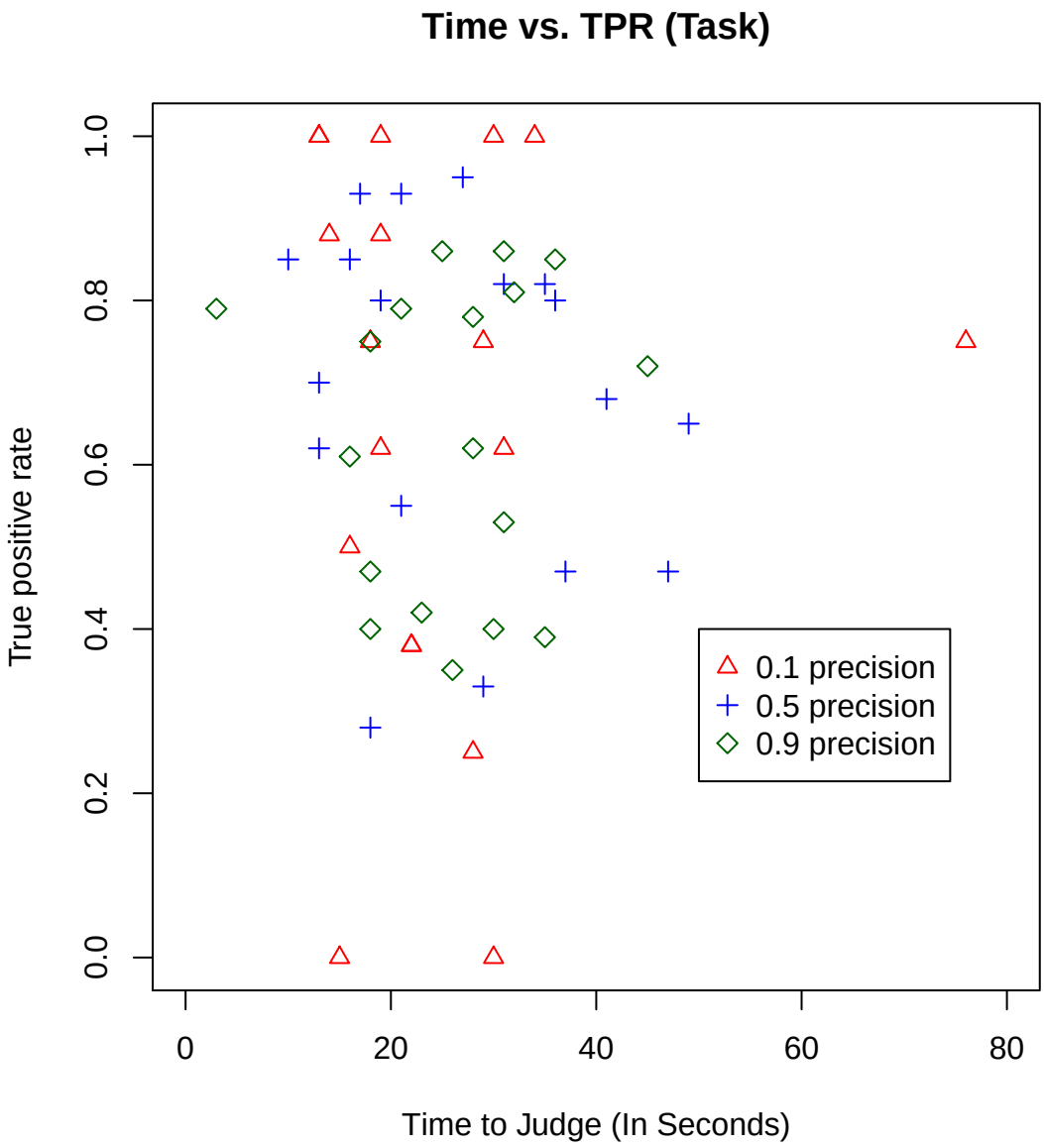


Figure 4.10: Time to judge vs. true positive rate for task phase.

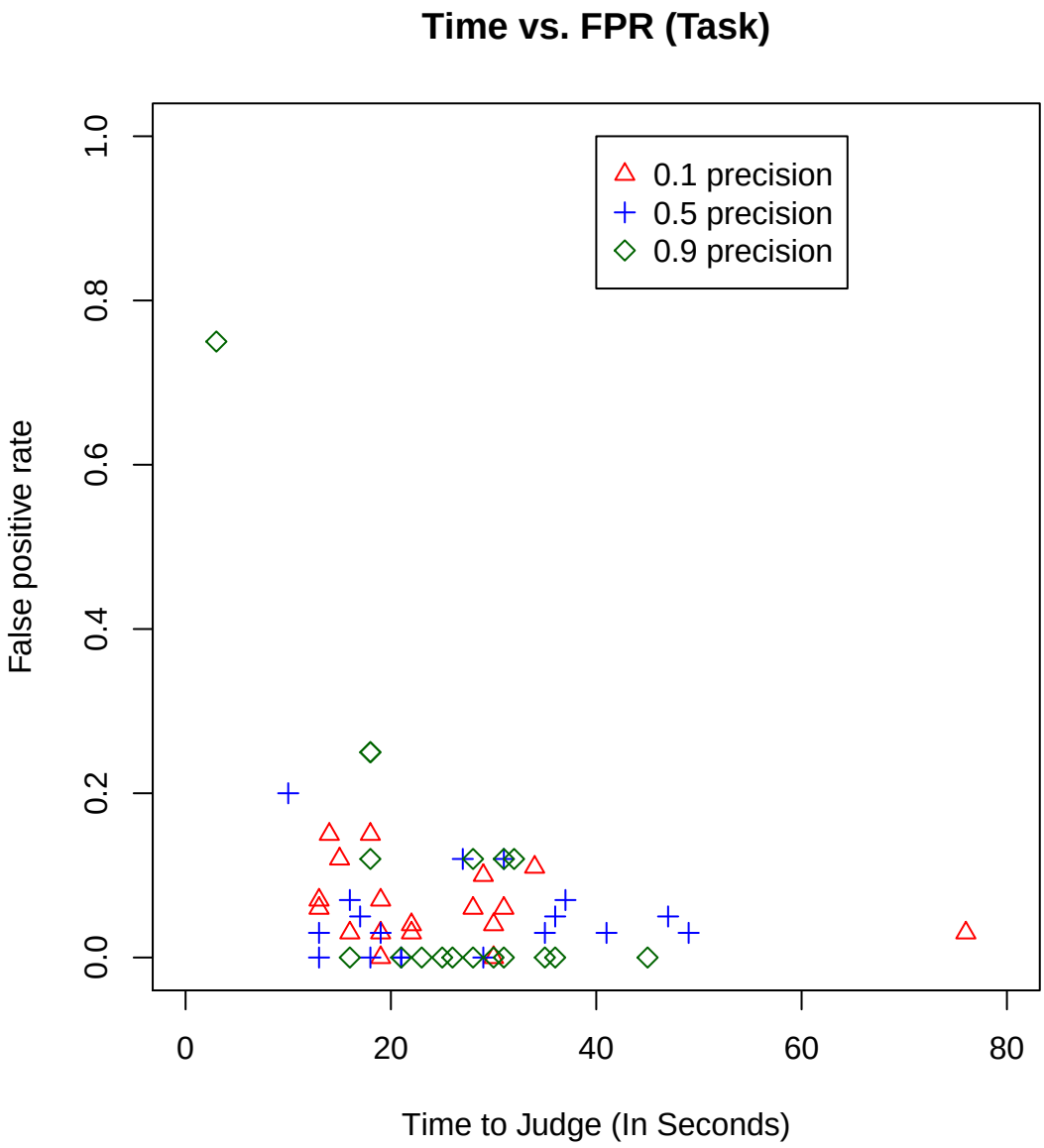


Figure 4.11: Time to judge vs. false positive rate for task phase.

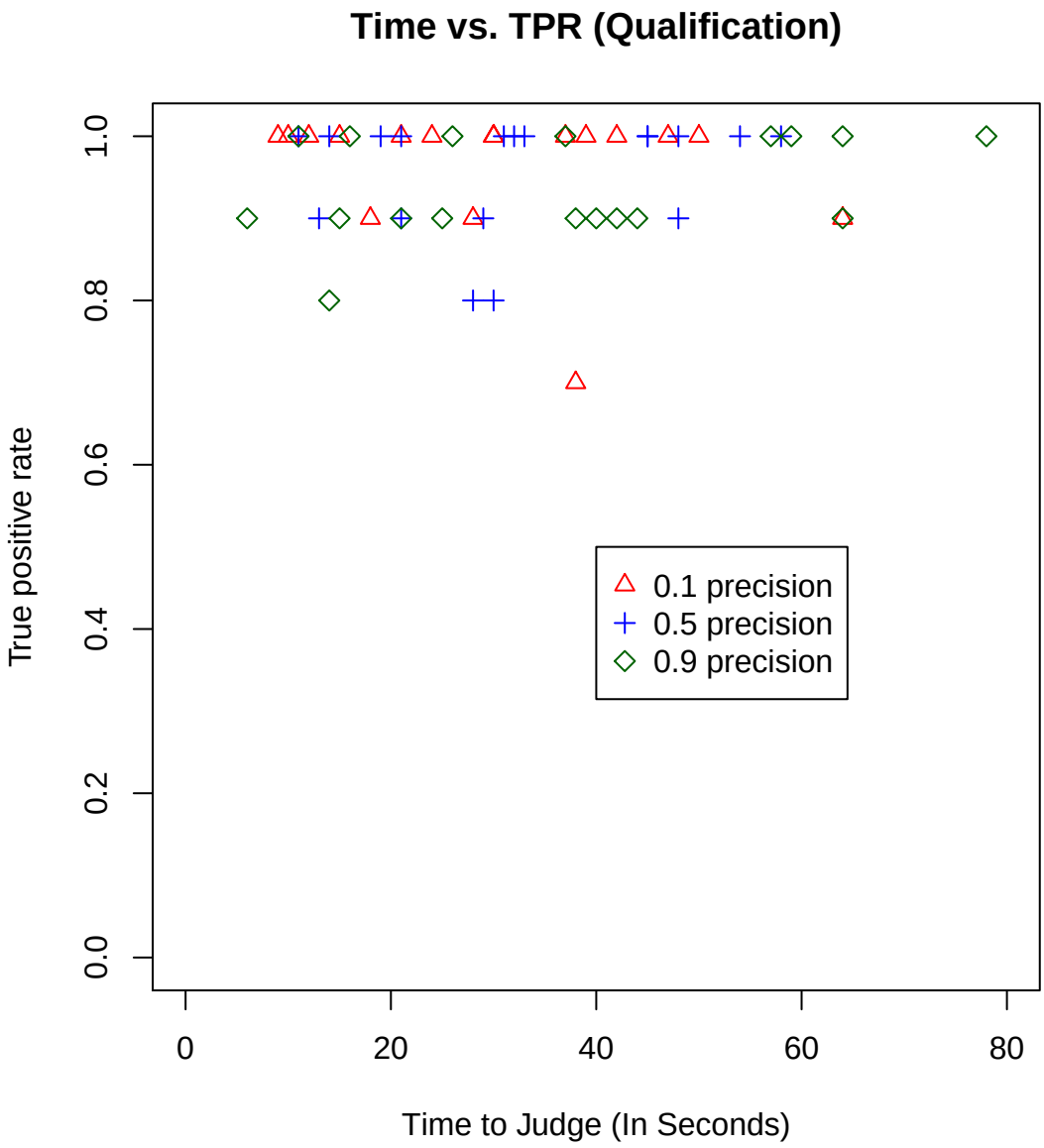


Figure 4.12: Time to judge vs. true positive rate for qualification phase.

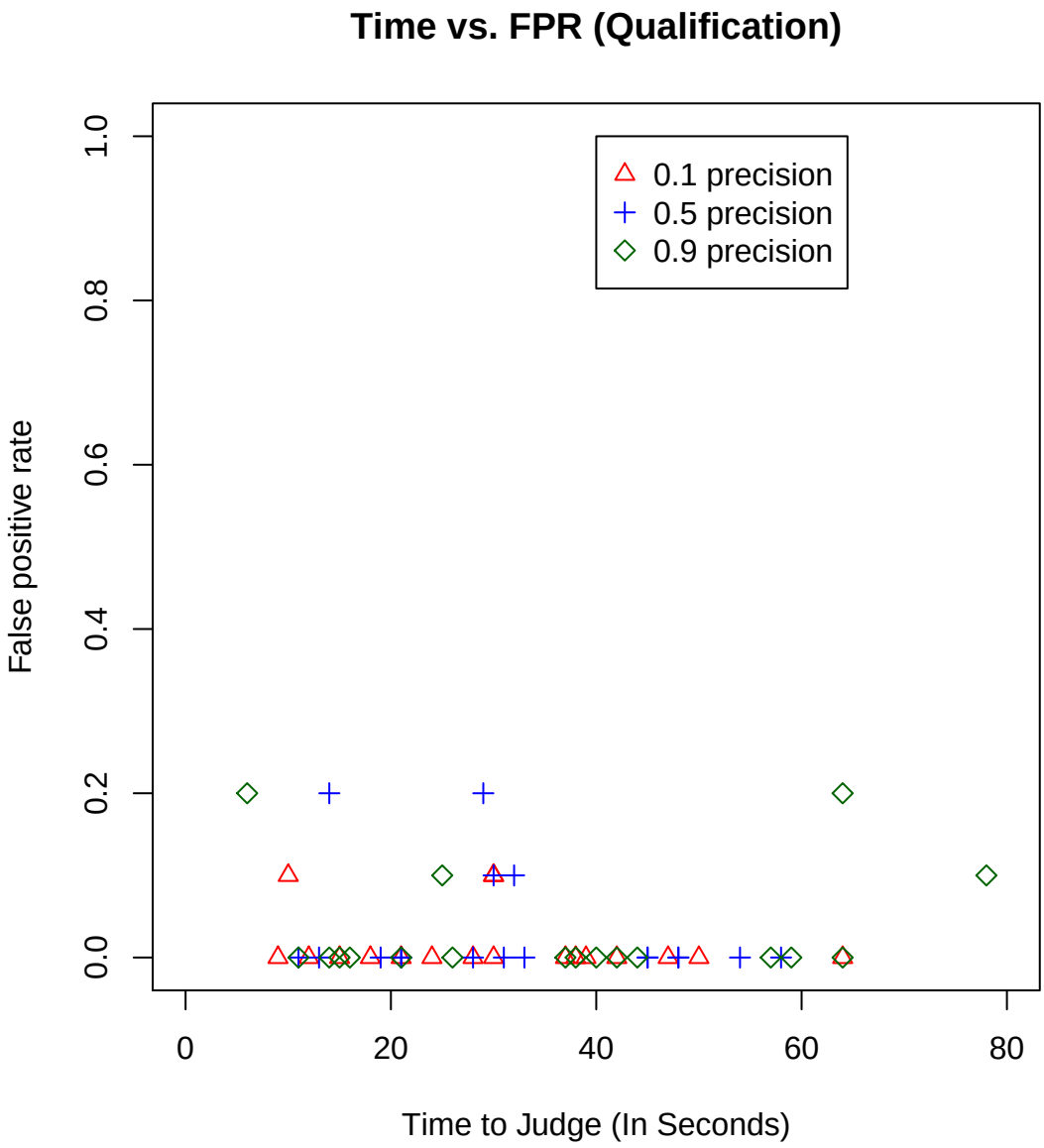


Figure 4.13: Time to judge vs. false positive rate for qualification phase.

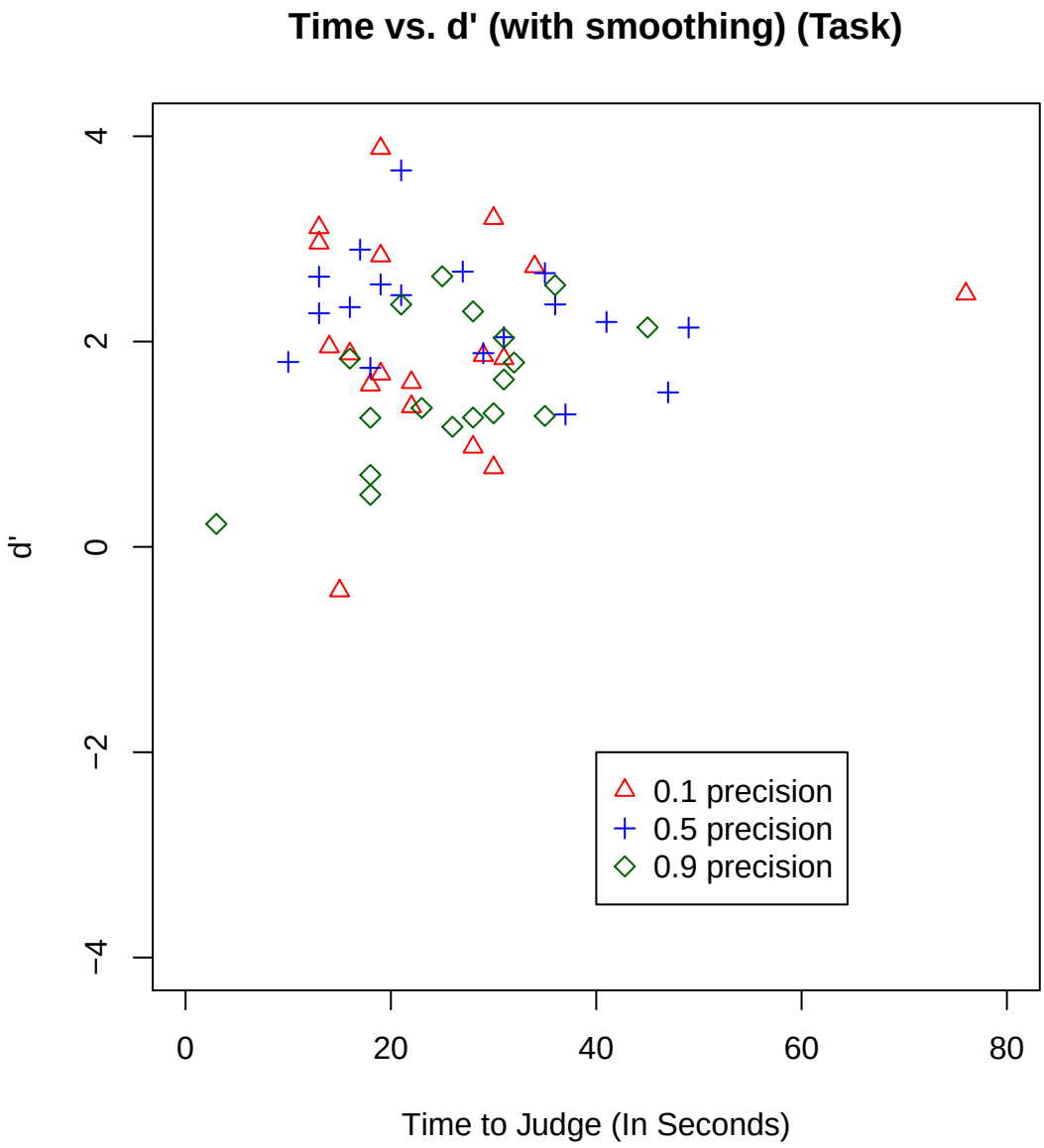


Figure 4.14: Time to judge vs. d' for task phase.

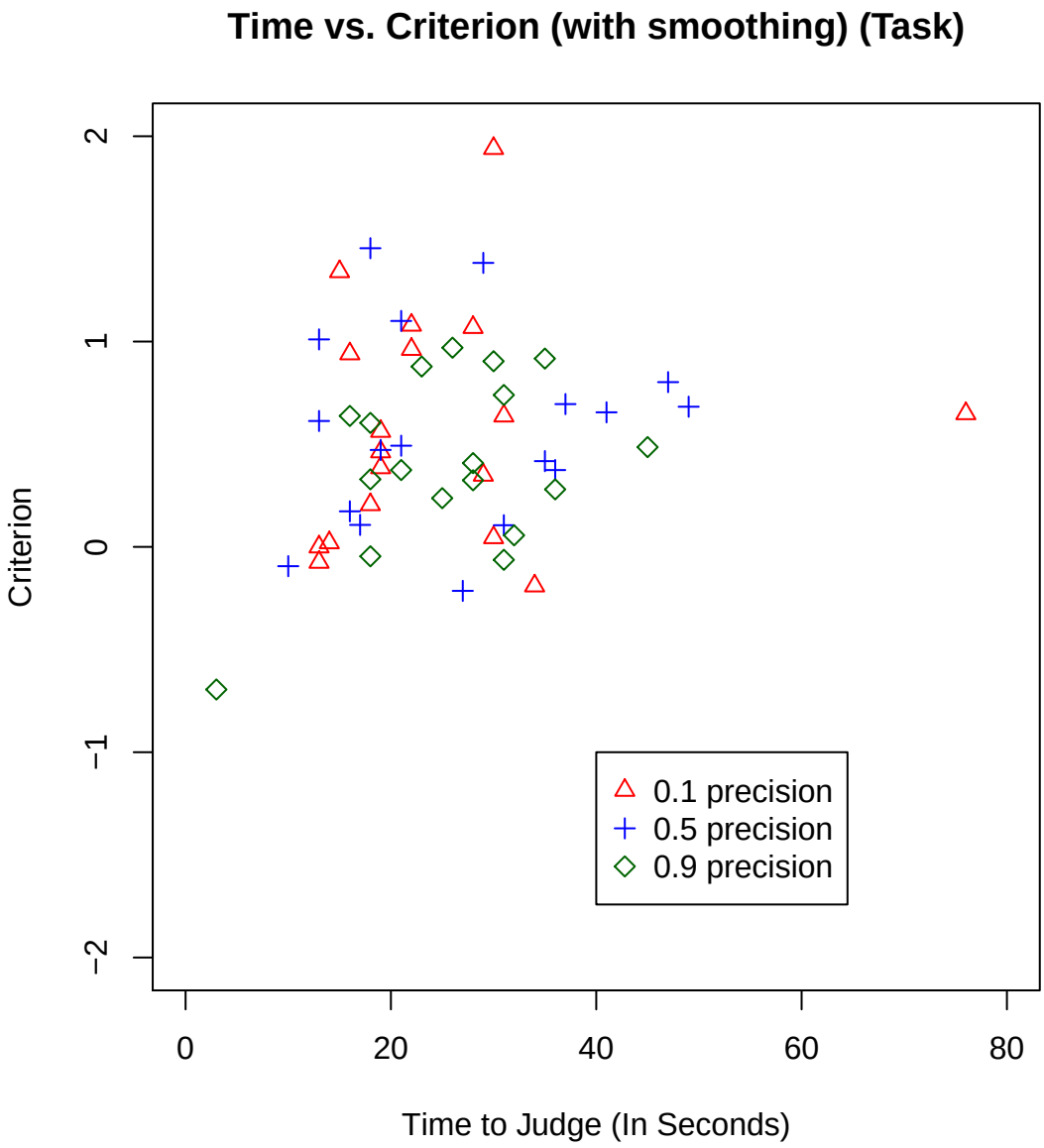


Figure 4.15: Time to judge vs. criterion for task phase.

Chapter 5

Conclusion

In this thesis, we presented a user study which was conducted to understand the behaviour of participants as the precision level changes while judging relevance of documents. We analyzed the behaviour of three different groups of participants by using various measures like true positive rate, false positive rate, accuracy, ability to discriminate (d') and criterion (c).

The key contributions of this thesis are:

- We showed that prevalence affects the quality of participants' judgements. Participants' true positive rates and false positive rates are the best when they work on 0.5 precision sets, although no statistically significant difference was found between the rates at 0.1 and the rates at 0.5.
- Prevalence also has an effect on amount of time spent on the documents while judging. If we consider NIST judgements as the gold standard, participants spend more time on relevant documents than on non-relevant documents. If we consider participants'

judgements as the source of truth, this behaviour changes for participants working on 0.9 precision sets: they spend more time on non-relevant documents. As prevalence increases, time spent on NIST relevant documents decreases.

- We showed that participants start slowly when the task begins and speed up and judge documents faster as the task proceeds.

- It appears that there is little to no correlation between the time to judge a document and the quality of judgements produced. Our data shows that participants appear to work at their own pace to produce a given quality of work.

In summary, we can say that relevance judging studies should avoid documents sets with high level of precisions. Precision 0.5 appears to be better than precision 0.9, but we did not find statistically significant difference between the judging quality at precision 0.1 and at precision 0.5.

APPENDICES

Appendix A

Ethics Application and Recruitment Email

APPLICATION FOR ETHICS REVIEW OF RESEARCH INVOLVING HUMAN PARTICIPANTS

Please remember to PRINT AND SIGN the form, and forward TWO copies to the Office of Research Ethics, Needles Hall, Room 1024, with all attachments.

A. GENERAL INFORMATION

1. Title of Project: A Study of Document Relevance Judging Behavior

2. a) Principal and Co-Investigator(s)
Name Department Ext: e-mail:

2. b) Collaborator(s)
Name Department Ext: e-mail:

3. Faculty Supervisor(s)

Name	Department	Ext:	e-mail:
Mark D. Smucker	Management Sciences	38620	mark.smucker@uwaterloo.ca

4. Student Investigator(s)

Name	Department	Ext:	e-mail:	Local Phone #:
Chandra Prakash Jethani	Computer Science, School of	34822	cpjethan@cs.uwaterloo.ca	+1-226-220-2400

5. Level of Project: MMath **Specify Course:**

Research Project/Course Status: New Project\Course

6. Funding Status (if there is an industry sponsor and procedures pose greater than minimal risk, then Appendix B is to be completed):

Is this project currently funded? Yes

- If Yes, provide Name of Sponsor and include the title of the grant/contract: NSERC
- If No, is funding being sought OR if Yes, is additional funding being sought? No
- Period of Funding: April 1, 2009 to March 31, 2015

7. Does this research involve another institution or site? No
If Yes, what other institutions or sites are involved:

8. Has this proposal been, or will it be, submitted to any other Research Ethics Board/Institutional

Review Board? No

9. For Undergraduate and Graduate Research:

Has this proposal received approval of a Department Committee? Not Dept. Req.

10. a) Indicate the anticipated commencement date for this project: 2/15/2011

b) Indicate the anticipated completion date for this project: 8/31/2011

B. SUMMARY OF PROPOSED RESEARCH

1. Purpose and Rationale for Proposed Research

a. Describe the purpose (objectives) and rationale of the proposed project and include any hypothesis(es)/research questions to be investigated. For a clinical trial/medical device testing summarize the research proposal using the following headings: Purpose, Hypothesis, Justification, and Objectives. Where available, provide a copy of a research proposal. For a clinical trial/medical device testing a research proposal is required:

Purpose: To collect human relevance judging data regarding full document relevance in the context of text retrieval as the prevalence of relevant documents is varied.
Hypothesis: We hypothesize that users ability to find relevant documents in information retrieval systems varies with prevalence of relevant documents in the system.
Justification for the Study: Evaluation of information retrieval systems requires the collection of relevance judgments from human assessors. In our previous research, we have seen some indication of a prevalence effect. This study is designed to verify if this effect exists, which if so, impacts how IR evaluation is done.
Objectives: The long term objective of this research is to improve evaluation of IR systems.

b. In lay language, provide a one paragraph (approximately 100 words) summary of the project including purpose, the anticipated potential benefits, and basic procedures used.

In this study, we will ask participants to make judgments regarding the relevance of documents to various search topics. After collecting this data, we will analyze it to determine if the prevalence of relevant documents affects the judging behavior of assessors. The study results may be used to adjust how IR researchers collect relevance assessments in the future.

C. DETAILS OF STUDY

1. Methodology/Procedures

a. Indicate all of the procedures that will be used. Append to form 101 a copy of all materials to be used in this study.

Computer-administered task(s) or survey(s) None are standardized.
Unobtrusive observations
Logging of computer usage

b. Provide a detailed, sequential description of the procedures to be used in this study. For studies involving multiple procedures or sessions, provide a flow chart. Where applicable, this section also should give the research design (e.g., cross-over design, repeated measures design).

For this user study we'll be using slight modification of the protocol given by Toms et al. "WiIRE: the Web interactive information retrieval experimentation system prototype," Information Processing and Management, 40, 2004, pp. 655-675.

Protocol

1. Introduction

2. Content Form

3. Demographic and Background Questionnaire

4. Overview of Experiment

5. Tutorial

6. Practice Interface

7. Qualification Task

if participant fails, go to 11.

8. Pre Task Questionnaire

9. Task

10. Post Task Questionnaire

11. Thank You

Study will involve the participants determining the relevance of documents to a given search topic. The participants will be shown documents and the search topic and asked whether or not they think the document is relevant to the search topic. We will vary the prevalence of relevant documents between three levels across the participants. Each participant will only view sets of documents at a given level of prevalence.

We will collect timing information and associated computer usage data unobtrusively during the study.

c. Will this study involve the administration/use of any drug, medical device, biologic, or natural health product? No

2. Participants Involved in the Study

a. Indicate who will be recruited as potential participants in this study.
UW Participants:
Undergraduate students
Graduate students
Faculty and/or Staff

b. Describe the potential participants in this study including group affiliation, gender, age range and any other special characteristics. Describe distinct or common characteristics of the potential participants or a group (e.g., a group with a particular health condition) that are relevant to recruitment and/or procedures (e.g., A group with asbestosis is included. People with this condition tend to be male, 50+ years, worked with asbestos.). If only one gender is to be selected for recruitment, provide a justification for this.

Adults fluent in English, familiar with web search (e.g. Google, Yahoo, Bing), and capable of unassisted use of a computer with keyboard, mouse, and LCD monitor.

c. How many participants are expected to be involved in this study? For a clinical trial, medical device testing, or study with procedures that pose greater than minimal risk, sample size determination information is to be provided, as outlined in Guidance Note C2c.

24 to 60 plus a couple of participants during the pilot phase. Study will involve 4 to 8 topics. We know that human performance in text retrieval varies across both humans and the search topics. Each participant will be completing 2 out of 8 topics and since we need each topic to be judged by at least 3 participants because of the 3 levels of prevalence studied, we would need 8X3/2 or 12 participants to complete a block. 12 participants is on the low side given the known variability of human behavior in search tasks. We may increase the number of participants to up to 60 if needed in blocks of 12. Thus we would need 60 participants at most plus a couple of pilot testing participants. This will be a convenience sample of students and other adults of the University of Waterloo community.

3. Recruitment Process and Study Location

a. From what source(s) will the potential participants be recruited?
Other UW sources: We will send email on various mailing uw mailing lists & posters across the campus.

b. Describe how and by whom the potential participants will be recruited. Provide a copy of any materials to be used for recruitment (e.g. posters(s), flyers, cards, advertisement(s), letter(s), telephone, email, and other verbal scripts).
We will send email on various University of Waterloo mailing lists & post posters across the campus.

c. Where will the study take place? On campus: On campus: CPH 4335

4. Remuneration for Participants
Will participants receive remuneration (financial, in-kind, or otherwise) for participation? Yes
If Yes, provide details:
Participants who complete the full study will be paid $25 for the user study, which should take 2 hours to complete. Participants will be asked to go through a qualification task which will be very similar to actual tasks. This task will take approximately 40 minutes to complete. If participants successfully complete this task, they'll be asked to complete the main tasks which should take 80 minutes to complete. If participants fail to complete the qualification task, they'll not be allowed to work on the main tasks and will be paid $7 for their participation in the qualification task. Should the participants need to leave or are asked to leave in case of obvious non-compliance with study protocol (e.g. reading emails, surfing web), they will be paid on prorated basis rounded up to the nearest dollar.

5. Feedback to Participants

Describe the plans for provision of study feedback and attach a copy of the feedback letter to be used. Wherever possible, written feedback should be provided to study participants including a statement of appreciation, details about the purpose and predictions of the study, restatement of the provisions for confidentiality and security of data, an indication of when a study report will be available and how to obtain a copy, contact information for the researchers, and the ethics review and clearance statement.
Refer to the Checklist for Feedback Sheets on ORE web site:
http://iris.uwaterloo.ca/ethics/human/application/samples/checklistfeedback.htm
Participants will be advised that if they are interested in the outcomes of the study, they may contact the principal investigator at a later time to learn about any resulting publications.

D. POTENTIAL BENEFITS FROM THE STUDY

1. Identify and describe any known or anticipated direct benefits to the participants from their involvement in the project.

There are no known direct benefits to the participants from their involvement in the project.

2.Identify and describe any known or anticipated benefits to the scientific community/society from the conduct of this study.

Information retrieval (text search) has become part of daily life for many Canadians, as well as people around the world. This study has the long term potential to allow researchers to better evaluate retrieval systems. With better evaluation tools that allow for faster and more accurate evaluations, the rate at which retrieval systems improve should increase. With better retrieval systems, people are able to find information previously hidden and the more relevant information people have, the better decisions they are able to make.

E. POTENTIAL RISKS TO PARTICIPANTS FROM THE STUDY

1. For each procedure used in this study, describe any known or anticipated risks/stressors to the participants. Consider physiological, psychological, emotional, social, economic risks/stressors. A study–specific current health status form must be included when physiological assessments are used and the associated risk(s) to participants is minimal or greater.

Minimal risks anticipated.
Participants will be asked to use a computer with keyboard, mouse, and LCD monitor to answer brief questionnaires as well as to read and make decisions about documents and document summaries. These activities are common to everyday life and pose no greater risk. The search topics that will be utilized are those that might be used by an analyst and none of them deal with matters outside of what is commonly found in major newspapers. All documents come from either major newswire services (Associate Press, etc.) or from U.S. governmental agencies.

If the risk is greater than minimal and the study is industry sponsored, then Appendix B is to be completed.

2. Describe the procedures or safeguards in place to protect the physical and psychological health of the participants in light of the risks/stressors identified in E1.

As the study involves only minimal risk, no explicit procedures or safeguards will be in place other than to provide a safe, usable computer system in a university computing lab commonly used by students.

F. INFORMED CONSENT PROCESS

Researchers are advised to review the Sample Materials section of the ORE website

Refer to sample information letters and consent forms:
http://iris.uwaterloo.ca/ethics/human/application/101samples.htm

1. What process will be used to inform the potential participants about the study details and to obtain their consent for participation?

Information letter with written consent form

2. If written consent cannot be obtained from the potential participants, provide a justification for this.

3. Does this study involve persons who cannot give their own consent (e.g. minors)? No

G. ANONYMITY OF PARTICIPANTS AND CONFIDENTIALITY OF DATA

1. Provide a detailed explanation of the procedures to be used to ensure anonymity of participants and confidentiality of data both during the research and in the release of the findings.

All participants will be issued an anonymous identifier (ID). The mapping from a participant's name to the ID will be maintained for the length of the study. This mapping will kept in a locked cabinet in a secure location during the study and will be destroyed at the completion of the study. After the study concludes, there will be no way to identify a participant to the data. All computer usage will be with computers in a University of Waterloo computer lab and not with personally identifiable computers, i.e. participants will not use their own computer. All data collected will be retained indefinitely and will be used for research purposes. We may refer to individual participants when describing the results or the study, and in these cases, we will always refer to "participant 1" or some other similar anonymous name. Participants' names will never appear in any publication that results from this study.

2. Describe the procedures for securing written records, video/audio tapes, questionnaires and recordings. Identify (i) whether the data collected will be linked with any other dataset and identify the linking dataset and (ii) whether the data will be sent outside of the institution where it is collected or if data will be received from other sites. For the latter, are the data de-identified, anonymized, or anonymous?

The document test collection that we use comes from the U.S. National Institute of Standards and Technology (NIST). This is a publicly available dataset. By our very use of this dataset, we will "link" with it, but we will not be linking your information collected here to any other information that concerns the you personally. We may choose to distribute the data collected to other researchers. All data will be anonymized at the conclusion of the study and prior to any distribution, but each participant's data will remain identifiable as coming from an individual, i.e. "participant 1", "participant 2", etc. We will not publicly share this data, i.e. the data would only be made available to other researchers for research purposes.

3. Indicate how long the data will be securely stored and the method to be used for final disposition of the data.

Paper Records

Data will be retained indefinitely in a secure location.

Electronic Data

Data will be retained indefinitely in a secure location.

Location: Location: Principal investigator's office (paper) and on secure computers.

4. Are there conditions under which anonymity of participants or confidentiality of data cannot be guaranteed?

Yes

If Yes, please provide details:

We will conduct the study in groups of more than one participant at a time. Fellow participants will know those that participated, but the anonymity of the resulting data is guaranteed.

H. DECEPTION

1. Will this study involve the use of deception? No

Researchers must ensure that all supporting materials/documentation for their applications are submitted with the signed, hard copies of the ORE form 101/101A. Note, materials shown below in bold are normally required as part of the ORE application package. The inclusion of other materials depends on the specific type of

projects.

Researchers are advised to review the Sample Materials section of the ORE web site:
http://iris.uwaterloo.ca/ethics/human/application/101samples.htm

Protocol Involves a Drug, Medical Device, Biologic, or Natural Health Product

If the study procedures include administering or using a drug, medical device, biologic, or natural health product that has been or has not been approved for marketing in Canada then the researcher is to complete Appendix A, a Word document. Appendix A is to be attached to each of the two copies of the application that are submitted to the ORE. Information concerning studies involving a drug, biologic, natural health product, or medical devices can be found on the ORE website.

Drug , biologic or natural health product http://iris.uwaterloo.ca/ethics/human/researchTypes/clinical.htm

Medical devices: http://iris.uwaterloo.ca/ethics/human/researchTypes/devices.htm

Appendix A http://iris.uwaterloo.ca/ethics/human/application/101samples.htm

Please **check** below all appendices that are attached as part of your application package:

- Recruitment Materials: A copy of any poster(s), flyer(s), advertisement(s), letter(s), telephone or other verbal script(s) used to recruit/gain access to participants.
- Information Letter and Consent Form(s)*. Used in studies involving interaction with participants (e.g. interviews, testing, etc.)
- Data Collection Materials: A copy of all survey(s), questionnaire(s), interview questions, interview themes/sample questions for open-ended interviews, focus group questions, or any standardized tests.
- Feedback letter *

* Refer to sample letters:
http://iris.uwaterloo.ca/ethics/human/application/101samples.htm

NOTE: The submission of incomplete application packages will increase the duration of the ethics review process.

To avoid common errors/omissions, and to minimize the potential for required revisions, applicants should ensure that their application and attachments are consistent with the *Checklist For Ethics Review of Human Research Application*

http://iris.uwaterloo.ca/ethics/form101/checklist.htm

Please note the submission of incomplete packages may result in delays in receiving full ethics clearance. We suggest reviewing your application with the Checklist For Ethics Review of Human Research Applications to minimize any required revisions and avoid common errors/omissions.
http://iris.uwaterloo.ca/ethics/form101/checklist.htm

INVESTIGATORS' AGREEMENT

I have read the Tri-Council Policy Statement (TCPS): Ethica Conduct for Research Involving Humans

and agree to comply with the principles and articles outlined in the TCPS. In the case of student research, as Faculty Supervisor, my signature indicates that I have read and approved this application and the thesis proposal, deem the project to be valid and worthwhile, and agree to provide the necessary supervision of the student.

Signature of Principal Investigator/Supervisor	**Date**
Signature of Student Investigator	**Date**

FOR OFFICE OF RESEARCH ETHICS USE ONLY:

Susan E. Sykes, Ph.D., C. Psych. **Director, Office of Research Ethics** ***OR*** ***Susanne Santi, M.Math*** ***Senior Manager, Research Ethics*** ***OR*** ***Julie Joza, B.Sc.*** ***Manager, Research Ethics***	**Date**

ORE 101
Revised August 2003

School of Computer Science
University of Waterloo

Participants Needed for Research in Text Search

- We are looking for volunteers to take part in a study of **text search** (if you use Google, Yahoo! or Bing to search the web, that is text search).
- As a participant in this study, you would be asked to complete demographic and task-related questionnaires, and judge the relevance of documents.
- The study will involve a qualification task and judging the relevance of documents for 2 search topics. The qualification task is estimated to take 40 minutes to complete and the judging of documents for 2 search topics is estimated to take 1 hour and 20 minutes to complete.
- In appreciation for your time, you will receive $25 for completing the full study. After going through the qualification task, if you fail to qualify for the full study, you will not continue in the study session and will be paid $7 for your participation.
- For more information about this study, or to volunteer for this study, please contact:

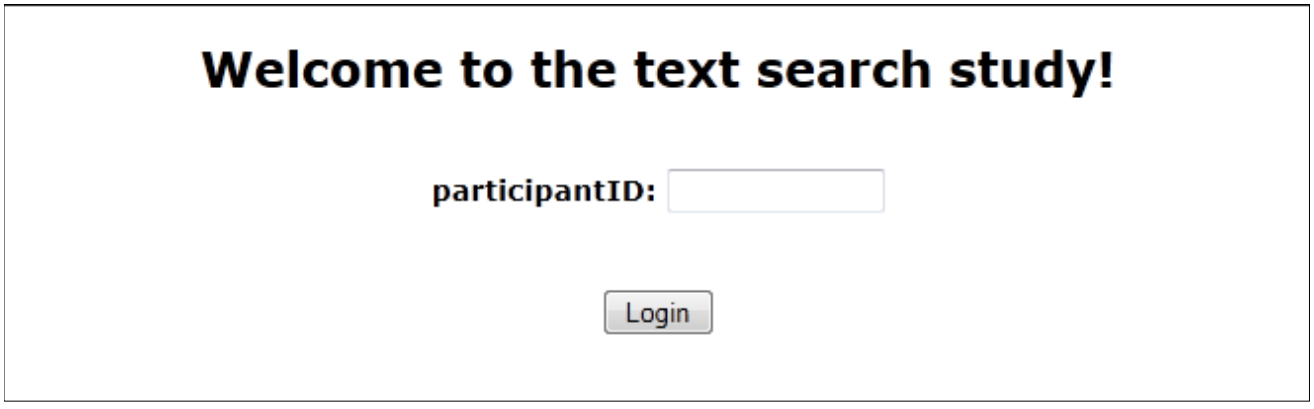
Chandra Prakash Jethani
School of Computer Science
at
Email: cpjethan@cs.uwaterloo.ca

This study has been reviewed by, and received ethics clearance through, the Office of Research Ethics, University of Waterloo.

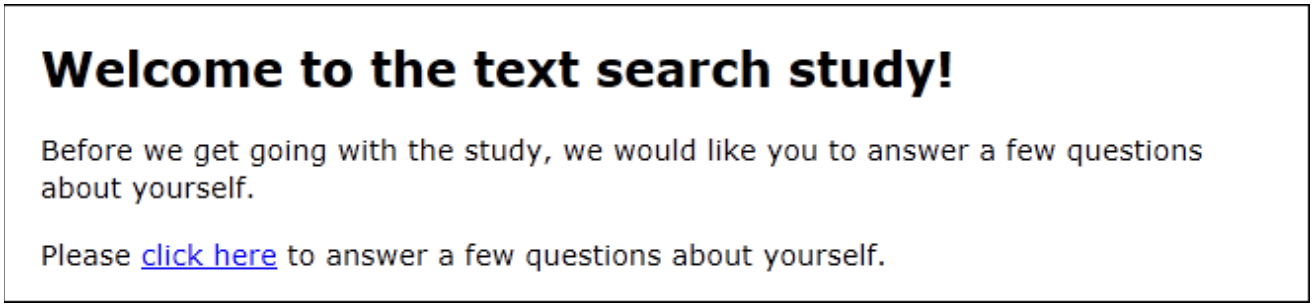
Appendix B

User Study Interfaces

B.1 Welcome Screen



B.2 After login



B.3 Demographic information form

Demographic Information Form

1. What is your age?

2. Are you male or female?
- Male
- Female

3. Are you:
- An undergraduate student
- A graduate student
- Other. Please specify

4. If you are a student, are you
- an arts student
- a science, technology, engineering, or math student
- other

5. How often do you search the internet for information using a search engine such as Google, Yahoo Search, or Microsoft Bing?
- Several times a day
- At least once a day
- At least once a week
- At least once a month
- Rarely (less than one search a month on average)

6. Are you a fluent speaker and reader of English?
- Yes
- No

7. How much do you agree with the following statements?

7a. I am an expert at finding information using search engines like Google, Yahoo, and Microsoft Bing.
- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

7b. I often have trouble finding what I am looking for on the internet.
- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

7c. Friends and family turn to me to help them search the internet for answers to their questions.
- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

7d. I enjoy using search engines like Google, Yahoo, and Microsoft Bing.
- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

7e. I consider myself a fast reader of web pages, magazines, and books.

- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

7f. When I'm in a group and a handout is given for us to read, I'm one of the last to finish reading the handout.

- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

8a. Have you ever had special training or education in searching or information retrieval?

- Yes
- No

8b. If the answer to 8a. is **yes**, please describe the training or education.

Please click the submit button to continue. Submit

B.4 User study instructions

Introduction to the Text Search Study

First, thank you for your answers to those questions.

In this study, you will judge the relevance of documents to a given search topic.

Before you begin, you will be going through a tutorial and a qualification task. In the tutorial you'll practice judging 5 documents for each of 2 search topics. After the tutorial, you need to go through a qualification task where you will judge 10 documents for each of the 2 search topics, for a total of 20 documents. Based on how well you judge the relevance of these 20 documents, we will decide if you qualify for the full study.

If you qualify for the full study, you will complete 2 search tasks. For each task, you will judge 40 documents as relevant or not to a search topic. In total you will judge 80 documents. These 2 search topics will be different from the ones you'll work on in the tutorial and qualification task.

User Study Instructions

To participate in this study, you need to know both how to judge the relevance of a document and what the rules of the study are. Please read these instructions carefully. At the end of the instructions, we will present you with a short quiz about the instructions. You will have to retake the quiz until you pass

First we'll explain what makes a document relevant or not to a given search topic.

Judging the Relevance of a Document to a Search Topic

For each search task, we will present you with a search topic and a series of documents. You will view one document at a time. You are to judge the document as relevant or not to the search topic. After you judge a document, we will show you the next document to judge.

Each search topic has a title and a description. The description attempts to give you detailed instructions on what to consider a relevant document to the search topic. Here is an example search topic:

Search Topic Title: Journalist Risks
Description: Any document identifying an instance where a journalist or correspondent has been killed, arrested or taken hostage in the performance of his work is relevant.

For this search topic, documents should mention an actual incident where a journalist has been hurt or taken hostage. Describing the risks journalists take is not enough for a document to be relevant.

A document is relevant if any portion of the document is relevant. Some documents may contain many non-relevant parts, but if any part of the document is relevant, you should judge the document as relevant.

We will highlight in each document the words from the search topic's title. A document may be relevant even if it has no highlighted words.

In a moment, we will give you the chance to practice judging the relevance of documents. But first, we need to establish a few rules for this user study.

Study Rules

Please follow these simple rules for the duration of the study

- Work as quickly as possible while making as few mistakes as possible. It is important to accurately judge the relevance of documents while being efficient in making your judgments.
- Some participants may finish before other participants. Please focus on your work and continue to judge documents as accurately and as quickly as possible.
- Please work on a given search topic task from start to finish. If you need to take a break, please do so between tasks. We will inform you when it is appropriate to take a break.
- Once you have made a judgment, do not attempt to go back and change your judgment. All judgments are final.

This scientific research study requires your full attention. If you are unable to give this research your full attention, please excuse yourself from the study. In particular:

- Please turn off your mobile phones. Phones may not be used during the study.
- Please put all iPods and music players away. You may not listen to music during the study.
- Do not use the computer for checking email, viewing web pages, or other activities during the study.

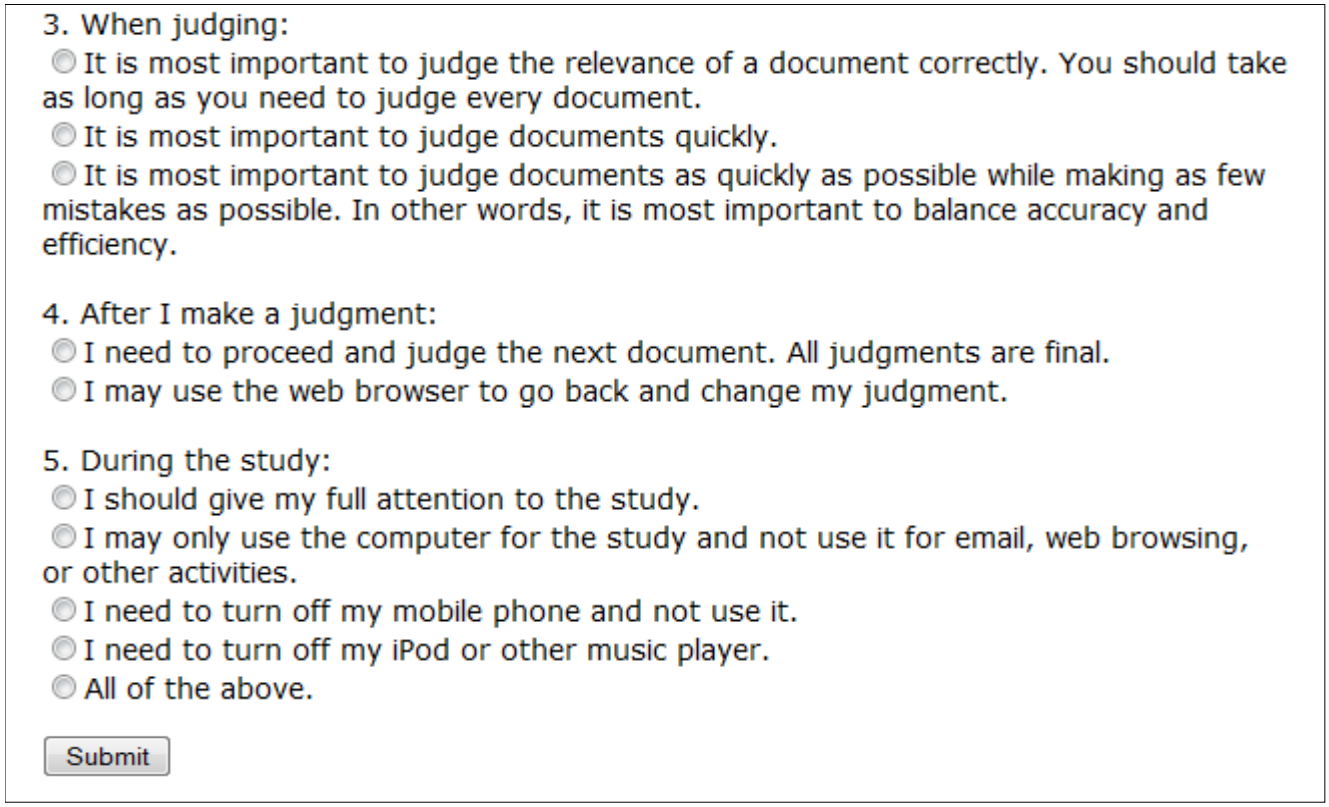
If you use the computer for non-study activities or use a phone or music player, we will end your participation and ask you to leave.

Quiz

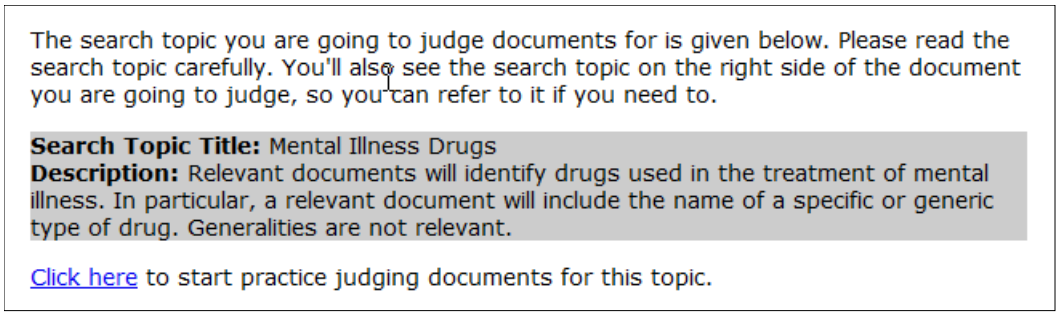
Please read the instructions above as many times as you want. Once you have understood the instructions, please take the quiz below. You will not be able to proceed until you answer all questions correctly.

1. A document should be considered relevant to a search topic when:
- ○ The document contains the words found in the search topic's title.
- ○ The document fits the description of relevance given by the search topic's "description".
- ○ A reasonable person would consider this a relevant document to a search engine query where the query is the search topic's title.

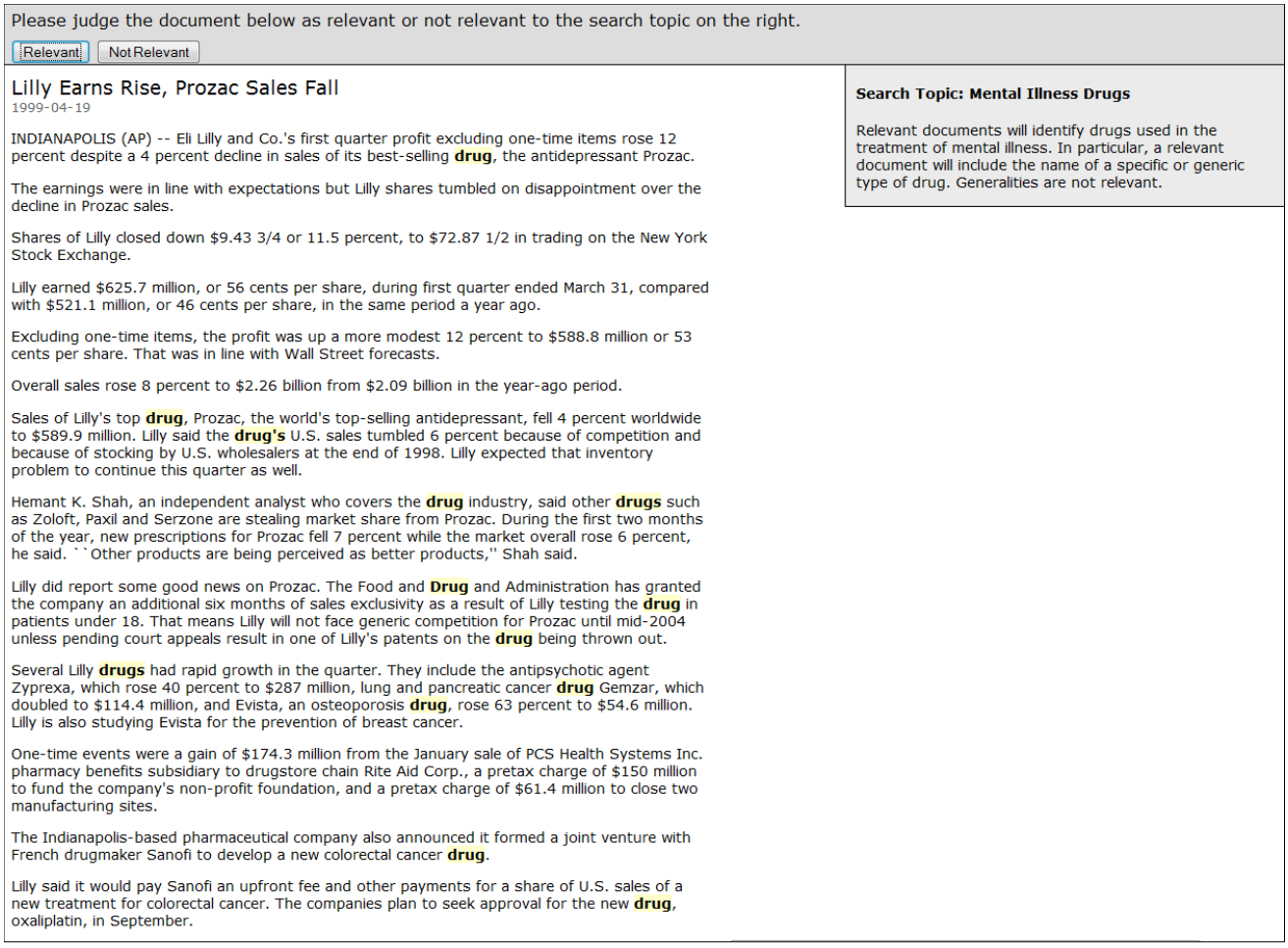
2. For a document to be relevant:
- ○ Any portion of the document must be relevant.
- ○ At least a paragraph of the document must be relevant.
- ○ The entire document must be relevant.



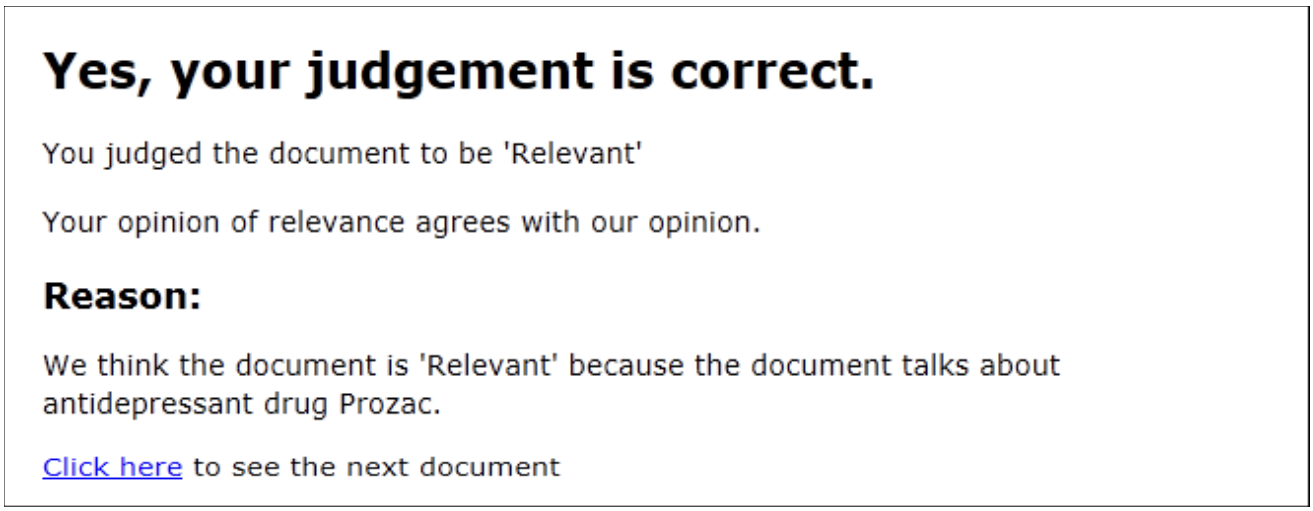
B.5 First topic for tutorial and qualification phases



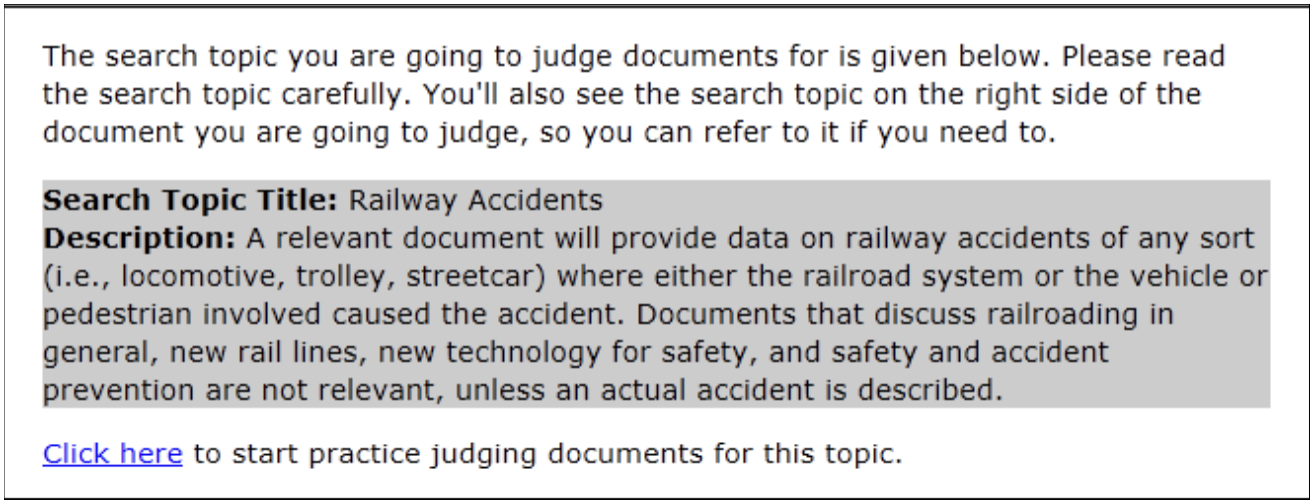
B.6 Example of a document on first topic for tutorial and qualification phases



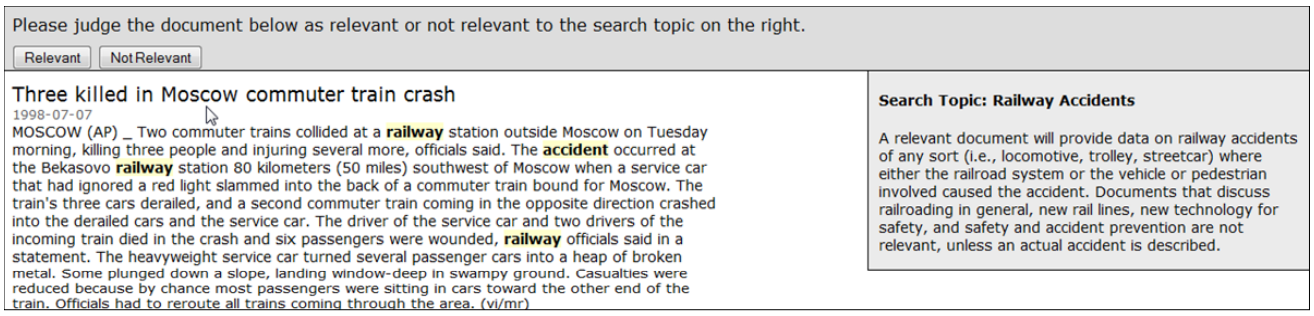
B.7 Example of a feedback on judgement given on first topic for tutorial and qualification phases



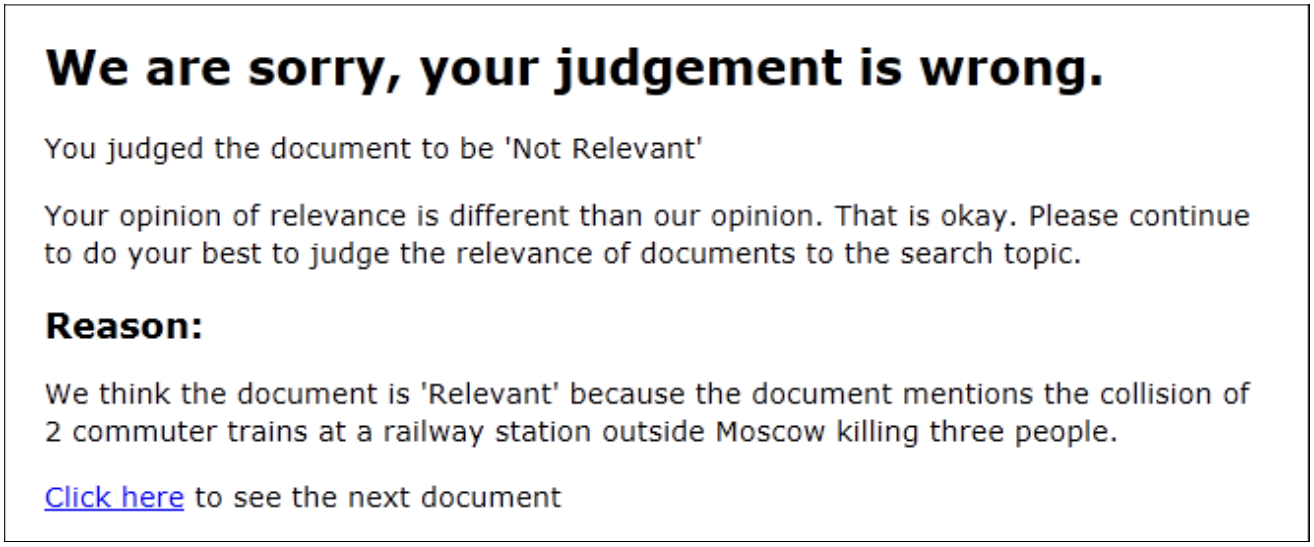
B.8 Second topic for tutorial and qualification phases



B.9 Example of a document on second topic for tutorial and qualification phases



B.10 Example of a feedback on judgement given on second topic for tutorial and qualification phases



B.11 End of tutorial phase

Yes, your judgement is correct.

You judged the document to be 'Relevant'

Your opinion of relevance agrees with our opinion.

Reason:

We think the document is 'Relevant' because the documents mentions derailing of a train killing 10 people and injuring more than 30 people.

Thank you for completing the tutorial.

You judged **7 out of 10** documents correctly.

You will now go through a qualification test. You need to judge 10 documents for each of the 2 topics. These 2 search topics are same as the search topics used in the tutorial.

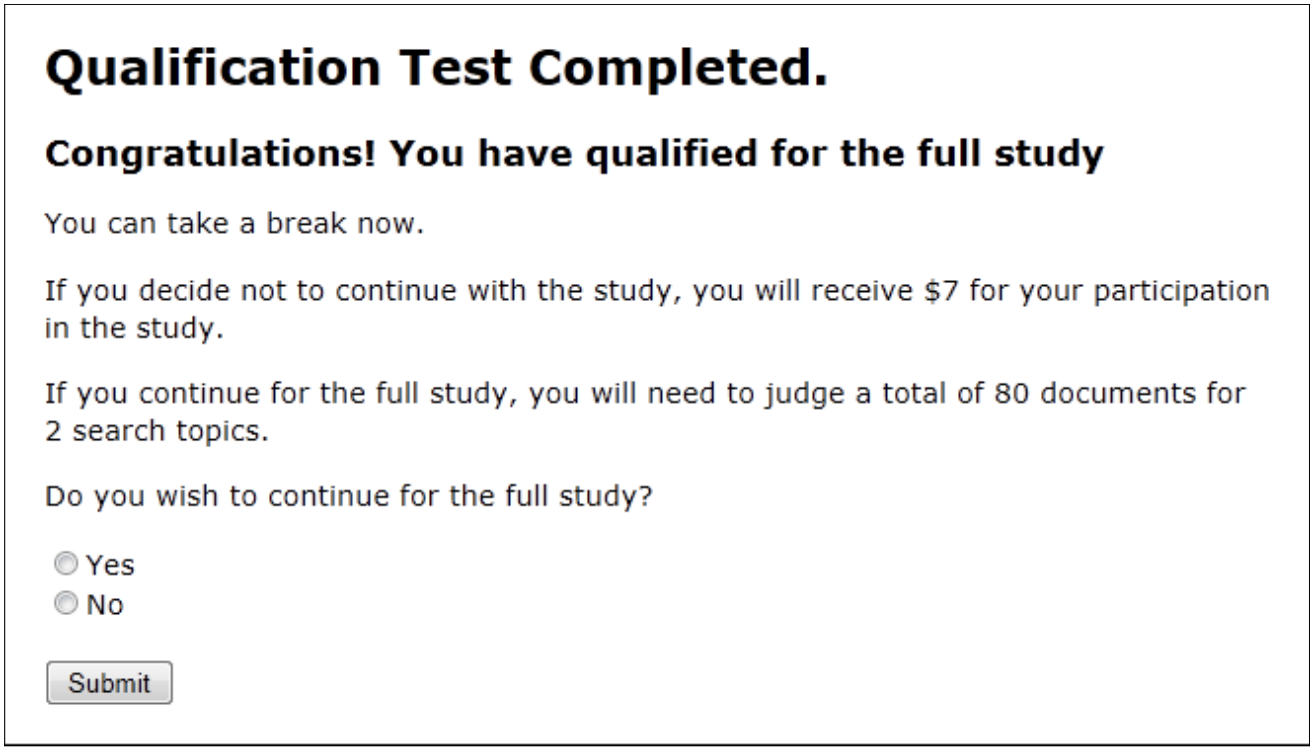
We will show you one topic at a time and ask you to judge 10 documents for that topic.

Work as quickly as possible while making as few mistakes as possible. It is important to accurately judge the relevance of documents while being efficient in making your judgments.

Based on your performance, we will decide whether you qualify for the full study.

Click here to start judging documents for the first topic.

B.12 End of qualification phase



B.13 Pre-task and post-task questionnaires

Pre-Task Questionnaire

Search Topic : Human Smuggling

A relevant document shows an incident of humans (at least ten) being smuggled. The smugglers would have to realize a monetary gain for their actions, while the people being smuggled may or may not be willing participants.

1. How much do you know about this topic ?
- ○ Nothing
- ○ Heard of it
- ○ Know generally about it
- ○ Quite familiar with topic
- ○ Know details about topic

2. How difficult do you think it will be to find relevant documents for this topic ?
- ○ Very Difficult
- ○ Difficult
- ○ Neutral
- ○ Easy
- ○ Very Easy

3. How relevant is this topic to your life ?
- ○ Not at all
- ○ Not much
- ○ Neutral
- ○ Somewhat
- ○ Very much

4. How interested are you to learn more about this topic ?
- ○ Not at all
- ○ Not much
- ○ Neutral
- ○ Somewhat
- ○ Very much

Please click the submit button to begin judging documents for relevance to the search topic. Thanks.

Submit

Post-Task Questionnaire

Search Topic : Human Smuggling

A relevant document shows an incident of humans (at least ten) being smuggled. The smugglers would have to realize a monetary gain for their actions, while the people being smuggled may or may not be willing participants.

1. How difficult was it to find relevant documents about this topic?
- Very Difficult
- Difficult
- Neutral
- Easy
- Very Easy

2. How would you rate your experience searching for information about this topic?
- Very Unenjoyable
- Unenjoyable
- Neutral
- Enjoyable
- Very Enjoyable

3. How would you rate your mood while you searched?
- Very Bored
- Bored
- Neutral
- Engaged
- Very Engaged

4. How hard was it to concentrate while you searched?
- Very Hard
- Hard
- Neutral
- Easy
- Very Easy

5. Did you encounter any issues while completing this task? If yes, please describe.

Please click the submit button to continue. Submit

References

Omar Alonso and Stefano Mizzaro. Can we get rid of trec assessors? using mechanical turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on The Future of IR Evaluation*, Boston, Massachusetts, 2009. 15

Omar Alonso, Daniel E. Rose, and Benjamin Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, November 2008. 15

Earl W. Bailey, Diane Kelly, and Karl Gyllstrom. Undergraduates' evaluations of assigned search topics. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 812–813, Boston, MA, USA, 2009. 26

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, Singapore, Singapore, 2008. 11

George Buchanan and Fernando Loizides. Investigating document triage on paper and electronic media. In *European Conference on Digital Libraries*, pages 416–427, 2007. 10

Robert Burgin. Variations in relevance judgments and the evaluation of retrieval performance. *Information Processing & Management*, 28:619–627, July 1992. ISSN 0306-4573. 12

Cyril W. Cleverdon. The cranfield tests on index language devices. In *Aslib Proceedings*, volume 19, pages 172–192, 1967. 1

Cyril W. Cleverdon. The effect of variations in relevance assessments in comparative experimental tests of index languages. Technical Report Cranfield Library Report No. 3, Cranfield Institute of Technology, October, 1970. 12

C. Cool, N. J. Belkin, O. Frieder, and P. Kantor. Characteristics of texts affecting relevance judgments. In *Proceedings of the 14th National Online Meeting*, pages 77–84, 1993. 10

William S. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7(1):19–37, 1971. 9

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, Boston, Massachusetts, USA, 2009. 37

Carlos A. Cuadra and Robert V. Katter. Experimental studies of relevance judgments: Final report. Report TM-3520, System Development Corporation, Santa Monica, California, June, 1967. 3, 10

Mathias S. Fleck and Stephen R. Mitroff. Rare targets are rarely missed in correctable search. *Psychological Science*, Vol. 18, pages 943-947, 2007. 17

Maura R. Grossman and Gordon V. Cormack. Inconsistent assessment of responsiveness in e-discovery: Difference of opinion or human error? In *ICAIL 2011 Workshop on Setting Standards for Searching Electronically Stored Information in Discovery*, Pittsburgh, PA, USA, 2011. 13

David Gur, Howard E. Rockette, Derek R. Armfield, Arye Blachar, Jennifer K. Bogan, Giuseppe Brancatelli, Cynthia A. Britton, Manuel L. Brown, Peter L. Davis, James V. Ferris, Carl R. Fuhrman, Sara K. Golla, Sanj Katyal, Joan M. Lacomis, Barry M. McCook, F. Leland Thaete, and Thomas E. Warfel. Prevalence effect in a laboratory environment. *Radiology*, 228(1):10-4, 2003. 16

Donna Harman. *Information Retrieval Evaluation*, pages 32–33. Morgan & Claypool Publishers, 2011. 33

Stephen P. Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47:37–49, January 1996. ISSN 0002-8231. 12

Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, Florence, Italy, 2008. 15

John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, Geneva, Switzerland, 2010. 15

Michael E. Lesk and Gerard Salton. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, pages 343–359, 1968. 11

Neil A. Macmillan and C. Douglas Creelman. *Detection Theory: A User's Guide*. Lawrence Erlbaum Associates, 2005. 31, 32

T. Saracevic. Comparative effects of titles, abstracts and full text on relevance judgments. *Journal of the American Society for Information Science*, 22:126–139, 1969. 10

Mark D. Smucker and Chandra Prakash Jethani. Human performance and retrieval precision revisited. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 595–602, Geneva, Switzerland, 2010a. ix, 3, 5, 12, 13, 23, 24, 36, 37

Mark D. Smucker and Chandra Prakash Jethani. Impact of retrieval precision on perceived difficulty and other user measures. In *Proceedings of the Fourth Workshop on Human-Computer Interaction and Information Retrieval*, New Brunswick, NJ, USA, 2010b. 13, 59

Mark D. Smucker and Chandra Prakash Jethani. Measuring assessor accuracy: A comparison of NIST assessors and user study participants. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Beijing, China, 2011. 13

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, 2008. 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, Melbourne, Australia, 1998. 12

Ellen M. Voorhees. Overview of the TREC 2005 robust retrieval track. In *Proceedings of the Fourteenth Text REtrieval Conference*, TREC 2005, Gaithersburg, MD, USA, 2005a. 6, 20

Ellen M. Voorhees and Donna K. Harman, editors. *TREC : Experiment and Evaluation in Information Retrieval*. Digital Libraries and Electronic Publishing series. MIT Press, 2005b. 2

Jianqiang Wang. Accuracy, agreement, speed, and perceived difficulty of users relevance judgments for ediscovery. In *Proceedings of the SIGIR 2011 Workshop on E-Discovery*, Beijing, China, 2011. 14

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 - Volume 47*, Pittsburgh, Pennsylvania, 2010. 14

Patrick Wilson. Situational relevance. *Information Storage and Retrieval*, pages 457–471, August 1973. 9

Jeremy M. Wolfe and Michael J. Van Wert. Varying target prevalence reveals two dissociable decision criteria in visual search. *Current Biology*, Vol. 20, Issue 2, pages 121-124, 2010. 5, 17

Jeremy M. Wolfe, Todd S. Horowitz, and Naomi M. Kenner. Rare items often missed in visual searches. *Nature*, Vol. 435, pages 439-440, 2005. 5, 16

Jeremy M. Wolfe, Todd S. Horowitz, Michael J. Van Wert, Naomi M. Kenner, Skyler S. Place, and Nour Kibbi. Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology*, Vol. 136, No. 4, pages 623-638, 2007. 17