

The Effects of Time Constraints and Document Excerpts on Relevance Assessing Behavior

by

Shahin Rahbariasl

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, 2018

© Shahin Rahbariasl 2018

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

Assessors who judge the relevance of documents to the search topics and perform the relevance assessment process are one of the main parts of Information Retrieval (IR) systems evaluations. They play a significant role in making test collections which can be used in evaluations and system designs. Relevance assessment is also highly important for e-discovery where relevant documents and materials should be found with acceptable cost and in an efficient way. In order to study the relevance judging behavior of assessors better, we conducted a user study to further examine the effects of time constraints and document excerpts on relevance behavior. Participants were shown either full documents or document excerpts that they had to judge within 15, 30, or 60 seconds time constraint per document. For producing document excerpts or paragraph-long summaries, we have used algorithms to extract what a model of relevance considers most relevant from a full document. We found that the quality of judging slightly differs within each time constraint but not significantly. While time constraints have little effect on the quality of judging, they can increase the judging speed rate of the assessors. We also found that assessors perform as good and in most cases better if shown a paragraph-long document excerpt instead of a full document, therefore, they have the potential to replace full documents in relevance assessment. Since document excerpts are significantly faster to judge, we conclude that showing document excerpts or summaries to the assessors can lead to better quality of judging with less cost and effort.

Acknowledgements

I would like to thank all the people who made this possible. At the top of the list, my supervisor, Professor Mark D. Smucker. He guided me patiently and always gave me insightful advices throughout the course of my degree. I would like to express my deepest respect and appreciation for everything I have learned from him during these two years and his support and understanding whenever I needed help or advice either on my research or outside the research area. It was a genuine honor to work with him.

I would also like to thank and express my sincere gratitude to Professors Gordon V. Cormack and Maura R. Grossman for being my thesis readers and providing valuable and helpful comments and suggestions to improve this thesis.

I thank all my office-mates and friends in the data systems lab and information retrieval group. I enjoyed all the time I spent doing research with them.

I also thank all my friends in Canada for the great time I have had with them.

I would also like to thank all the participants who took part in my study and helped me with achieving the goal of this thesis.

At last, my special thanks and appreciation extend to my parents. They were not here with me, but their thought was always in my mind. I thank them for all their support and love since day one.

Dedication

To my parents.

Table of Contents

List of Tables	x
List of Figures	xii
1 Introduction	1
1.1 Background Overview	2
1.1.1 IR Evaluation and Test Collections	2
1.1.2 Evaluation Metrics	3
1.1.3 Differences in Primary and Secondary Assessors	3
1.2 Problem Statement	4
1.3 Thesis Overview	5
1.4 Contributions	6
2 Literature Review	7
2.1 The Definition of Relevance	7
2.2 Precision and Uncertainty in Relevance Assessment	10
2.3 Disagreement in Relevance Assessment	10
2.4 Crowdsourcing in Relevance Assessments	12
2.5 Speed and Accuracy in Relevance Assessment	13

3	Data Set	15
3.1	Topic Selection	15
3.2	Topic 310: Radio Waves and Brain Cancer	16
3.3	Topic 336: Black Bear Attacks	16
3.4	Human Smuggling	16
3.5	Topic 367: Piracy	17
3.6	Topic 426: Law enforcements, Dogs	17
3.7	Topic 427: UV damage, eyes	17
3.8	Topic 436: Railway Accidents	18
3.9	Document Selection	18
3.10	Using Document Excerpts Instead of Full Docs	19
4	User Study	21
4.1	Study Design	21
4.1.1	Study Protocol	21
4.1.2	Tutorial	22
4.1.3	Main Task	22
4.2	User Interface	23
4.2.1	Document length for the task phase	24
4.2.2	Time constraints for the task phase	24
4.3	Balanced Design	24
4.4	Participants	25
5	Results	26
5.1	Analysis of Accuracy	26
5.2	Analysis of Rates	30
5.3	Analysis of Time	31
5.4	Analysis of Assessor Criterion and Ability to Discriminate	36

5.5	Time Vs. TPR and FPR	37
5.6	Time Vs. d' and Criterion	39
5.7	Speed, Accuracy and topics	39
5.8	Inconsistency in Relevance Judgments	45
5.9	Participants Perceived Difficulty	45
5.10	Participants Experience and Mood during the Relevance Assessment	46
5.11	Concentration and Stress under Time Pressure	47
5.12	Participants Perceived Accuracy	47
5.13	General Analysis of Post Task Questionnaires	48
5.14	Comparison with Former Studies	48
5.15	Analysis of the four classes of document types	50
6	Conclusion	51
	References	61
	APPENDICES	61
A	Forms	62
A.1	Ethics	62
A.2	Recruitment Email	72
A.3	Information and Consent Forms	74
A.4	Demographic information form	77
A.5	Tutorial and Instructions	79
A.6	Questionnaire	81
A.7	Feedback Letter	85

B	User Interfaces	86
B.1	Example of Topic descriptions	86
B.2	Example of a document and Information for users	87
B.3	Example of a Feedback for a Judgment	88
B.4	Example of a Questionnaire	89
C	Codes	91
C.1	Latin Square	91
C.2	Stratified Sampling from the RRF lists	94

List of Tables

3.1	Topics used in the study	15
4.1	Treatments used in the study to judge 20 documents	22
4.2	6*6 Latin Square used in the study	24
5.1	Confusion matrix	26
5.2	Average accuracy. F stands for Full documents and S for Summaries or document excerpts.	27
5.3	Estimated true positive rates for different treatments of the study	31
5.4	Estimated false positive rates for different treatments of the study	31
5.5	Average Time in seconds	34
5.6	Average accuracy and time spent on judging for different time constraints	34
5.7	Number of times participants exceeded from their time limit.	36
5.8	Average time participants spent after their time limit passed and the document became hidden.	36
5.9	Average estimated ability to discriminate (d')	37
5.10	Assessors Criterion	37
5.11	Average accuracy within each topic and time constraint	44
5.12	Average accuracy within each topic and document length	44
5.13	Participants perceived difficulty and average accuracy.	45
5.14	Participants perceived difficulty and average time.	46
5.15	Participants Experience and Mood during the Relevance Assessment.	46

5.16	Participants Concentration and Stress under Time Pressure.	47
5.17	Participants perceived accuracy and actual average accuracy.	47
5.18	Post task questionnaire analysis. Sections A.6 displays the details of the questions.	48
5.19	The comparison between a former study with the same topics and the current thesis.	49
5.20	The comparison between five topics of a former study and the current thesis.	49
5.21	Analysis of the 4 classes of document types	50

List of Figures

3.1	Topic 310	16
3.2	Topic 336	16
3.3	Topic 362	17
3.4	Topic 367	17
3.5	Topic 426	17
3.6	Topic 427	18
3.7	Topic 436	18
4.1	Main task user interface	23
5.1	Average accuracy for all the participants	28
5.2	Participant's Accuracy for a given Treatment. F stands for Full documents and S for Summaries or document excerpts.	29
5.3	True positive rates for all the 6 treatments. F stands for Full documents and S for Summaries or document excerpts.	32
5.4	False positive rates for all the 6 treatments. F stands for Full documents and S for Summaries or document excerpts.	33
5.5	Participant's average time for a given Treatment. F stands for Full documents and S for Summaries or document excerpts.	35
5.6	Participant's estimated ability to discriminate (d') for a given Treatment. F stands for Full documents and S for Summaries or document excerpts.	38
5.7	Assessor Criterion for all the participants.	39

5.8	Average Time vs. True Positive Rate for all the participants. The orange dot is the overall mean.	40
5.9	Average Time vs. False Positive Rate for all the participants. The orange dot is the overall mean.	41
5.10	Average Time vs. d' for all the participants. The orange dot is the overall mean.	42
5.11	Average Time vs. Criterion for all the participants. The orange dot is the overall mean.	43

Chapter 1

Introduction

Information Retrieval and search engines are an important part of our daily life. For any information need that we have, we use a search engine to perform a query and look for the answer in a list of ranked documents. Due to this importance of information retrieval systems, researchers and companies are constantly trying to improve their IR systems. This development cannot be done without evaluating new ranking algorithms with different corpora of already judged documents. Even existing systems should constantly get evaluated with new sets of documents for improvements.

The term Relevance plays a big role in any IR driven research. This is due to the fact that the ultimate goal of any IR system is to retrieve the most relevant documents that satisfy the information need of the users. Almost all work in the area of Information Retrieval is somehow related and linked to the relevance of documents and texts. However, providing users and searchers with the most relevant documents is a complex task and it gets more complicated every day as the new technologies evolve and the volume of data grows.

While there are many applications for information retrieval systems, we were most interested in tasks such as high-recall retrieval where high volume of relevant documents should be retrieved with acceptable cost and effort. This requires many items and documents to be judged for relevance given a topic statement. High-recall retrieval is important in tasks such as e-discovery, systems review, and test collection construction. The aim of this work is to study if we can speed up the judging process while maintaining the quality of judging.

There are mainly two avenues of research in IR. First, the system-driven research which focuses on algorithms and fundamentals of storing and retrieving data in an efficient way

and using different techniques such as matching words in both queries and documents to retrieve relevant information. Second is the user-oriented research which emphasizes on users' decision making and the cognitive processes they go through when judging the relevance of retrieved documents. This approach focuses more on the behavior of the users and tries to understand their information need better. The goal in this type of research is to study user-oriented view of relevance and their evaluation of the usefulness of the retrieved documents according to their information need. System-driven research is useful and easier but it does not include everything related and no system has been shown to be perfect. This has led to increase in the number of user-oriented research. However, user-oriented research is much more expensive. It is extremely hard to control the environment for this type of research. There are many aspects that should be taken to account such as the number of actual users, proper user interface, learning effects of the users and so on (Jones, 1997). While on the other hand, laboratory system-driven research is much less expensive since controlling different variables are way easier (Voorhees, 2001).

Both groups of IR researchers take advantage of the results of the system-driven and user-oriented approaches to develop search systems that better address users information needs. In this thesis, we mainly focus on the user-oriented approach.

1.1 Background Overview

1.1.1 IR Evaluation and Test Collections

While IR systems improve as new technologies emerge, we still need a way to determine if an invented algorithm is better than the others. Moreover, current systems need to be optimized and checked over new metrics. Cyril W. Cleverdon started a series of projects called the Cranfield Project which took about 10 years. The project goal was to provide a foundation for IR evaluations. According to Cranfield paradigm, in order to evaluate an IR system, researchers need test collections and evaluation metrics. A test collection consists of three main components including the documents set, topics set, and relevance assessments set (Baeza-Yates et al., 1999, Baeza-Yates et al., 1999). Documents get retrieved by running each topic against the set of documents and then top-k documents get pooled to be judged by the assessors. After having the relevance assessments, evaluation metrics can be used to compare IR systems, search algorithms, and different search strategies (Sanderson et al., 2010, Voorhees et al., 2005; Sanderson, 2010). The first pioneering Cranfield collection only contained 1398 documents and 250 topics while more recent test collections such as the ones used in TREC Ad Hoc consist of millions of documents and hundreds of topics.

The Text REtrieval Conference (TREC) is an annual workshop funded by the National Institute of Standards and Technology (NIST) and U.S. Department of Defence which started in 1992 and its purpose is to support the research in information retrieval by providing tools and infrastructure for large-scale IR evaluation. The number of documents in test collections will continue to increase dramatically as large text searches are required for better evaluating current and future IR systems.

1.1.2 Evaluation Metrics

There is a large number of evaluation metrics that researchers use in Cranfield paradigm to evaluate IR systems after running them through test collections. Metrics such as precision (the fraction of documents retrieved that are relevant) and recall (the fraction of relevant documents retrieved) are two of the most important IR evaluation metrics. But here in relevance assessment, we focus on already retrieved documents and how users find them relevant or non-relevant. We define accuracy based on what we call true positive rates and false positive rates which represent if the users agree or disagree with the ground truth judgments. We explain in detail the metrics and measures we used in later chapters.

1.1.3 Differences in Primary and Secondary Assessors

One of the main problems in Cranfield paradigm is what so-called incompleteness in relevance assessment. Since relevance is subjective, different assessors have different opinions which lead to non-consistent test collections. This has been a problem since the first Cranfield experiments in 1960s ([Harter, 1996](#); [Taube, 1965](#)). Primary assessors who originate the search topics and can be any kind of users such as experts, people who work in e-discovery or researchers from other fields, may have different opinions about assessing specific documents than secondary assessors. Secondary assessors are individuals who get hired to judge the relevance of documents to specific topics. They may have different backgrounds and they do the relevance assessment just based on the descriptions of the topics and the guidance given to them. In fact, in a study performed by NIST on TREC-4 collection, there was less than 50% agreement between primary and secondary assessors ([Voorhees, 2001](#)). While primary assessors may have a better background and a good level of knowledge of the search topics, they still make mistakes in assessing the relevance of documents ([Grossman and Cormack, 2011](#), Grossman and Cormack, 2011). This is true for secondary assessors as well since they might even have less background knowledge of the search topics and they might not fully understand the topics given to them ([Kinney et al., 2008](#), Kinney et al., 2008).

An important point to keep in mind in relevance assessment behavior studies is that many aspects can affect the way people judge documents. Cuadra (1967) described 38 parameters in 5 categories which affect relevance judgments of documents performed by different people. For example, we know users background knowledge, previous experiences, topics and documents difficulty levels, documents length and so on, will affect relevance assessment.

The subjectivity of relevance assessment and the disagreement between primary and secondary assessors has been an open research problem over the past years but, in this thesis, we focus on secondary assessors since they play an important role in producing test collections for IR evaluations and we are interested in their behavior in terms of speed and accuracy. Also, by studying the behavior of users performing search tasks, we can learn and predict their behavior in real world situations. This way we can evaluate and develop better IR systems. Assessors also have similar behavior when judging various documents. By controlling effective variables on assessors behavior, we can develop more errorless systems.

1.2 Problem Statement

With the advance in technology and the significant increase in the number of documents produced and test collections sizes, the process of making relevance judgments has become more expensive and time-consuming. So the need for recruiting assessors and making relevance judgments in an effective way has been increased. Therefore, more researchers have become interested in studying the effects of different variables on assessors relevance judging behavior. Furthermore, since not all the assessors have the expertise in all the topic domains, creating a better judging environment can help them to have a better performance and accuracy in their judgments.

Studying the relevance behavior can also help in designing better systems for collecting relevance assessments, therefore, creating better test collections and ground truths that can help IR researchers conducting more accurate studies.

Relevance assessment is also important in e-discovery where assessors should be recruited to find responsive or relevant documents to the production requests. Production request is the legal term in e-discovery for search topics.

Due to the importance of studying the relevance judging behavior, we believe that more qualitative research should be done to better understand this behavior. We designed an experiment along with questionnaires to study this behavior better, especially the effects of

applying time constraints and showing document excerpts instead of full documents. Our questionnaires also helped us to study other aspects such as assessors level of confidence while judging, their ability to discriminate between relevant and non-relevant documents, their perception of accuracy and difficulty in relevance judging, and in general what were some of the variables that could affect their performance in judging. We mainly aimed for answering these questions that can we force people to work faster but with the same performance by applying time constraints? Can users have the same performance if we show them shorter document excerpts instead of full documents?

To the best of our knowledge, there has not been enough research to study the trade-off between speed and accuracy in relevance judging behavior. Also, studying the effects of using document excerpts instead of full documents can lead to interesting findings regarding relevance assessment. All these different variables can help IR researchers to better understand the relevance behavior and design better systems.

1.3 Thesis Overview

In this thesis, we conducted a user study to find the effects of variables such as time constraint and document excerpts on relevance assessment. Our main focus was to find if using time limitations and also shorter document excerpts instead of full documents will lead to better and faster relevance assessment with less cost and effort.

While chapter 2 represents the related work in relevance assessment area, we discuss in chapter 3 the data set we used for the user study and how we chose our topics and documents.

Since there are lots of aspects that can affect the relevance assessment process, chapter 4 explains in detail our protocols and how we designed the study to be completely balanced so that we have less biased data. We also explain the different phases of our study and the user interfaces we designed.

In chapter 5, we represent our results and findings based on all the users' data that we collected. This includes the analysis of accuracy, analysis of time, analysis of assessor criterion and ability to discriminate and the difference in results when using document excerpts instead of full documents.

Finally, in chapter 6, we discuss our conclusions and future directions.

1.4 Contributions

We make the following contributions in this thesis:

- We show that assessors' quality of relevance judging when shown document excerpts, is not only as good as when shown full documents but also in most of the cases better. We had three different time constraints that we applied when assessors had to judge the relevance of the documents. In all three cases, assessors performed better when shown document excerpts instead of full documents however, the difference was not statically significant.
- We show that giving assessors more time to judge the documents leads to slightly but not a significant increase in their accuracy and performance. The average time spent on judging the documents in this study was 12.8 seconds.
- Assessors average time spent on judging increases when given more time but this does not lead to significant change in accuracy and performance.
- The previous contributions lead to a more general statement that we can speed up the judging process by using document excerpts and time limitations while maintaining the quality of judging.
- We show that assessors ability to discriminate between relevant and non-relevant documents increases when given more time.
- We show that assessors criterion does not change within different time constraints. Our participants showed liberal behavior when judging documents meaning that they preferred to commit false positive mistakes to avoid missing relevant documents.
- We see that there is no correlation between assessors actual accuracy and their perceived difficulty of the judging tasks.

Chapter 2

Literature Review

Many studies have been conducted to study users relevance assessment behavior. Being aware of users behavior helps in designing and evaluating information retrieval systems better. Prior knowledge of their behavior also can help researchers and IR system designers to predict users behavior and their search performance on new IR systems.

2.1 The Definition of Relevance

The very first issue that researchers face when conducting this type of research is how to define relevance? What is the concept of relevance in different search tasks? There has been considerable amount of research trying to address this problem in details (([Froehlich, 1994](#)); ([Borlund, 2003](#))); ([Green, 1995](#)); ([Harter, 1992](#)); ([Schamber et al., 1990](#)); ([Cosijn and Ingwersen, 2000](#)); ([Mizzaro, 1998](#)); ([Park, 1994](#)); ([Saracevic, 1996](#)); ([Croft et al., 2010](#))). [Cooper \(1971\)](#) defined relevance logically as the relation between portions of stored information and information need. Part of the stored information that satisfies the users information need gets classified as relevant and the rest non-relevant. This can be done by logically answering questions that divide sentences or statements to redundant and non-redundant information. However, relevance is not limited to topical relevance (([Borlund, 2003](#))). [Wilson \(1973\)](#) categorized relevance meanings and defined them as evidential and situational. According to him, users' personal knowledge and situation affect the concept of relevance for them. A number of aspects should be taken to account when defining relevance. Situational relevance is user-based and more practical for interactive information retrieval (([Borlund, 2000](#)); ([Borlund and Ingwersen, 1997](#))). In situational relevance, there can be a document that perfectly matches the words in users query but

still not relevant. There can be reasons such as non-reliable source, outdated document or previous knowledge of the user. While much research has been done to fully understand the concept of relevance, the problem is still open. (Borlund, 2003) classifies relevance to levels and classes. She also states that relevance is dynamic and dependent on the searching process. Kuhlthau (1991) also mentions that the concept of relevance can be dynamic and varies during the information seeking process. Büttcher et al. (2016) mentions that a document is relevant to a topic only if its content satisfies the information need of the user and its query.

However, when it comes to practice, finding a relevant document is much more complex. Buchanan and Loizides (2007) found that users have a first impression of the document's relevance when they first see it and if they spend more time on that document they are looking for confirming their first impression. Users mostly pay attention to titles, snippets, and abstracts and mostly less attention to the main context of the document ((Cool et al., 1993), (Saracevic, 1969)). Barry and Schamber (1998) studied user's relevance judgment behavior by focusing on relevance values and criteria. Wang and Soergel defined a cognitive model of users' decision making in selecting relevant documents (Wang and Soergel, 2010).

Cuadra (1967) defined various variables which affect users relevance judgment abilities and decision makings. These variables can be related to the document itself, the users information need, users personal situation, knowledge, and judgment conditions. For example, contents diversity, length of the documents, the volume of the information, documents difficulty level, textual attributes, the order of presentations, users background knowledge and so on can affect the concept of relevance. Cool et al. (1993) conducted two user studies to find all the factors that affect users relevance assessment. In their first study, they recruited university students to do search tasks and fill questionnaires about the relevance of documents to find the factors that impact their decisions. In the second study, they asked scholars in humanities to search for relevant documents. Then interviews were done to understand the factors that affected their judgments. They found several factors including topic, content, format, presentation, values, and oneself. Schamber (1994) also categorized 80 factors of relevance assessment to six groups of judges, requests, documents, information system, judgments condition, and choice of scale. Barry (1994) conducted a user study by asking users in an academic environment to evaluate documents and mark the portions that affected their decisions the most. Users were also interviewed to discuss those portions in detail. He reported 23 factors and seven categories such as subjective qualities (agreement with the content/information provided) and situational factors (e.g, time constraints). Eisenberg and Barry (1988) also found that the order of presentation affects assessors relevance judgment. They reported that when ranks were sorted from high to low, top highly relevant documents were underestimated and when sorted low to

high, low and medium range documents were overrated. [Ruthven et al. \(2007\)](#) found that assessors level of knowledge of the search topics, their interest in the search topics, and their confidence level all are factors that have an impact on relevance assessment. ([Taylor et al. \(2007\)](#), [Spink et al. \(1998\)](#); [Tang and Solomon \(1998\)](#); [Vakkari \(2000\)](#); [Vakkari and Hakala \(2000\)](#); [Wang and White \(1999\)](#)) all concluded that users relevance assessment varies during the information seeking process.

[Xu and Chen \(2006\)](#) found that topicality and novelty are critical factors in relevance assessment while understandability, reliability, and scope have some importance. [Bales and Wang \(2005\)](#) did a research on 16 empirical studies and reported 230 factors under four categories of situations and context, user criteria, document information elements and value judgments.

A number of researchers have studied the effects of different types of assessors on relevance judgments. [Bailey et al. \(2008\)](#) recruited three types of participants for three different types of assessors. Assessors of type gold standard were both originators of the topics and IR experts. Silver standard assessors were not originators of the topics but experts in IR. The third group was bronze standard which did not originate the topics and were not experts in IR. According to their study, there was a significant difference in judgments results from all 3 assessors types and that, the level of agreement between these 3 groups was significantly low. Lesk and Salton studied how information retrieval systems rankings change by differences in relevance assessment ([Lesk and Salton, 1968](#)). They studied the judgments from topic authors and non-authors and they found that while there was 30% agreement between them, they still produced the same rankings. [Harter \(1996\)](#) also studied the relation between relevance assessment and retrieval effectiveness. He concluded that in all empirical and existing literature there are variations in relevance assessment but this variation does not affect the measures of retrieval effectiveness which again means they lead to same rankings.

Overall, the concept of relevance plays an important role when studying the accuracy of assessors and it is important to this thesis. Therefore, we will explain exactly what we mean by a relevant document in the remaining chapters to understand the meaning of accuracy better.

2.2 Precision and Uncertainty in Relevance Assessment

There has been much work done to study other aspects of users cognition process while doing relevance assessment. [Smucker and Jethani \(2010\)](#) studied the effect of precision on assessors. They found that high precision result list leads to a lower rate of true positive (the fraction of relevant documents judged as relevant). They also captured the data regarding the mood of the participants along with various other parameters. For example, from this data, they concluded that users find the judging task less difficult when they were given a high precision rank list and that they were also more concentrated during working with high precision result lists. In a similar study, [Smucker and Jethani \(2011b\)](#) compared the behavior of NIST assessors and the participants. They found that while NIST assessors had better rates of false positive (fraction of non-relevant documents judged as relevant), they had the same true positive rates as the participants in the study.

Aiman Al Harbi and Mark D. Smucker([ALHARBI, 2016](#)),([Al-Harbi and Smucker, 2014](#))) conducted a think a loud user study to better understand the relevance judging behavior of secondary assessors and to capture the uncertainty in secondary assessors' relevance judgments. They found that the level of certainty differs from low to high in relevance judgments. The interactions between the search topic, the secondary assessor, and the document being judged can result in these differences. They also showed that the amount of detail stated in a search topic's description can affect the relevance judging behavior. They report the reasons in more detail in ([Al-Harbi and Smucker, 2013](#)) and recommend improvements in judging interfaces to capture assessors' certainty besides their judgments.

2.3 Disagreement in Relevance Assessment

Disagreement among the assessors in relevance judgments is another important issue since it can directly affect effectiveness measures in IR systems. There is no doubt that this disagreement exist among the judges and it has been reported in many studies ([Gull, 1956](#)); ([Barhydt, 1964](#)), ([Barhydt, 1967](#)); ([Rees and Schultz, 1967](#)); ([O'Connor, 1969](#)). Not everyone agrees on the relevance of documents to a specific topic. There has been lots of research to study this issue and its reasons. The question is if this can affect the validity of the results and evaluations of IR systems.

Barhydt (1967) compared two groups of subject experts and system specialists as secondary assessors and did not find any correlation between the types of groups and their performance.

Janes and McKinney (1992) found that there is no significant difference between primary assessors (topic originators) and secondary assessors (non-originators of the topic) in terms of judgments and performances. In another study, Janes (1994) also recruited 48 assessors from the school of information/library and found that secondary assessors can produce comparable judgments to primary assessors. In a user study conducted by Al-Maskari et al. (2008) a high agreement of 63% were found among the participants (secondary assessors) and TREC assessors. Some topics were reported to be difficult by the participants.

However, Chouldechova and Mease (2013) performed a user study for search tasks and reported that primary assessors, users who originated the queries, performed better than secondary assessors.

In a study by Efthimiadis and Hotchkiss (2008), assessors without legal expertise performed better than legal experts in judging legal documents.

Grossman and Cormack (2011) studied if the disagreement in relevance assessment in legal documents is due to ambiguity or inconsistency in applying the criteria for responsiveness to particular documents or due to human error. They used the TREC Legal Track 2009 corpus and the assessments by law students and topic authorities. They randomly selected a sample of documents that the topic authorities and other assessors did not agree on. While (Grossman) was the topic authority for one of the topics, (Cormack) did a re-assessment and he reported clear disagreement for one document and arguable disagreement for nine other documents. Then (Grossman) the topic authority, did a re-assessment and for five out of ten documents changed her opinion for assessment. They did similar experiments for other topics as well and ended up with similar results. They concluded the vast majority of cases of disagreement are a product of human error rather than documents that fall in some gray area of responsiveness. Scholer et al. (2011) also studied the inconsistency in relevance judgment made by TREC assessors. They studied assessors relevance judgments when judging duplicate documents and found that when the number of documents between duplicates is less, assessors are more consistent in their judgments. They also report that assessors tend to judge the relevance of the next document the same as the preceding document which they call it judgment inertia. In another study, Scholer et al. (2011) reports that during the relevance assessment process, same assessors might produce different judgments at different times and this depends on using binary-relevance or three-level scale. In another study by Roegiest, Adam (2017), he concludes that "when

evaluating the results of a classifier trained by one assessor using an independent set of assessments, a large part of what is measured is a difference of opinion rather than a magnification of error”.

[Chandar et al. \(2013\)](#) also found interesting results regarding assessors disagreements. They ended up with three conclusions: “longer documents will provoke more disagreement, less coherent documents will provoke more disagreement, and documents that are easier to read are the ones that provoke more assessors disagreement”. While their initial hypothesis was that harder documents to read will end up in more disagreements, this hypothesis was found to be wrong.

In a recent study, [Wakeling et al. \(2016\)](#) tried to compare the primary and secondary relevance judgments using what they called real-world topics. The topics represented the information need of the staff and students at the university which they recruited the participants from. They found that the secondary assessors find the relevance task more challenging and less interesting. They also took longer time, to judge the documents but overall, the agreement between primary and secondary assessors reported to be high. They also found that topic descriptions were sometimes confusing for the secondary assessors and that was the reason they took longer to assess some documents. The unfamiliar vocabulary was reported to be one of the main reasons for confusion. The agreement percentage between the primary and secondary assessors was 79%. Primary assessors had higher interest and more knowledge of the search topics. The level of confidence was also reported higher for primary assessors since they were the originator of the search topics.

[Villa and Halvey \(2013\)](#) found that assessors pay more attention to the documents with greater length. Based on their results, highly relevant or non-relevant documents needed less effort for judging. They also concluded that the length of the documents does not impact the accuracy but the degree of relevance of the documents (highly relevant or non-relevant), impacts the level of accuracy.

2.4 Crowdsourcing in Relevance Assessments

With the emerge of crowdsourcing technology in the last few years, lots of research has been done to study relevance judgments made by crowdsourced workers. There have been papers to verify the reliability of the test collections and relevance judgments made by crowdsourcing. [Clough et al. \(2013\)](#) reported that there was a delicate difference between the relevance judgments by the crowdsourced workers and the results from the two search engines that they used in their study. However, in order to have reliable relevance judgments from crowdsourcing, special techniques should be applied ([Alonso et al., 2008](#)).

Alonso and Mizzaro (2009) achieved highly reliable results by careful design of their experiment. In a similar study, Alonso and Mizzaro (2012) discuss that crowdsourced workers can be reliable and much less expensive only if we design the experiments carefully and control its execution and quality. However, in another study, Alonso and Baeza-Yates (2011) argue that due to different conditions, producing reliable relevance judgments with crowdsourcing can be challenging and different groups of HCI, engineering, and IR should work together to overcome these challenges since the crowdsourcing technology has the potential and it is much less expensive than traditional in-lab user studies. Snow et al. (2008) with the use of Amazon MechanicalTurk, collected annotations on natural language tasks and reported that they had high agreement with gold standard annotations. Eickhoff et al. (2012) divided the crowdsourced workers into money-driven assessors and entertainment-driven assessors and argued that if we could attract entertainment-driven assessors, we will end up with more reliable results. They also propose their technique for designing crowdsourced experiments better.

2.5 Speed and Accuracy in Relevance Assessment

In a recent study by Maddalena et al. (2016), they report how the time available to judge documents, affect the quality of judging. Based on their results, it is possible to reduce the cost of crowdsourced evaluation by applying time constraints and reducing the time available to make judgments with no loss of quality. They also report that "top judgment quality is obtained with 25-30 seconds to judge a topic-document pair."

We preferred in-lab user study in this work since we wanted to be able to monitor the users carefully and study their cognition process and decision makings with emphasize on their speed and accuracy.

Wang and Soergel (2010) performed a user study on different parameters in relevance assessment including agreements, speed, and accuracy. They recruited four law students and four students from library and information studies (LIS) to judge the relevance of 100 documents. The study consisted of relevance task, entry questionnaire, and exit interview. They found that law students judged the relevant documents as accurately as LIS students while LIS students judged non-relevant documents slightly more accurately than law students. Moderate to a substantial agreement was reported between the two groups of students. There was no significant difference in the speed of the two groups of assessors but the speed varied between the individuals. They also concluded that background knowledge is not needed for assessors if they are given accurate guidelines to judge the documents. However, their study was limited by the number of participants and the topics they used.

In a similar study, Wang (2011) studied the accuracy, agreement, speed and perceived difficulty in relevance judgments for E-discovery. Again, four law students and four students from library and information studies were recruited to judge the e-discovery test collection of TREC. All the participants had perceived the difficulty of the judgments differently. Participants also agreed with TREC assessors more than 70% of the time. A number of influencing factors were reported such as document subjects, length, and legibility, assessors background knowledge and reading skills, relevance guidelines, and learning effects. While relevance judgments speed varied among the participants, some of the documents took noticeably longer time to judge. Based on the results: Strong correlations were observed between relevance judgment speed and perceived difficulty and between perceived difficulty and relevance judgment accuracy, but not between relevance judgment accuracy and speed. That means that if an assessor finds a document easy to read, he or she can judge that document faster and more accurate. Participants showed various different relevance judgment speeds (ranging from 13 to 83 documents per hour). Assessors judged non-relevant documents about as fast as relevant documents which was an interesting finding since it was against the traditional thought that assessors can judge relevant documents much more faster because it is easier to find the relevant piece of text in a document and there is no need to read the rest. Regarding the individual speeds, only small fraction of documents slowed down the assessors. The speed also varied significantly among different topics. Slightly positive correlation between speed and accuracy was reported for one topic, negative correlation for another, and no correlation for two other topics. Therefore, the author mentions that there was no detectable correlation between speed and accuracy for the participants in this study.

However, these previous studies had limitations such as the number of participants and number of documents and topics. In this thesis, we intend to focus on the speed and accuracy of the participants in relevance judgments closely and also report other aspects of users cognition progress while doing search tasks such as relevance assessment.

Chapter 3

Data Set

3.1 Topic Selection

In total, we used seven topics from the New York Times annotated corpus (Sandhaus, 2008). Six for the main task, and one for the tutorial phase. Table 3.1 shows the topics that we used and their titles. The reason that we chose these specific topics was that we had a good insight into them. Lots of data had been collected based on these topics in the previous studies. Every search topic comes with a title, a description, and a narrative. Description defines what document can be relevant to the search topic. A narrative is just a more detailed description of what information to look for in a relevant document. For our study, we combined the description and narrative not to make participants confused. We call the combined description and narrative "Search Topic's Description".

Topic	Title	Task
310	Radio Waves and Brain Cancer	Main
336	Black Bear Attacks	Main
362	Human smuggling	Main
367	Piracy	Tutorial
426	Law enforcement, dogs	Main
427	UV damage, eyes	Main
436	Railway accidents	Main

Table 3.1: Topics used in the study

3.2 Topic 310: Radio Waves and Brain Cancer

This topic is about Radio Waves and Brain Cancer. Figure 3.1 shows the topics description that assessors should read in order to be able to judge the relevance of the documents. In order to assess the documents regarding this topic accurately, a good knowledge of electrical radio waves and a good level of reading comprehension is required since users have to find the part of the information that relates radio waves and brain cancer together. Assessors should look for any articles or studies which report on the incidence of brain cancer being as a cause of radio waves.

Relevant documents will reflect evidence for or against the notion that radio waves from radio towers or car (mobile/cell) phones can cause or affect brain cancer occurrence.

Figure 3.1: Topic 310

3.3 Topic 336: Black Bear Attacks

This topic is about Black Bear Attacks. The description of this topic is presented in figure 3.2. A relevant document would mention any worldwide attack by black bears, its cause, and any attempts by wildlife officials to control these attacks. Assessors should decide whether the document addresses the information need described by the description or not.

A relevant document would discuss black bear attacks, the frequency of black bear attacks worldwide, and the possible causes for these attacks.

Figure 3.2: Topic 336

3.4 Human Smuggling

Topic 362 is about human smuggling. Figure 3.3 shows the description of this topic. The description of this topic seems short and easy. However, assessors should pay attention to the information need here since there is explicit information about what is it going to be considered as human smuggling.

A relevant document shows an incident of humans being smuggled. Relevant incidents should involve 10 or more people.

Figure 3.3: Topic 362

3.5 Topic 367: Piracy

Figure 3.4 shows the description of topic 367, piracy. Five documents regarding this topic were used for the tutorial phase. We picked this topic for the tutorial phase randomly out of seven selected topics for the study. Assessors should pay attention to the fact that not all the documents about piracy are relevant here. Only instances of piracy in the sea are considered to be relevant for this topic.

What modern instances have there been of old fashioned piracy, the boarding or taking control of boats? Documents discussing piracy on any body of water are relevant.

Figure 3.4: Topic 367

3.6 Topic 426: Law enforcements, Dogs

As shown in figure 3.5, topic 426 is about law enforcements and dogs. The description of this topic is simple and clear. Any document about training or using dogs in law enforcements worldwide is considered as relevant to this topic.

Relevant documents discuss the use of dogs worldwide for law enforcement purposes.

Figure 3.5: Topic 426

3.7 Topic 427: UV damage, eyes

Figure 3.6 shows the description of topic 427, UV damage and eyes. This topic can be confusing when the documents talk about any other kind of radiation damages for eyes.

Only UV damages caused by the sun or light, treatments, and educational programs or preventions are considered relevant to this topic. Assessors should have good reading comprehension skill to link the source of damage to the eyes correctly.

Relevant documents discuss the damaging ultraviolet (UV) light from the sun can do to eyes. Relevant topics include 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.

Figure 3.6: Topic 427

3.8 Topic 436: Railway Accidents

Topic 436 is about railway accidents. Figure 3.7 presents the description of this topic. Documents regarding any source of accidents in railways are considered as relevant. Assessors should be careful while judging documents about this topic since talking about railways in general, accident preventions and other safety technologies are not relevant. The description is concise and provides enough information about relevant documents.

Relevant documents discuss the causes of railway accidents throughout the world. Railway accidents include locomotives, trolleys, streetcars, etc. where either the railway or another vehicle or person caused the accident.

Figure 3.7: Topic 436

3.9 Document Selection

In order to select the best documents for our study, we used Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) which is a method for combining the document rankings from multiple IR systems. This method has shown better results than individual IR systems. RRF combines the results of IR systems by assigning scores to documents based on the

majority of votes from different systems and reorders the documents result. This means that a higher ranked document is less vague and there is more confidence in knowing that the document is relevant or not relevant to the search topic. A higher RRF score means that there is more general agreement on the relevance of the document to the search topic among actual IR systems. For calculating the RRF scores, we used the parameters given by [Cormack et al. \(2009\)](#). We used all the runs submitted to TREC 2017 Core Track. After creating the RRF list, We did a stratified sampling by creating four lists of higher rrf relevant, lower rrf relevant from the list of NIST relevant documents and higher rrf non-relevant, lower rrf non-relevant from the list of NIST non relevant documents. In order to create these four groups, we took the rrf rank lists for relevant and non-relevant documents and we splitted each in half to get higher and lower rrf scores. Then with having these four subgroups or strata we did a simple random sample and chose five documents from each to get 20 documents. For each topic, we sampled 20 (10 relevant and 10 non-relevant) documents. Overall, we sampled 120 documents for the six topics of the main task and five documents for the tutorial phase. Documents from higher rrf relevant and lower rrf non-relevant tend to be easier to judge and documents from higher rrf non-relevant and lower rrf relevant tend to be harder to judge.

3.10 Using Document Excerpts Instead of Full Docs

For three out of the six main tasks, we showed the users document excerpts or paragraph-long summaries instead of full documents. This decision was made based on a study by [Zhang et al. \(2018\)](#). In that study, they used a summary selection algorithm to generate one-sentence long summaries. In another study by [ZHANG et al. \(2017\)](#), the authors modified the method to generate paragraph-long summaries instead of one-sentence summaries. They first indexed all paragraphs from all the documents by calculating tf-idf features for each paragraph. They used the same document frequency corresponding to full documents instead of counting each paragraph separately. For the original study they randomly chose 100 documents and labeled them as non-relevant. These documents were then passed to the classifier to be trained on. For our study, we trained the classifier based on NIST judgements. The classifier then ranks all the paragraphs and selects the highest ranking paragraph and its corresponding document.

In the study by [ZHANG et al. \(2017\)](#), the authors tested that if short document excerpts can reduce assessment time and effort for high-recall retrieval. They conducted a controlled user study with 10 participants. They designed a high-recall retrieval system using continuous active learning (CAL). They either displayed full documents or short

document excerpts to the participants and they found that participants were able to find more relevant documents when they were shown document excerpts as opposed to full documents. In this thesis, we wanted to further study this hypothesis that not only assessors are able to find more relevant documents when shown document excerpts, but also their overall judging performance would be higher. Based on the data that we collected we were able to validate this hypothesis and we think more research should be done regarding this matter.

Chapter 4

User Study

4.1 Study Design

The study consisted of the tutorial phase and the main task. The tutorial phase helped participants to get familiar with judging the relevance of documents and understand the relevance concepts. For the main task, participants had to judge the documents under different time constraints.

4.1.1 Study Protocol

The study was held in a private lab at the University of Waterloo. Participants were provided with an information letter and a consent form before starting the study. The information letter explained all the rules and information that participants needed to know prior to the study. Only after signing the consent form, they were able to continue and login to the system. After logging in, they were provided with information and instructions on how to perform the study. After reading the instructions, participants had to pass a simple quiz of the rules of the study. They also had to fill the demographic information form so that we could gain data about their previous experiences with IR systems. The questions included their age, English fluency, academic background, education and if they had received previous training in information retrieval or not. After filling the demographic information form and passing the quiz, they could start the training phase and then the actual task phase.

We collected all the users relevance judgments and time spent to judge each document.

4.1.2 Tutorial

The aim of the tutorial was to train the participants on the document judging process so that they produce more reliable data and feel confident doing the main task. All the settings and user interfaces for this tutorial was the same as the main task. Five documents of the same topic were given to each participant and they were required to make a judgment for each. 3 of the documents were relevant and two non-relevant. Each participant saw the same order of the documents. After each judgment by the users, they were given a feedback and an explanation regarding their choice. We collected their judgments and time spent on each document. Participants were told that they should ask questions if they needed any clarification.

4.1.3 Main Task

For the main task, participants judged 120 documents from six topics that were different from the topic used in the tutorial phase. Six different treatments were provided for each participant. Two sets had 15 seconds time constraint, two had 30 seconds, and two 60 seconds per document. Three sets had full documents and three had document excerpts. Each participant judged 20 documents of a topic in each set with a different treatment. Participants were provided with a pre-questionnaire about their familiarity with the topic before each set. After each set, they were asked to fill a post-questionnaire about their experience doing the previous task. The precision of each set was 0.5 (10 relevant and 10 non-relevant documents). This decision was made due to the previous result from the study by [Smucker and Jethani \(2010\)](#) where they found that users perform the best with 0.5 precision for the document sets. Each participant saw a randomized order of the topics and the documents. Table 4.1 shows all the different treatments used for each set of 20 documents.

Time Constraint(s)	Document Length
15	Full
30	Full
60	Full
15	Excerpt
30	Excerpt
60	Excerpt

Table 4.1: Treatments used in the study to judge 20 documents

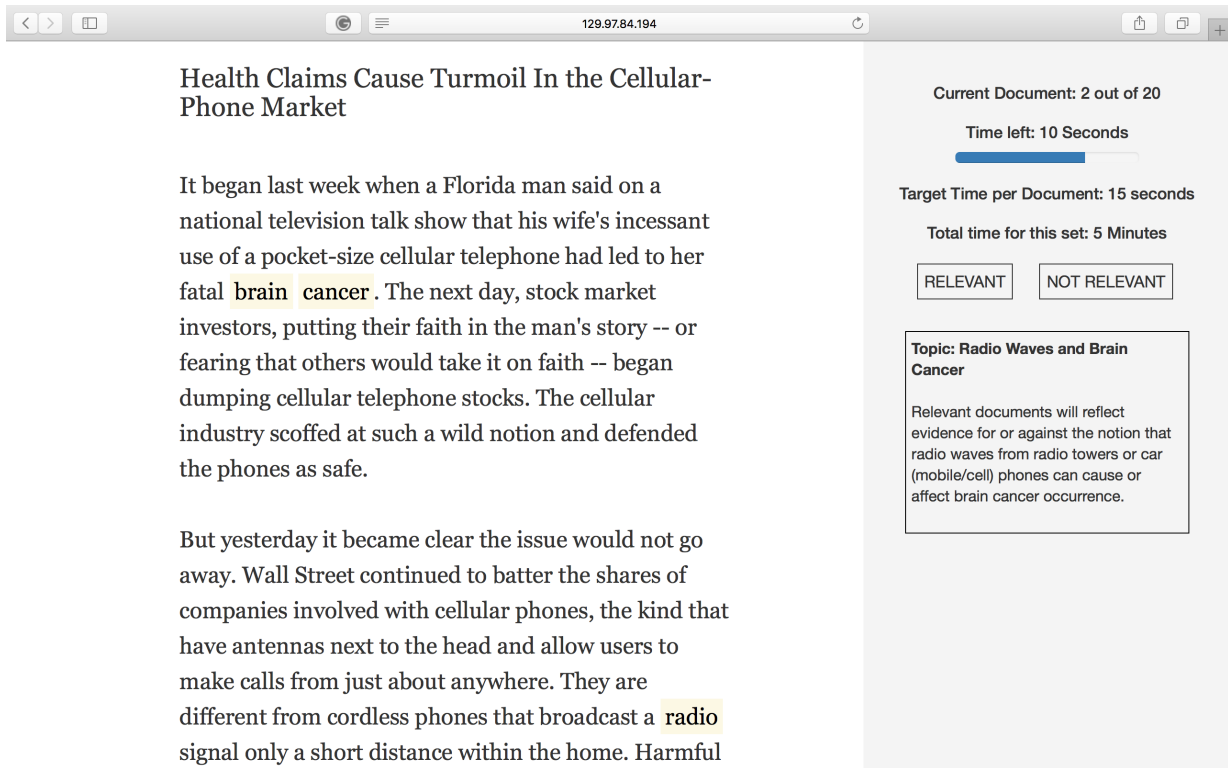


Figure 4.1: Main task user interface

4.2 User Interface

We designed the user interface to show the participants one document at time. Participants could see the title of the document and either the full document or a short document excerpt. They only could make binary judgments and could click on relevant or non-relevant buttons. This is the most standard way to collect relevance judgments. Through the whole study, participants could see the topic and its description for that set so that they do not forget the search topic. Participants were also provided with information about the number of documents and the time left for judging each document. Figure 4.1 shows an example of the user interface for the main task.

4.2.1 Document length for the task phase

Each participant was shown document excerpts instead of full documents for three out of the six sets with the aim of studying the difference in their accuracy and speed when given document excerpts. All the study settings and the user interfaces were the same when we used document excerpts.

4.2.2 Time constraints for the task phase

The study involved three different time constraints for sets of 20 documents: 15 seconds per doc, 30 seconds per doc, and 60 seconds per doc. When participants ran out of time, the document got hidden and they had to finalize their judgment. We recorded the overall time they spent on judging the document including the time after hiding the document.

4.3 Balanced Design

In order to eliminate the effects of learning curves and to make sure all the participants experience every condition, we used Latin squares to design a fully balanced and randomized ordering for the topics and the treatments. First, we created a 6x6 Latin square with six letters, randomized the rows and columns and assigned the treatments to the letters randomly. Each row then represented the ordering of the treatments. For the topics, we created another 6x6 Latin square, randomized the rows and columns and assigned the topics to the letters randomly. But this time, each row of the square represented the ordering of the topics for a whole block of six users. So each block of six users had the same ordering of topics but the ordering was shuffled for the other blocks. For each topic, the participants were presented with the same 20 documents but the ordering of displaying them was randomized.

A	B	C	D	E	F
B	A	F	E	C	D
C	F	B	A	D	E
D	C	E	B	F	A
E	D	A	F	B	C
F	E	D	C	A	B

Table 4.2: 6*6 Latin Square used in the study

4.4 Participants

After getting the approval from the Office of Research Ethics (ORE), we recruited 62 participants, 14 for the pilot study and 48 for the main study. The aim of the pilot study was to run the study and find any possible bug or design flaw in practice. After we ran the pilot study, we found that our design was not balanced enough so we changed our Latin square designs as explained in the previous section. The results from the pilot study were not included in the final analyses. Participants were selected on first come first serve basis. Participants were between the ages of 18 and 49. Participants in the main task consisted of 24 males and 24 females. All of the participants were students except two where one of them was a research assistant and the other one an art studies alumni. 24 were graduate and 22 undergraduate students. All participants were fluent in speaking and reading English. Three students were art students while the rest were a science, technology, engineering, or math student. All of the participants used search engines several times a day. Except three, all the other participants considered themselves as experts at finding information using search engines. 9 participants mentioned that they have trouble finding information on the internet.

Chapter 5

Results

5.1 Analysis of Accuracy

Based on the signal detection theory (Abdi, 2007), we can consider the task of judging documents as relevant or non-relevant as a classic signal detection task where the choices for decisions are only Yes and No. Therefore, we can study the performances of the assessors by calculating the number of hits (number of times the assessor choose relevant correctly), and the number of false alarms (number of times the assessor is wrong in saying relevant). In information retrieval, we consider hits as true positives and false alarms as false positives. We can also compute true negatives and false negatives based on the same theory. Table 5.1 shows the confusion matrix for the type of responses that we get from the assessors.

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

Table 5.1: Confusion matrix

Now to study users' behavior and measure their performance, we can use the true positive rate or TPR (the fraction of relevant documents judged as relevant) and the false positive rate or FPR (the fraction of non-relevant documents judged as relevant). We can also compute the accuracy based on TP, TN, FP, and FN.

The true positive rate is measured as:

$$TPR = \frac{|TP|}{|TP|+|FN|} \quad (5.1)$$

The false positive rate is measured as:

$$FPR = \frac{|FP|}{|FP|+|TN|} \quad (5.2)$$

The accuracy is measured as:

$$Accuracy = \frac{|TP|+|TN|}{|TP|+|FP|+|TN|+|FN|} \quad (5.3)$$

doc.factor	time.factor			Mean	p-value
	15	30	60		
F	0.69	0.70	0.72	0.70	$p = 0.1980$
S	0.72	0.71	0.73	0.72	
Mean	0.70	0.70	0.72	0.71	
<i>p-value</i>	$p = 0.1779$				

Table 5.2: Average accuracy. F stands for Full documents and S for Summaries or document excerpts.

Table 5.2 shows the average accuracy results for all the 48 participants. Each participant judged six sets of 20 documents, overall, 120 documents. As shown in the table, the overall accuracy of all the participants was 71%. As we were expected, their accuracy got slightly better when they had more time to judge documents but the difference was not significant, 70%, 70%, and 72% for time constraints of 15, 30, and 60 seconds. We calculated the p-value using linear mixed effects models. In our equations we used doc.factor (full doc or document excerpt) and time.factor (different time constraints) as our fixed variables. Topics and participant ids were our random effects. As it can be seen in the table, the p-values of 0.1779 and 0.1980 for time.factor and doc.factor indicate that the effects were not statically significant. Equation 5.4 and 5.5 show our mixed linear models.

An interesting point to mention is that participants performed not only as good as but even better when shown document excerpts instead of full documents. The overall average accuracy for document excerpts was 72% where full documents had an average accuracy

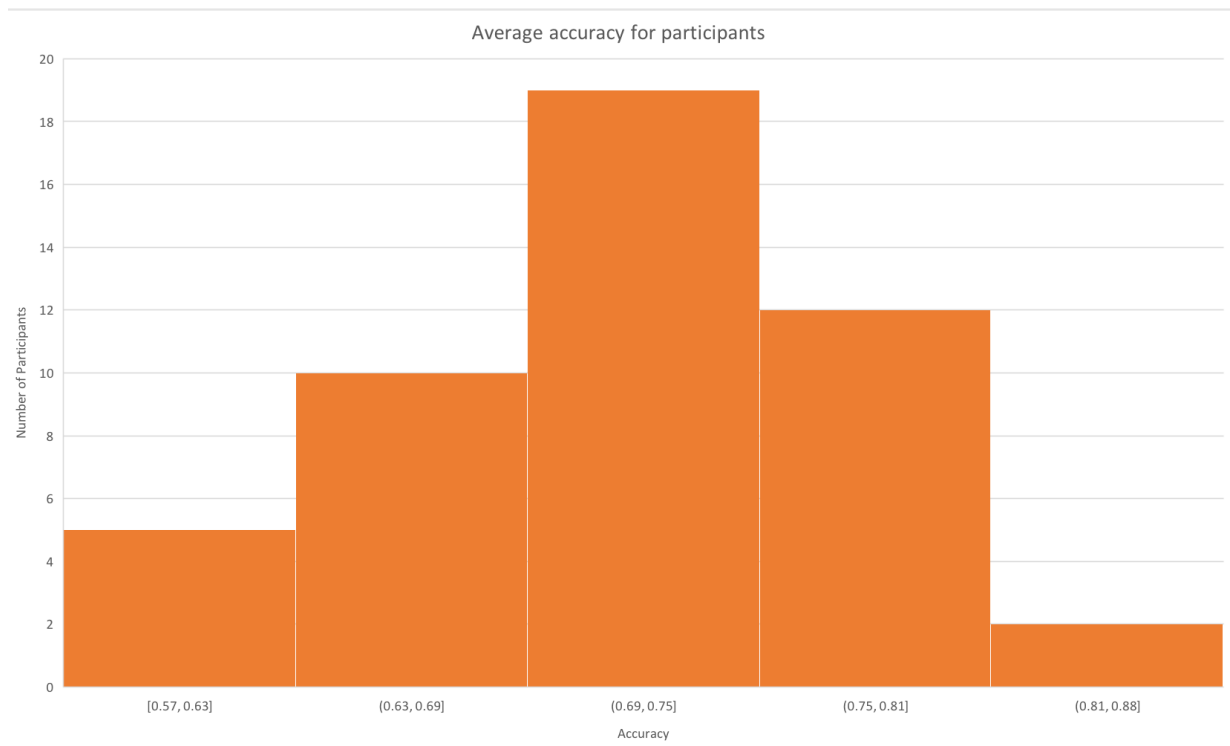


Figure 5.1: Average accuracy for all the participants

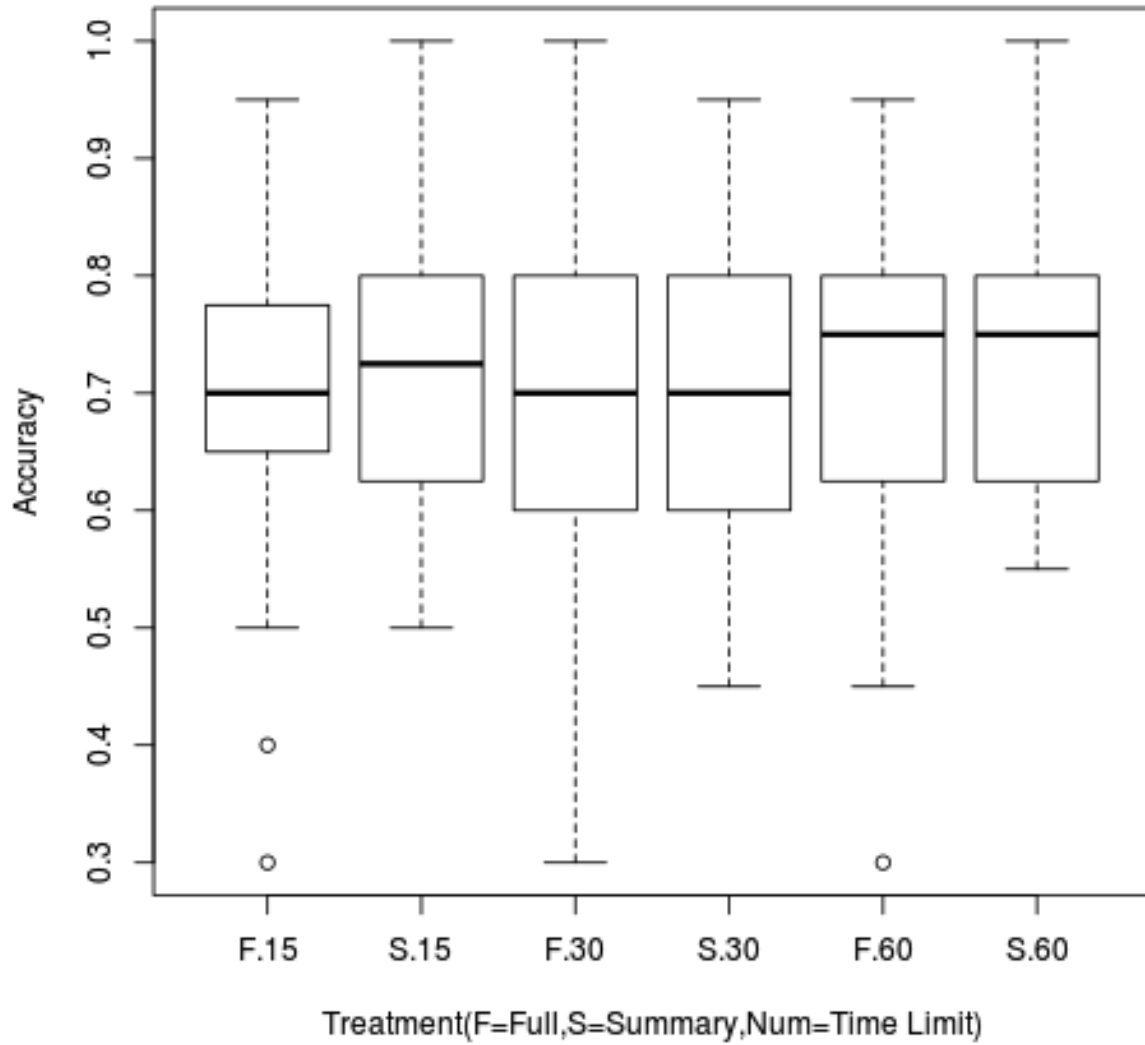


Figure 5.2: Participant's Accuracy for a given Treatment. F stands for Full documents and S for Summaries or document excerpts.

of 70%. This follows the findings by an earlier study by ZHANG et al. (2017) where they found assessors working within a given amount of time, would find more relevant documents by viewing paragraph-length document excerpts rather than full documents.

$$lmer(accuracy \sim time.factor + (1|pid) + (1|topic), data = allData)) \quad (5.4)$$

$$lmer(accuracy \sim doc.factor + (1|pid) + (1|topic), data = allData)) \quad (5.5)$$

5.2 Analysis of Rates

In order to understand participants judging behavior better, we calculated and analyzed their true positive and false positive rates for different tasks of the study. The true positive rate (TPR) is defined as:

$$TPR = \frac{|TP|}{|TP|+|FN|} \quad (5.6)$$

and false positive rate (FPR) is calculated as:

$$FPR = \frac{|FP|}{|FP|+|TN|} \quad (5.7)$$

In later sections we also analyze and calculate the assessor’s criterion which represents how liberal or conservative the assessor is while judging. The assessor’s criterion is defined as:

$$\frac{1}{2}(z(TPR) + z(FPR)) \quad (5.8)$$

Since true positive or false positive rates of 1 and 0 can lead to infinities in the function z , we employed a standard smoothing mechanism by adding a pseudo-document to the count of documents judged. Therefore, the estimated true positive rate (eTPR) is defined as:

$$eTPR = \frac{|TP|+0.5}{|TP|+|FN|+1} \quad (5.9)$$

and estimated false positive rate (eFPR) is calculated as:

$$eFPR = \frac{|FP|+0.5}{|FP|+|TN|+1} \quad (5.10)$$

Table 5.3 and 5.4 show the results for eTPR and eFPR calculations. As it can be seen, participants true positive rates are way higher than false positive rates. This means that participants did better in finding relevant documents and had fewer mistakes in choosing non-relevant documents as relevant. The highest true positive rate was for documents excerpts and time constraint of 15s while the lowest was for full documents and time constraint of 60 seconds. Overall, document excerpts had a higher rate of true positives and slightly higher rate of false positives.

doc.factor	time.factor				p-value
	15	30	60	Mean	
F	0.64	0.63	0.62	0.63	$p = 0.0335$
S	0.71	0.66	0.66	0.68	
Mean	0.68	0.64	0.64	0.65	
<i>p-value</i>	$p = 0.2070$				

Table 5.3: Estimated true positive rates for different treatments of the study

doc.factor	time.factor				p-value
	15	30	60	Mean	
F	0.30	0.26	0.22	0.26	$p = 0.2752$
S	0.31	0.27	0.24	0.27	
Mean	0.30	0.26	0.23	0.26	
<i>p-value</i>	$p = 8.338715e-06$				

Table 5.4: Estimated false positive rates for different treatments of the study

5.3 Analysis of Time

In order to study users' judgments speed, we collected precise data regarding their time spent on each document. Table 5.5 shows the average time participants spent on judging documents within each treatment. As shown in the table, the overall average time spent on judging the documents was 12.8 seconds. On average, the more time the participants had, the more they spent on judging (9.0, 12.3, and 17.0 seconds for 15, 30, and 60 second time constraints). They also spent significantly less time on judging when shown document excerpts instead of full documents. The overall average time spent on document excerpts was 10.59s while they spent 14.93 seconds on average when shown full documents. Now

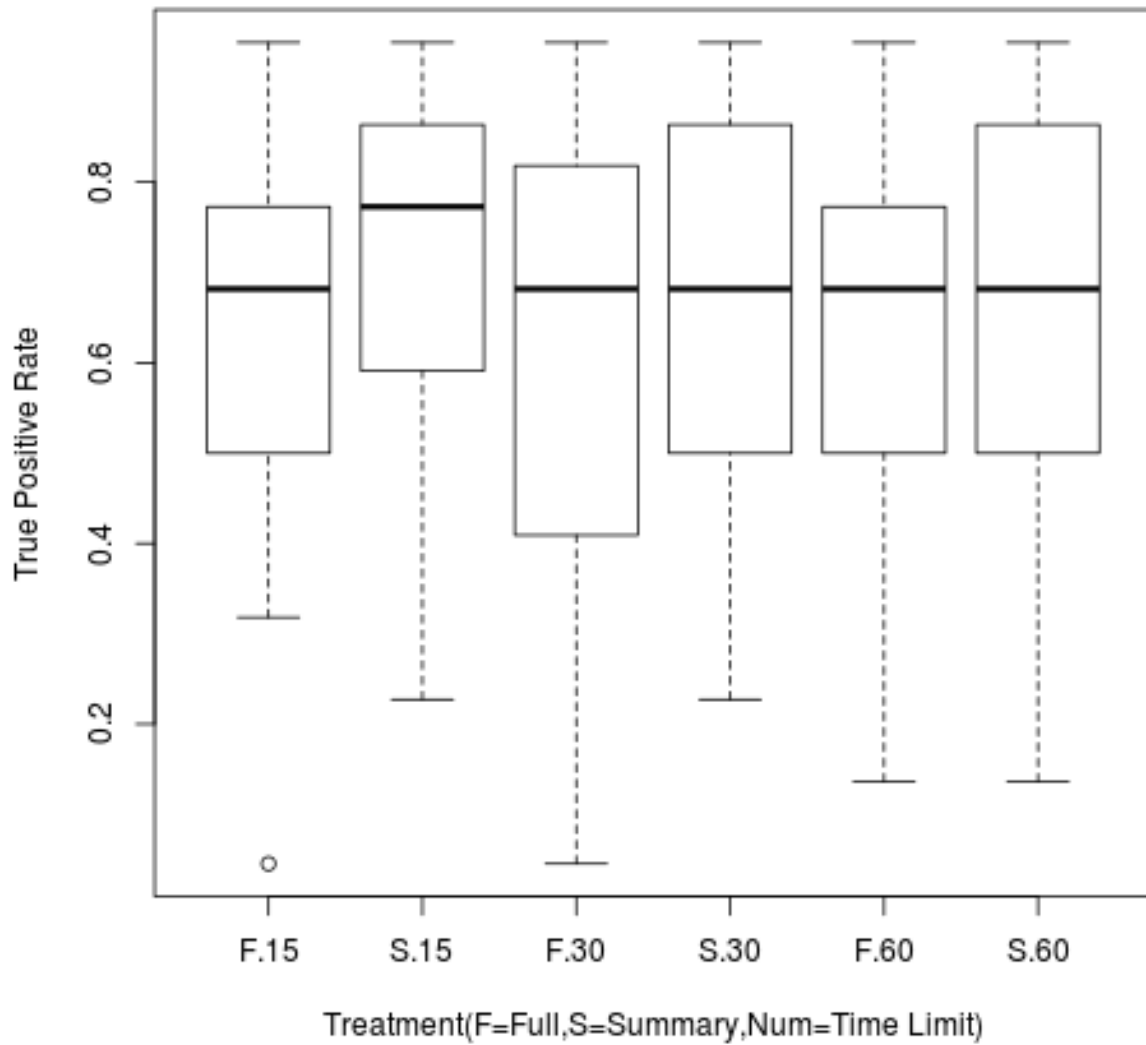


Figure 5.3: True positive rates for all the 6 treatments. F stands for Full documents and S for Summaries or document excerpts.

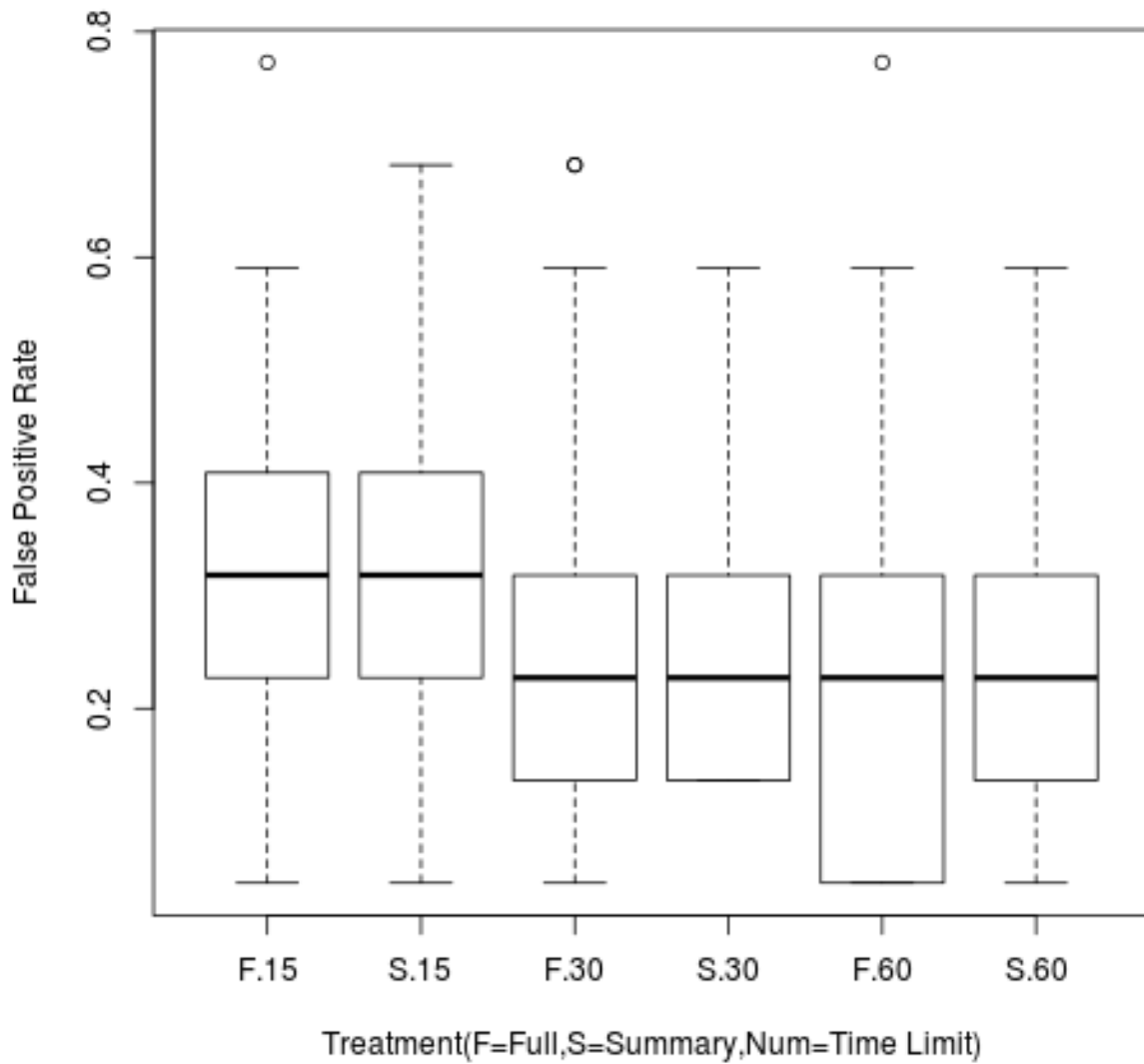


Figure 5.4: False positive rates for all the 6 treatments. F stands for Full documents and S for Summaries or document excerpts.

the important question here is that although they spent less time on document excerpts and that reduces the cost of relevance assessment, but how did that affect their quality of judging?

doc.factor	time.factor				<i>p</i> -value
	15	30	60	Mean	
F	9.33	14.06	21.39	14.93	<i>p</i> = 3.69518e-09
S	8.65	10.44	12.68	10.59	
Mean	8.99	12.25	17.04	12.76	
<i>p</i> -value	<i>p</i> = 8.935529e-18				

Table 5.5: Average Time in seconds

Table 5.6 shows the average accuracy and time spent for the time constraints of 15, 30, and 60 seconds. As shown in the table, while the average times spent on judging for 15 and 30 seconds are less than 60 seconds, the accuracy is almost the same meaning that by forcing assessors to work faster, we can still have the same quality of judging but with less cost for relevance assessment. We can also see that the average time spent on judging for 60s time constraint is 17.04 seconds which is way less than the 60s time allowance. This can also mean that after a certain time threshold, no matter how much time given to the assessors, they judge the documents within that certain amount of time. The overall average time of 12.76s can follow this conclusion.

time.factor	Accuracy	Average time
15	0.70	8.99
30	0.70	12.25
60	0.72	17.04
Mean	0.71	12.76

Table 5.6: Average accuracy and time spent on judging for different time constraints

Table 5.7 shows the number of times participants exceeded from their time limit to judge the document. As it can be seen, participants exceeded their time limit way more when shown full documents. Table 5.8 shows the average time they spent after their time limit passed and the document became hidden.

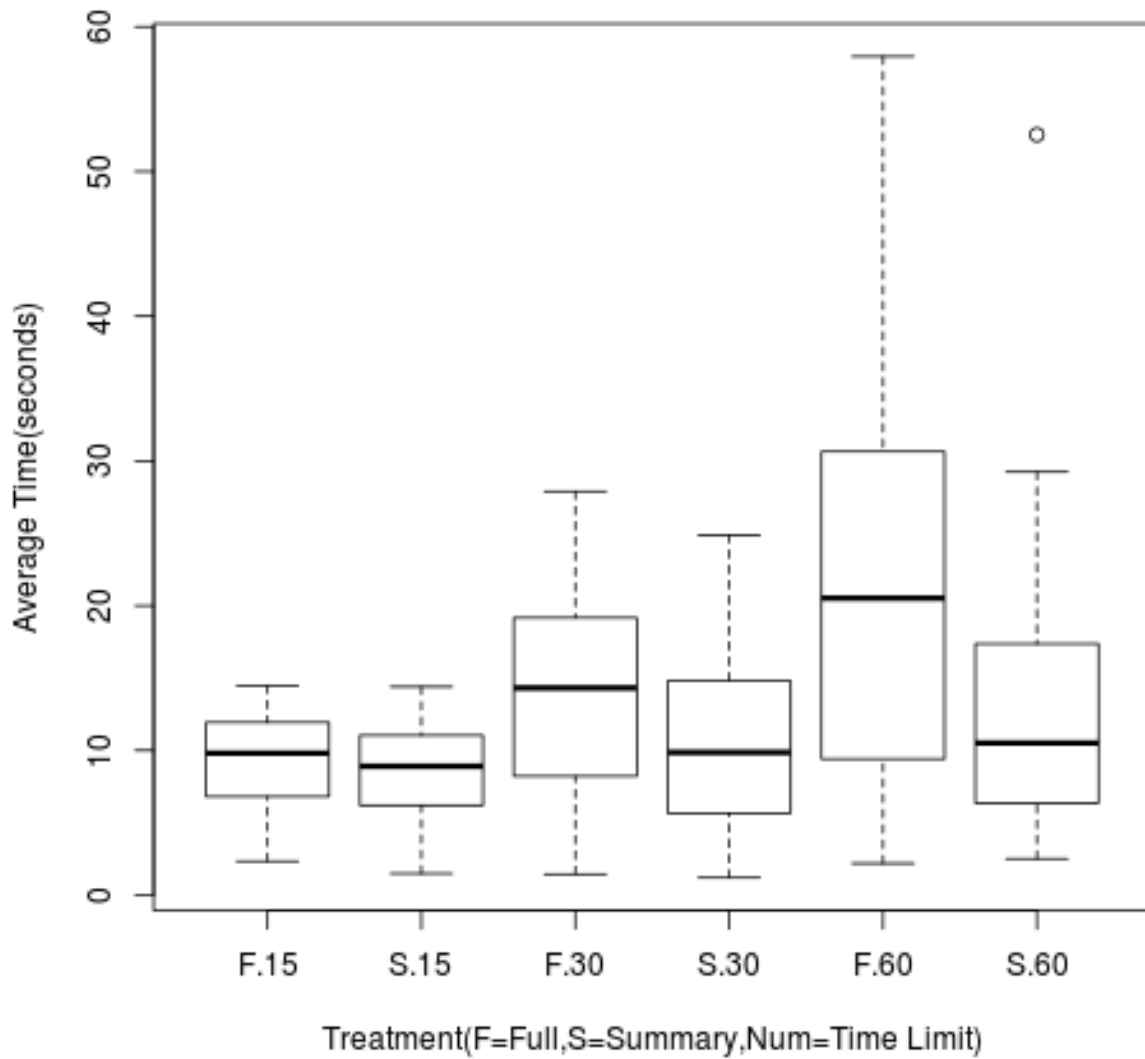


Figure 5.5: Participant's average time for a given Treatment. F stands for Full documents and S for Summaries or document excerpts.

time.factor				
doc.factor	15	30	60	Sum
F	204	109	39	352
S	140	30	0	170
Sum	344	139	39	522

Table 5.7: Number of times participants exceeded from their time limit.

time.factor				
doc.factor	15	30	60	Mean
F	1.89	1.53	0.99	1.47
S	2.45	0.70	0.00	1.05
Mean	2.17	1.12	0.49	1.26

Table 5.8: Average time participants spent after their time limit passed and the document became hidden.

5.4 Analysis of Assessor Criterion and Ability to Discriminate

In order to study users' judging behavior better, we also used two other measures from the signal detection theory called assessors ability to discriminate between relevant and non-relevant documents or d' , and assessors criterion or c .

The assessors ability to discriminate is defined as:

$$d' = z(TPR) - z(FPR) \quad (5.11)$$

where the function z is the inverse of the normal distribution function.

While a d' value of 0 indicates random behavior, a greater value of d' indicates better ability to discriminate between relevant and non-relevant documents.

The assessors criterion is defined as:

$$\frac{1}{2}(z(TPR) + z(FPR)) \quad (5.12)$$

A criterion represents user's strategy in judging documents. A positive criterion represents that the user is more conservative, meaning that they will miss relevant documents

but they will have a lower false positive rate. On the other hand, a negative criterion represents that the user has a more liberal behavior, meaning that they will commit false positive mistakes so that they do not miss relevant documents.

doc.factor	time.factor				p-value
	15	30	60	Mean	
F	1.01	1.16	1.27	1.15	$p = 0.1870$
S	1.24	1.21	1.31	1.25	
Mean	1.12	1.19	1.29	1.20	
<i>p-value</i>	$p = 0.0844$				

Table 5.9: Average estimated ability to discriminate (d')

Table 5.9 shows the average estimated ability to discriminate for different time constraints and document lengths. Participants ability to discriminate between relevant and non-relevant documents increased when given more time. This increase also applies when shown document excerpts instead of full documents. As it can be seen the d' value is 1.25 for document excerpts while it is 1.15 for full documents.

doc.factor	time.factor				p-value
	15	30	60	Mean	
F	-0.10	-0.17	-0.27	-0.18	$p = 0.0015$
S	0.04	-0.05	-0.14	-0.05	
Mean	-0.028	-0.11	-0.20	-0.11	
<i>p-value</i>	$p = 0.0006$				

Table 5.10: Assessors Criterion

Table 5.10 shows that most of the participants had a liberal behavior when judging the documents. The average criterion for all the treatments was negative showing that assessors preferred to commit false positive mistakes so that they do not miss relevant documents. The only positive criterion was for document excerpts with 15 seconds time constraint. However, we can see that participants' liberal behavior increases when given more time to judge the documents.

5.5 Time Vs. TPR and FPR

Figure 5.8 and figure 5.9 show the average time spent on judging documents versus true positive and false positive rates. Participants were prone to have higher false positive rate

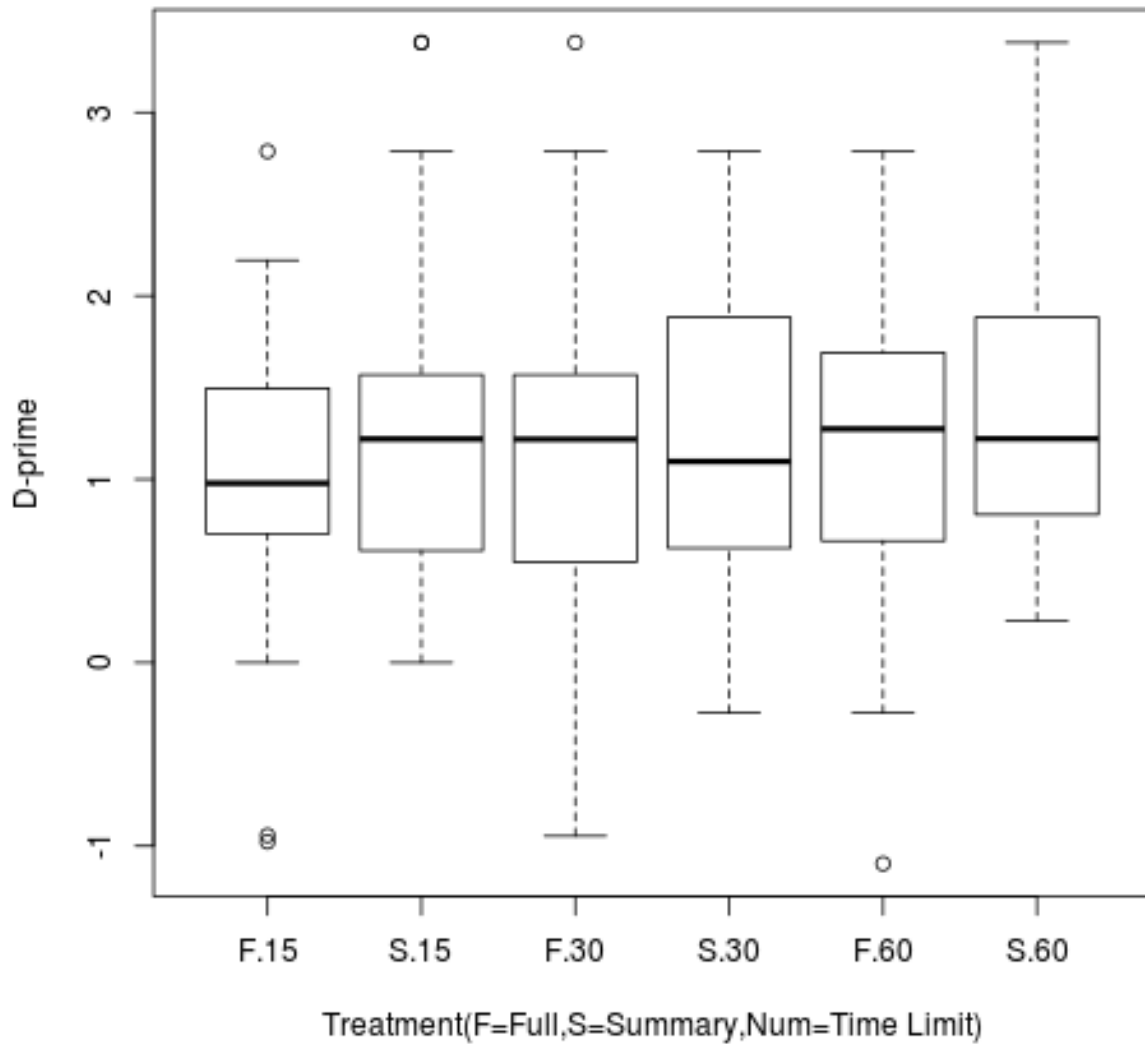


Figure 5.6: Participant's estimated ability to discriminate (d') for a given Treatment. F stands for Full documents and S for Summaries or document excerpts.

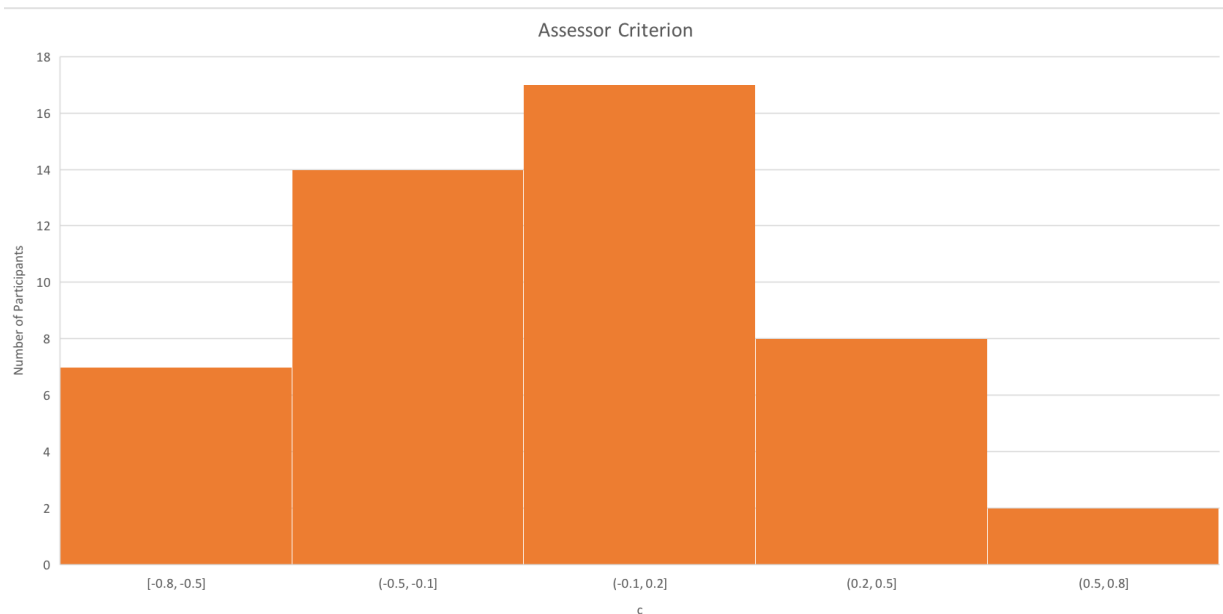


Figure 5.7: Assessor Criterion for all the participants.

when they judged faster.

5.6 Time Vs. d' and Criterion

Figure 5.10 and figure 5.11 show the average time spent on judging the documents versus the ability of discriminate (d') and assessors' Criterion. When participants judged faster, they tended to have lower value of d' .

5.7 Speed, Accuracy and topics

We analyzed the accuracy and speed of the participants within each topic.

Table 5.11 shows that overall, topic 436 (Railway Accidents) had the highest average accuracy (83%) and it was the easiest for the assessors to judge. Topic 310 (Radio Waves and Brain Cancer) had the lowest accuracy (66%) and it was the hardest to judge. The

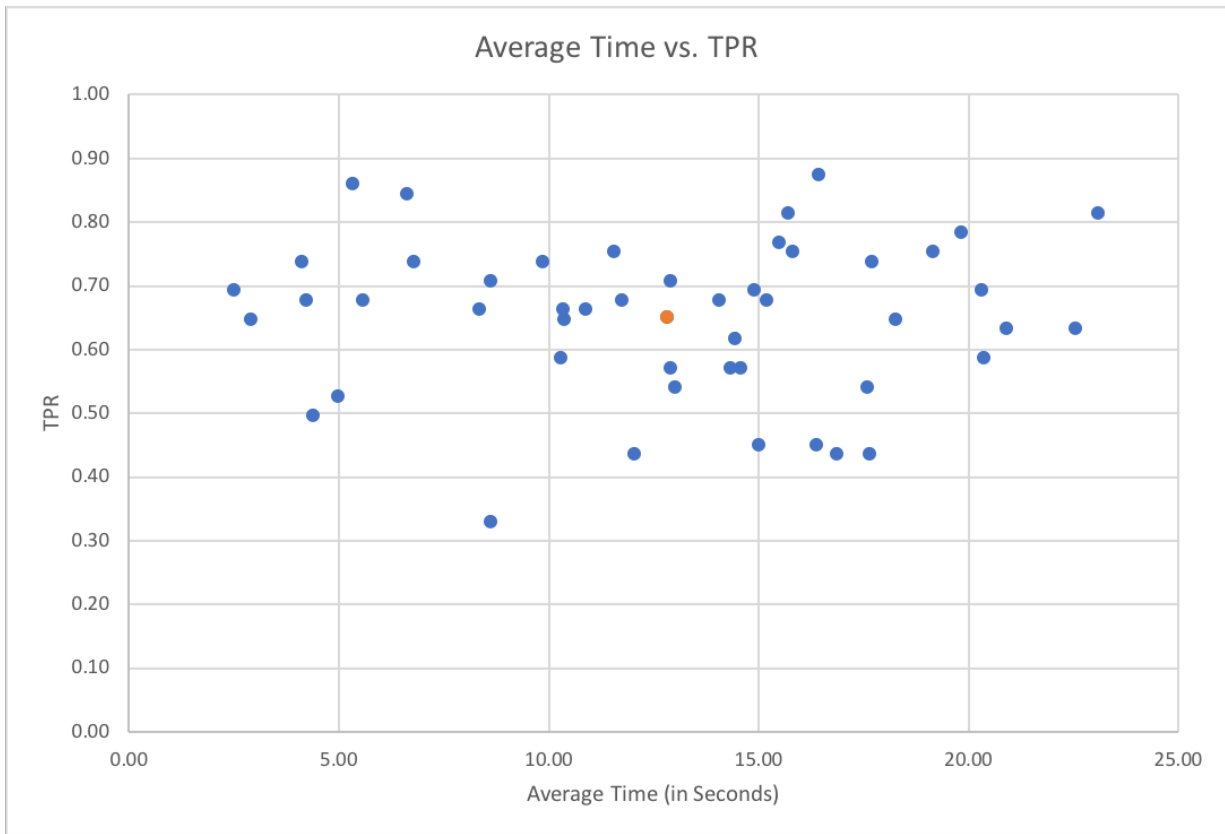


Figure 5.8: Average Time vs. True Positive Rate for all the participants. The orange dot is the overall mean.

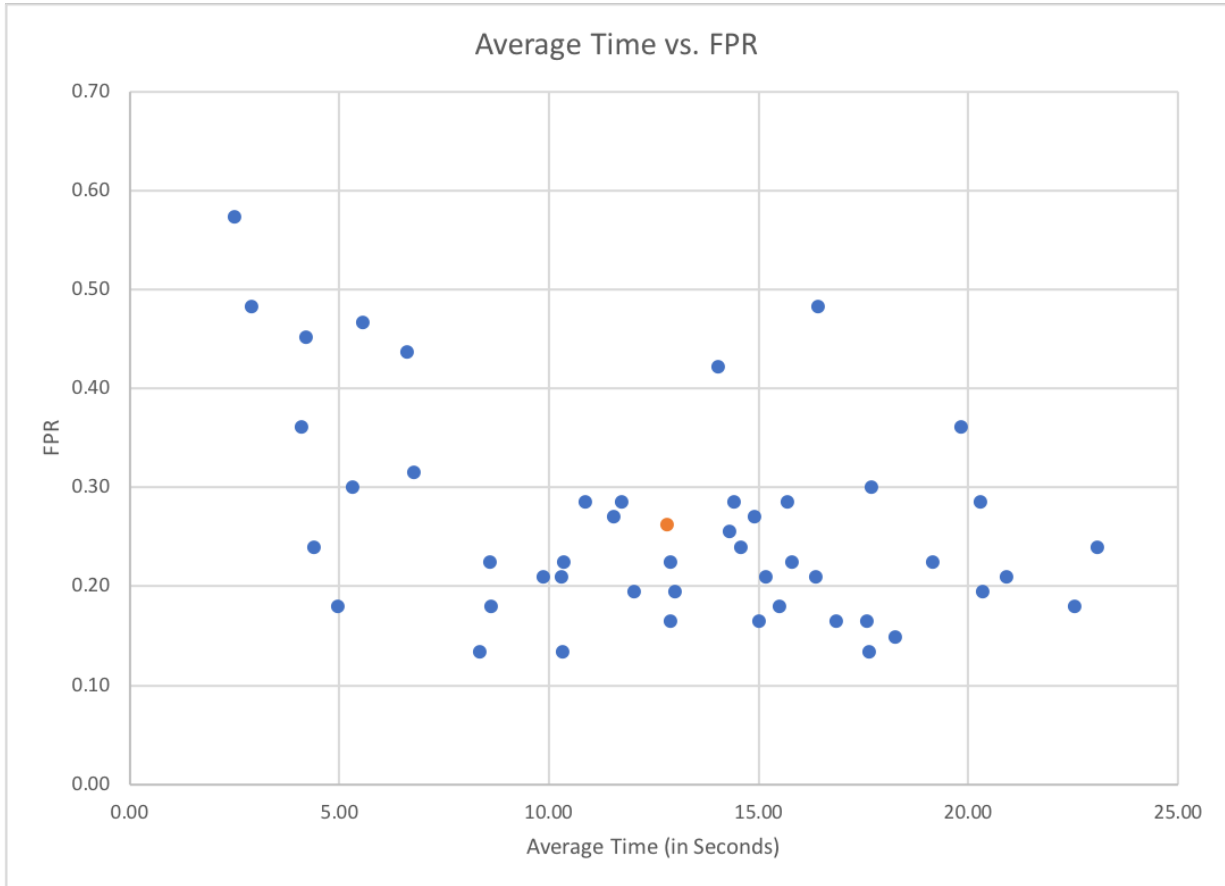


Figure 5.9: Average Time vs. False Positive Rate for all the participants. The orange dot is the overall mean.

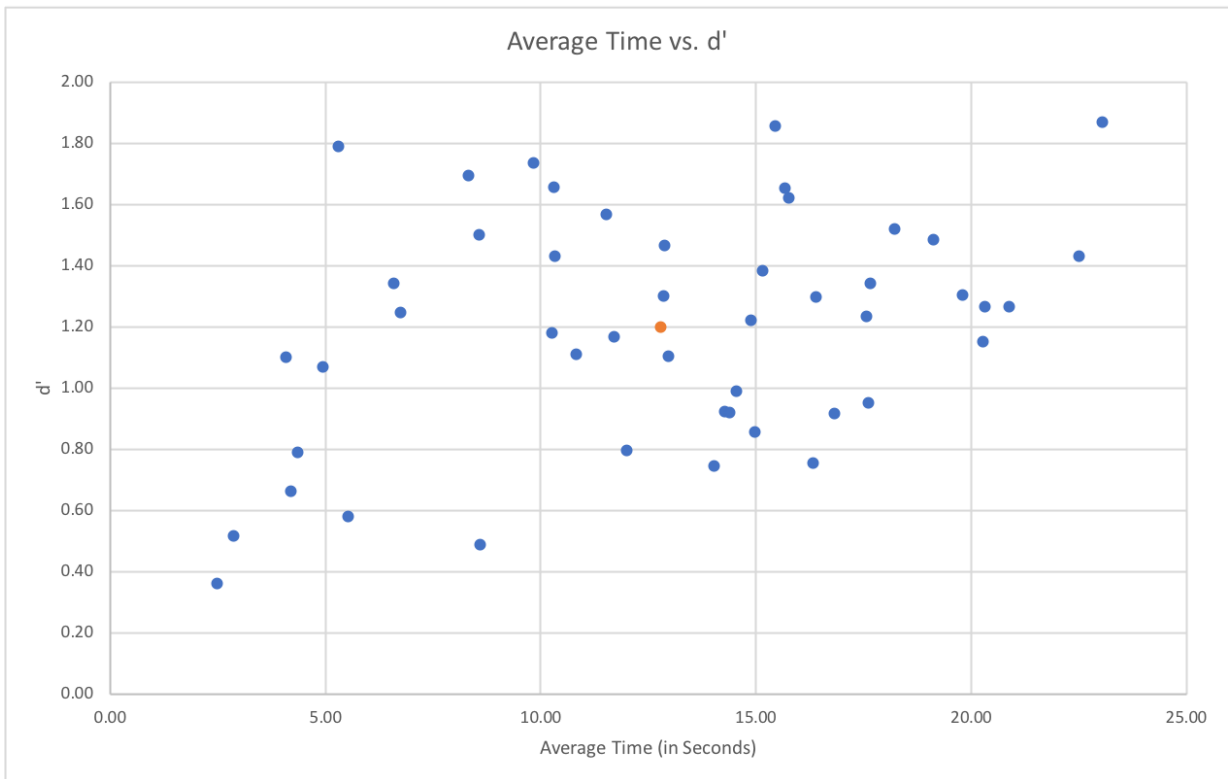


Figure 5.10: Average Time vs. d' for all the participants. The orange dot is the overall mean.

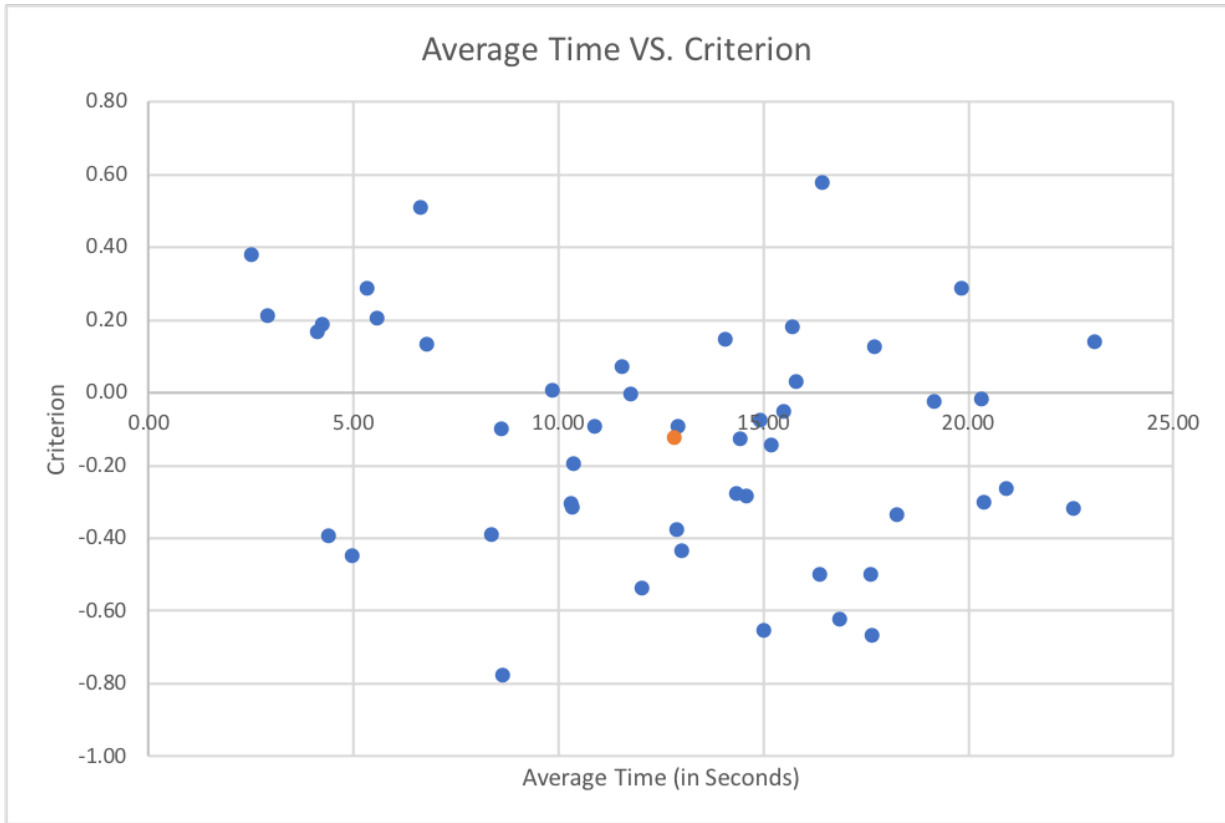


Figure 5.11: Average Time vs. Criterion for all the participants. The orange dot is the overall mean.

highest average accuracy within a time constraint was 85% for topic 436 and time constraints of 15 and 60 seconds while the lowest was 63% for topic 310 and time constraint of 30 seconds.

time.factor				
topic	15	30	60	Mean
310	0.68	0.63	0.67	0.66
336	0.67	0.70	0.66	0.68
362	0.65	0.72	0.70	0.69
426	0.67	0.72	0.75	0.71
427	0.69	0.67	0.70	0.68
436	0.85	0.79	0.85	0.83
Mean	0.70	0.70	0.72	0.71

Table 5.11: Average accuracy within each topic and time constraint

Table 5.12 shows the average accuracy for all the topics with different document lengths. The highest average accuracy of 84% was obtained when participants were shown document excerpts for topic 436 (Railway Accidents). Participants had the lowest average accuracy were shown document excerpts for topic 310 (Radio Waves and Brain Cancer).

doc.factor			
topic	F	E	Mean
310	0.70	0.62	0.66
336	0.70	0.66	0.68
362	0.73	0.66	0.69
426	0.63	0.80	0.71
427	0.63	0.73	0.68
436	0.82	0.84	0.83
Mean	0.70	0.72	0.71

Table 5.12: Average accuracy within each topic and document length

5.8 Inconsistency in Relevance Judgments

We used NIST judgments for the seven topics that we used in this user study (one for the tutorial phase and six for the main task) as the source of comparison. However, according to the study by [Harman \(2011\)](#) NIST judgments are known to have inconsistency and therefore, they don't make the best gold standard. The reason is that a document judged by an assessor may be judged differently by another assessor so there can be errors in the judgments. Measures such as true and false positive rates can get affected because of these errors. However, all the documents that we selected for the study were selected based on a stratified sampling and every participant was shown the same 120 documents. Therefore, although the exact numbers reported for true positive rates, false positive rates and accuracies may not be absolutely correct, but the comparison between different treatments of the study will remain valid.

5.9 Participants Perceived Difficulty

Table 5.13 shows the average accuracy and participants perceived difficulty for each topic. We calculated the perceived difficulty percentage by counting the number of participants that said the judging of that specific topic was difficult or very difficult in the post questionnaires. For example, a perceived difficulty of 25% for topic 310 means that 25% of the participants found the documents of topic 310 to be hard to judge while 75% found them easy to judge. We couldn't find any significant correlation between the quality of judging and perceived difficulty but the hardest topic was reported to be topic 427 (UV damage, eyes) based on the questionnaires results.

topic	accuracy	perceived difficulty
310	0.66	0.25
336	0.68	0.37
362	0.69	0.35
426	0.71	0.29
427	0.68	0.39
436	0.83	0.25

Table 5.13: Participants perceived difficulty and average accuracy.

Table 5.14 shows the average time and participants perceived difficulty for each topic. Topic 362 (Human Smuggling) which had the highest average time spent on judging, did

not have the highest perceived difficulty. Topic 436 had the lowest average time and along with topic 310, found to be the easiest to judge by the participants.

topic	time	perceived difficulty
310	12.80	0.25
336	14.39	0.37
362	14.84	0.35
426	11.09	0.29
427	12.52	0.39
436	10.93	0.25

Table 5.14: Participants perceived difficulty and average time.

5.10 Participants Experience and Mood during the Relevance Assessment

Table 5.15 shows the mood of the participants and their experience of relevance assessment tasks. The results are based on the post-questionnaires after each topic. The values show the percentage of the participants that found the task of judging enjoyable and they felt engaged during the tasks. We couldn't find any significant correlation between the quality of judging and participants mood doing the tasks or their experience of the relevance assessment process.

topic	enjoyable	engaged
310	0.50	0.64
336	0.39	0.43
362	0.41	0.54
426	0.29	0.50
427	0.54	0.52
436	0.43	0.43

Table 5.15: Participants Experience and Mood during the Relevance Assessment.

5.11 Concentration and Stress under Time Pressure

Table 5.16 shows the percentage of participants who felt stressed under time pressure and find it hard to concentrate while doing the relevance judging. The rest, found it to be easy and they felt relaxed doing the judging tasks.

topic	difficult	stressed
310	0.18	0.45
336	0.39	0.54
362	0.27	0.52
426	0.12	0.39
427	0.31	0.54
436	0.16	0.43

Table 5.16: Participants Concentration and Stress under Time Pressure.

5.12 Participants Perceived Accuracy

Table 5.17 shows the percentage of the participants who felt they judged the documents accurately along with their actual average accuracy in comparison to NIST judgments. For example for topic 310, 62% of the participants felt they judged the documents accurately while the actual average accuracy for topic 310 was 66%.

topic	actual accuracy	perceived accuracy
310	0.66	0.62
336	0.68	0.45
362	0.69	0.52
426	0.71	0.60
427	0.68	0.54
436	0.83	0.66

Table 5.17: Participants perceived accuracy and actual average accuracy.

5.13 General Analysis of Post Task Questionnaires

Table 5.18 shows our preliminary analysis on the post task questionnaires. Section A.6 displays our detailed questions in the questionnaires. We used 5 point Likert scale to map each question to values 1 through 5 with the most negative answer mapped to 1 (Very Difficult) and the most positive answer mapped to 5 (Very Easy). The neutral response was mapped to 3. Based on the results from table 5.18, we can see that full documents with 15 seconds time constraint found to be the hardest by the participants.

Question	Treatment						
	(15, F)	(30, F)	(60, F)	(15, S)	(30, S)	(60, S)	All
Difficulty	2.5	3.1	2.9	3.3	3.5	3.7	3.2
Experience	3.0	3.1	3.0	3.3	3.5	3.5	3.2
Mood	3.3	3.4	3.3	3.5	3.5	3.3	3.4
Concentration	2.9	3.1	3.1	3.5	3.5	3.7	3.3
Confidence	2.9	2.8	2.5	2.6	2.4	2.1	2.5
Task Became Easier	2.7	2.6	2.6	2.5	2.3	2.3	2.5
Useful Guidelines	2.1	2.2	2.0	2.2	2.1	2.0	2.1
Time Pressure	2.0	2.3	2.9	2.4	3.0	3.1	2.6
Accuracy	2.8	2.6	2.5	2.5	2.2	2.2	2.5

Table 5.18: Post task questionnaire analysis. Sections A.6 displays the details of the questions.

5.14 Comparison with Former Studies

In a study by [Smucker and Jethani \(2010\)](#), the authors studied if the relevance judging behaviour of the assessors change when the prevalence of relevant documents in a set of documents changes. In one of their settings, they had a 0.5 prevalence which means half of the documents were relevant in the set of documents. We had the same level of prevalence (0.5) for all of our treatments. They had 18 participants and each judged 80 documents. Topics used in both studies were the same but the documents were different. We compared the two results from the former study and the current thesis. Table 5.19 shows both results.

As it can be seen in table 5.19, the overall average accuracy in the former study is higher but the average time spent on judging the documents was way less in the current study. While participants judged less documents in the former study, they also had to

study	accuracy	time	TPR	FPR	d'	c
former study	0.82	27.00	0.69	0.05	2.17	0.58
current thesis	0.71	12.76	0.65	0.26	1.20	-0.11

Table 5.19: The comparison between a former study with the same topics and the current thesis.

pass a qualification phase before starting the main task. True positive rates were similar in both studies (0.69 and 0.65) but the false positive rate in the current study was significantly higher which means our participants had more trouble judging the non-relevant documents. Both the ability of discriminate and assessor criterion were less in the current study, meaning that our participants had more liberal behaviour and preferred to commit false positive mistakes so that they do not miss relevant documents and also it was harder for them to discriminate between the relevant and non-relevant documents.

In another study by [Smucker and Jethani \(2011a\)](#), they did a comparison between the crowd-sourced and University laboratory participant behavior in relevance assessment. They conducted two experiments, one was a laboratory-based study with 18 participants and the other one was with 202 crowd-sourced participants. While 100% of the laboratory participants qualified for inclusion in the final group of participants, 30% of the crowd-sourced workers qualified. Each participant judged 80 documents and one of their topics was different than ours. Table 5.20 shows the comparison between our results and their study for the 5 common topics. The missing topic is topic 436 which they used for the tutorial phase.

study	accuracy	time	TPR	FPR	d'	c
former study (crowd)	0.79	14.8	0.75	0.17	1.94	0.17
former study (lab)	0.81	27.2	0.66	0.07	2.16	0.52
current thesis	0.71	12.76	0.65	0.26	1.20	-0.11

Table 5.20: The comparison between five topics of a former study and the current thesis.

As it can be seen in table 5.20, the judging behavior was similar between the two groups of laboratory participants and crowd-sourced workers in the former study. However, crowd-based workers had higher false positive rate but judged the documents twice as fast as laboratory participants. Both groups had higher accuracy and true positive rates than the current study but their overall judgment speed was lower. The false positive rate in the current study was way higher than both groups of the former study which means non-relevant documents in the current study were harder to judge. Both the ability

of discriminate and assessor criterion were lower in the current study which means the participants of the current study had more liberal behaviour and committed more false positive mistakes.

5.15 Analysis of the four classes of document types

We also did an analysis of the four document classes that we had. Using the RRF lists that were explained in section 3.9, we created four lists of higher rrf relevant, lower rrf relevant, higher rrf non-relevant, and lower rrf non-relevant. We then chose five samples from each of these lists to make a set of 20 documents for each of our topics. Table 5.21 shows the accuracy, True Positive Rate, False Positive Rate, the ability to discriminate, and the assessor criterion for each of these document classes. We computed these values by assigning each document to one of the four groups and calculated the average accuracy, TPR, FPR, d' , and c for each class of documents.

docClass	Accuracy	TPR	FPR	d'	c
higherNonRel	0.66	0.50	0.35	0.42	-0.21
higherRel	0.77	0.76	0.50	0.78	0.39
lowerNonRel	0.86	0.50	0.15	1.14	-0.57
lowerRel	0.57	0.56	0.50	0.18	0.09
Mean	0.71	0.58	0.37	0.63	-0.08

Table 5.21: Analysis of the 4 classes of document types

As it can be seen in table 5.21, higher rrf relevant and lower rrf non-relevant have the highest accuracy and tend to be easier to judge for the assessors while the other two lists have lower accuracies and tend to be harder to judge. Higher rrf relevant also had the highest TPR. Both higher rrf relevant and lower rrf non relevant had the higher false positive rate. Participants also had the highest ability to discriminate between relevant and non relevant documents for document class of lower rrf non-relevant. Participants had a liberal behaviour when they were shown documents from types of higher rrf non-relevant and lower rrf non-relevant. They had a conservative behaviour when they were shown higher rrf relevant and lower rrf relevant documents.

Chapter 6

Conclusion

In this thesis, we presented a user study to understand the assessors' relevance judging behavior better. We focused on two main factors, applying time constraints and document lengths. We analyzed the behavior of assessors for three different time constraints and two different length factors, full-length documents and document excerpts or paragraph-length summaries. We used various measures such as true positive rates, false positive rates, accuracy, ability to discriminate, assessors criterion and perceived difficulty.

The key conclusions that we make are:

- We showed that assessors' quality of relevance judging when shown document excerpts, is not only as good as when shown full documents but also in most of the cases better. We had three different time constraints that we applied when assessors had to judge the relevance of the documents. In all three cases, assessors performed better when shown document excerpts instead of full documents.
- We showed that giving assessors more time to judge the documents leads to slightly but not a significant increase in their accuracy and performance. The average time spent on judging the documents in this study was 12.8 seconds.
- Assessors average time spent on judging increased when given more time but this does not lead to significant change in accuracy and performance.
- In general, we can speed up the judging process by using document excerpts and time limitations while maintaining the quality of judging.

- We showed that assessors ability to discriminate between relevant and non-relevant documents increases when given more time.
- We showed that assessors criterion does not change within different time constraints. Our participants showed liberal behavior when judging the documents meaning that they preferred to commit false positive mistakes to avoid missing relevant documents.
- We saw that there is no correlation between assessors actual accuracy and their perceived difficulty of the judging tasks.

In summary, time limits have little effect on the quality of judging but they can speed up the judging process. Also, document excerpts have the potential to match or exceed full documents in performance and the quality of judging but most certainly the judging process is faster when using document excerpts which leads to less cost and effort in relevance assessment.

The final implication is that we can use full documents with time limits or document excerpts with no time limit and achieve similar quality and speed of judging.

Bibliography

- Abdi, H. (2007). Signal detection theory (sdt). *Encyclopedia of measurement and statistics*, 886–889.
- Al-Harbi, A. L. and M. D. Smucker (2013). User expressions of relevance judgment certainty. In *7th Annual Symposium on Human-Computer Interaction and Information Retrieval (HCIR 2013)*.
- Al-Harbi, A. L. and M. D. Smucker (2014). A qualitative exploration of secondary assessor relevance judging behavior. In *Proceedings of the 5th Information Interaction in Context Symposium*, pp. 195–204. ACM.
- Al-Maskari, A., M. Sanderson, and P. Clough (2008). Relevance judgments between trec and non-trec assessors. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 683–684. ACM.
- ALHARBI, A. (2016). Studying relevance judging behavior of secondary assessors.
- Alonso, O. and R. Baeza-Yates (2011). Design and implementation of relevance assessments using crowdsourcing. In *European Conference on Information Retrieval*, pp. 153–164. Springer.
- Alonso, O. and S. Mizzaro (2009). 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*, Volume 15, pp. 16.
- Alonso, O. and S. Mizzaro (2012). Using crowdsourcing for trec relevance assessment. *Information processing & management* 48(6), 1053–1066.
- Alonso, O., D. E. Rose, and B. Stewart (2008). Crowdsourcing for relevance evaluation. In *ACM SigIR Forum*, Volume 42, pp. 9–15. ACM.

- Baeza-Yates, R., B. Ribeiro-Neto, et al. (1999). *Modern information retrieval*, Volume 463. ACM press New York.
- Bailey, P., N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz (2008). Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 667–674. ACM.
- Bales, S. and P. Wang (2005). Consolidating user relevance criteria: A meta-ethnography of empirical studies. *Proceedings of the Association for Information Science and Technology* 42(1).
- Barhydt, G. C. (1964). A comparison of relevance assessments by three types of evaluator. In *Proceedings of the American Documentation Institute*, Volume 1, pp. 383–385. American Documentation Institute Washington, DC.
- Barhydt, G. C. (1967). Effectiveness of non-user relevance assessments.
- Barry, C. L. (1994). User-defined relevance criteria: an exploratory study. *Journal of the American Society for Information Science* 45(3), 149.
- Barry, C. L. and L. Schamber (1998). Users' criteria for relevance evaluation: a cross-situational comparison. *Information processing & management* 34(2-3), 219–236.
- Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of documentation* 56(1), 71–90.
- Borlund, P. (2003). The concept of relevance in ir. *Journal of the Association for Information Science and Technology* 54(10), 913–925.
- Borlund, P. and P. Ingwersen (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of documentation* 53(3), 225–250.
- Buchanan, G. and F. Loizides (2007). Investigating document triage on paper and electronic media. In *International Conference on Theory and Practice of Digital Libraries*, pp. 416–427. Springer.
- Büttcher, S., C. L. Clarke, and G. V. Cormack (2016). *Information retrieval: Implementing and evaluating search engines*. Mit Press.

- Chandar, P., W. Webber, and B. Carterette (2013). Document features predicting assessor disagreement. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 745–748. ACM.
- Chouldechova, A. and D. Mease (2013). Differences in search engine evaluations between query owners and non-owners. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 103–112. ACM.
- Clough, P., M. Sanderson, J. Tang, T. Gollins, and A. Warner (2013). Examining the limits of crowdsourcing for relevance assessment. *IEEE Internet Computing* 17(4), 32–38.
- Cool, C., N. Belkin, O. Frieder, and P. Kantor (1993). Characteristics of text affecting relevance judgments. In *National online meeting*, Volume 14, pp. 77–77. LEARNED INFORMATION (EUROPE) LTD.
- Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information storage and retrieval* 7(1), 19–37.
- Cormack, G. V., C. L. Clarke, and S. Buettcher (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 758–759. ACM.
- Cosijn, E. and P. Ingwersen (2000). Dimensions of relevance. *Information Processing & Management* 36(4), 533–550.
- Croft, W. B., D. Metzler, and T. Strohman (2010). *Search engines: Information retrieval in practice*, Volume 283. Addison-Wesley Reading.
- Cuadra, C. A. (1967). *Experimental Studies of Relevance Judgments. Final Report [by Carlos A. Cuadra and Others]*. System Development Corporation.
- Efthimiadis, E. N. and M. A. Hotchkiss (2008). Legal discovery: Does domain expertise matter? *Proceedings of the Association for Information Science and Technology* 45(1), 1–2.
- Eickhoff, C., C. G. Harris, A. P. de Vries, and P. Srinivasan (2012). Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 871–880. ACM.

- Eisenberg, M. and C. Barry (1988). Order effects: a study of the possible influence of presentation order on user judgements of document relevance. *Journal of the American Society for Information Science* 39(5), 293.
- Froehlich, T. J. (1994). Relevance reconsidered-towards an agenda for the 21st century: introduction to special topic issue on relevance research. *Journal of the American Society for Information Science* 45(3), 124.
- Green, R. (1995). Topical relevance relationships. i. why topic matching fails. *Journal of the American Society for Information Science (1986-1998)* 46(9), 646.
- Grossman, M. R. and G. V. Cormack (2011). Inconsistent assessment of responsiveness in e-discovery: Difference of opinion or human error? In *DESI IV: The ICAIL Workshop on Setting Standards for Searching Electronically Stored Information in Discovery Proceedings*, pp. 1–11. Citeseer.
- Gull, C. D. (1956). Seven years of work on the organization of materials in the special library. *Journal of the Association for Information Science and Technology* 7(4), 320–329.
- Harman, D. (2011). Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 3(2), 1–119.
- Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for information Science* 43(9), 602.
- Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *JASIS* 47(1), 37–49.
- Janes, J. W. (1994). Other people’s judgments: A comparison of users’ and others’ judgments of document relevance, topicality, and utility. *Journal of the American Society for Information science* 45(3), 160.
- Janes, J. W. and R. McKinney (1992). Relevance judgments of actual users and secondary judges: A comparative study. *The Library Quarterly* 62(2), 150–168.
- Jones, K. S. (1997). *Readings in information retrieval*. Morgan Kaufmann.
- Kinney, K. A., S. B. Huffman, and J. Zhai (2008). How evaluator domain expertise affects search result relevance judgments. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 591–598. ACM.

- Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American society for information science* 42(5), 361.
- Lesk, M. E. and G. Salton (1968). Relevance assessments and retrieval system evaluation. *Information storage and retrieval* 4(4), 343–359.
- Maddalena, E., M. Basaldella, D. De Nart, D. Degl'Innocenti, S. Mizzaro, and G. Demartini (2016). Crowdsourcing relevance assessments: The unexpected benefits of limiting the time to judge. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.
- Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting with computers* 10(3), 303–320.
- O'Connor, J. (1969). Some independent agreements and resolved disagreements about answer-providing documents. *Journal of the Association for Information Science and Technology* 20(4), 311–319.
- Park, T. K. (1994). Toward a theory of user-based relevance: A call for a new paradigm of inquiry. *Journal of the American society for information science* 45(3), 135.
- Rees, A. M. and D. G. Schultz (1967). A field experimental approach to the study of relevance assessments in relation to document searching. final report to the national science foundation. volume ii, appendices.
- Roegiest, Adam (2017). *On Design and Evaluation of High-Recall Retrieval Systems for Electronic Discovery*. Ph. D. thesis.
- Ruthven, I., M. Baillie, and D. Elswiler (2007). The relative effects of knowledge, interest and confidence in assessing relevance. *Journal of Documentation* 63(4), 482–504.
- Sanderson, M. et al. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval* 4(4), 247–375.
- Sandhaus, E. (2008). The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia* 6(12), e26752.
- Saracevic, T. (1969). Comparative effects of titles, abstracts and full texts on relevance judgments. *Proceedings of the American Society for Information Science* 6(1), 293–299.
- Saracevic, T. (1996). Relevance reconsidered. In *Proceedings of the second conference on conceptions of library and information science (CoLIS 2)*, pp. 201–218.

- Schamber, L. (1994). Relevance and information behavior. *Annual review of information science and technology (ARIST)* 29, 3–48.
- Schamber, L., M. B. Eisenberg, and M. S. Nilan (1990). A re-examination of relevance: toward a dynamic, situational definition. *Information processing & management* 26(6), 755–776.
- Scholer, F., A. Turpin, and M. Sanderson (2011). Quantifying test collection quality based on the consistency of relevance judgements. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 1063–1072. ACM.
- Smucker, M. D. and C. P. Jethani (2010). Human performance and retrieval precision revisited. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 595–602. ACM.
- Smucker, M. D. and C. P. Jethani (2011a). The crowd vs. the lab: A comparison of crowd-sourced and university laboratory participant behavior. In *Proceedings of the SIGIR 2011 Workshop on crowdsourcing for information retrieval*.
- Smucker, M. D. and C. P. Jethani (2011b). 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*, pp. 1231–1232. ACM.
- Snow, R., B. O’Connor, D. Jurafsky, and A. Y. Ng (2008). 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*, pp. 254–263. Association for Computational Linguistics.
- Spink, A., H. Greisdorf, and J. Bateman (1998). From highly relevant to not relevant: Examining different regions of relevance¹. *Information Processing & Management* 34(5), 599–621.
- Tang, R. and P. Solomon (1998). Toward an understanding of the dynamics of relevance judgment: An analysis of one person’s search behavior. *Information processing & management* 34(2-3), 237–256.
- Taube, M. (1965). A note on the pseudo-mathematics of relevance. *Journal of the Association for Information Science and Technology* 16(2), 69–72.

- Taylor, A. R., C. Cool, N. J. Belkin, and W. J. Amadio (2007). Relationships between categories of relevance criteria and stage in task completion. *Information Processing & Management* 43(4), 1071–1084.
- Vakkari, P. (2000). Relevance and contributing information types of searched documents in task performance. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 2–9. ACM.
- Vakkari, P. and N. Hakala (2000). Changes in relevance criteria and problem stages in task performance. *Journal of documentation* 56(5), 540–562.
- Villa, R. and M. Halvey (2013). Is relevance hard work?: evaluating the effort of making relevant assessments. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 765–768. ACM.
- Voorhees, E. M. (2001). The philosophy of information retrieval evaluation. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pp. 355–370. Springer.
- Wakeling, S., M. Halvey, R. Villa, and L. Hasler (2016). A comparison of primary and secondary relevance judgements for real-life topics. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pp. 173–182. ACM.
- Wang, J. (2011). Accuracy, agreement, speed, and perceived difficulty of users relevance judgments for e-discovery. In *Proceedings of SIGIR Information Retrieval for E-Discovery Workshop*, Volume 1.
- Wang, J. and D. Soergel (2010). A user study of relevance judgments for e-discovery. *Proceedings of the Association for Information Science and Technology* 47(1), 1–10.
- Wang, P. and M. D. White (1999). A cognitive model of document use during a research project. study ii. decisions at the reading and citing stages. *Journal of the Association for Information Science and Technology* 50(2), 98.
- Wilson, P. (1973). Situational relevance. *Information storage and retrieval* 9(8), 457–471.
- Xu, Y. C. and Z. Chen (2006, May). Relevance judgment: What do information users consider beyond topicality? *J. Am. Soc. Inf. Sci. Technol.* 57(7), 961–973.
- ZHANG, H., M. ABUALSAUD, N. GHELANI, A. GHOSH, M. D. SMUCKER, G. V. CORMACK, and M. R. GROSSMAN (2017). Uwaterloomds at the trec 2017 common core track.

Zhang, H., G. V. Cormack, M. R. Grossman, and M. D. Smucker (2018). Evaluating sentence-level relevance feedback for high-recall information retrieval. *arXiv preprint arXiv:1803.08988*.

APPENDICES

Appendix A

Forms

A.1 Ethics

ORE OFFICE USE ONLY

ORE # _____

APPLICATION FOR ETHICS REVIEW OF RESEARCH INVOLVING HUMAN PARTICIPANTS

Please remember to **PRINT AND SIGN** the form and **forward with all attachments** to the Office of Research Ethics, EC5, 3rd floor.

A. GENERAL INFORMATION

1. Title of Project: A study of Relevance Judging Behavior

2. a) Principal and Co-Investigator(s)

NEW As of May 1, 2013, all UW faculty and staff listed as investigation must complete the [Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans Tutorial, 2nd Ed. \(TCPS2\)](#) prior to submitting an ethics application. The tutorial takes at least three hours; it has start and stop features.

Name	Department	Ext:	e-mail:
------	------------	------	---------

2. b) Collaborator(s)

NEW As of May 1, 2013, all UW faculty and staff listed as investigation must complete the [Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans Tutorial, 2nd Ed. \(TCPS2\)](#) prior to submitting an ethics application. The tutorial takes at least three hours; it has start and stop features.

Name	Department	Ext:	e-mail:
------	------------	------	---------

3. Faculty Supervisor(s)

NEW As of May 1, 2013, all UW faculty and staff listed as investigation must complete the [Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans Tutorial, 2nd Ed. \(TCPS2\)](#) prior to submitting an ethics application. The tutorial takes at least three hours; it has start and stop features.

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 #:
------	------------	------	---------	----------------

Shahin Rahbariasl	Computer Science, School of		srahbari@uwaterloo.ca	+1-647-224-5144
-------------------	--------------------------------	--	-----------------------	-----------------

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

Research Project/Course Status: New Project\Course

6. Funding Status (If Industry funded and a clinical trial involving a drug or natural product or is medical device testing, 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 : NSERC : CRD, User Behavior Models for Information Access Evaluation, 2400-119733
- If No, is funding being sought OR if Yes, is additional funding being sought? No
- Period of Funding: September 2015 to August 2018

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

8. Has this proposal, or a version of it, been 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: 3/15/2018

b) Indicate the anticipated completion date for this project: 3/15/2019

11. Conflict of interest: [Appendix B](#) is attached to the application if there are any potential, perceived, or actual financial or non-financial conflicts of interest by members of the research team in undertaking the proposed research.

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 non-clinical study summarize the proposed research using the headings: Purpose, Aim or Hypothesis, and Justification for the Study. 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 document relevance in the context of text retrieval and to find the trade-off between users' speed and accuracy in case of judging relevant documents and topics.

Hypothesis: We hypothesize that users' accuracy in finding relevant documents in information retrieval systems varies as their speed changes and they may perform better within different time constraints. Also, for a number of documents, we are going to show summaries instead of full documents and we hypothesize that the accuracy may be less in this case.

Justification for the Study: Evaluation of information retrieval systems requires the collection of relevance judgments from human assessors. Also, in order to evaluate IR systems better, we need to understand users' behavior better in case of using search engines and doing relevance judgements. We can also balance the cost of hiring different numbers of human assessors and find the relation between the speed and accuracy of users.

Objectives: The long term objective of this research is to study users' behavior in order to improve evaluation of information retrieval 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 judgements regarding the relevance of documents to various search topics within different time constraints. After collecting this data, we will analyze it to determine the relation between users' speed and accuracy and to find the speed-accuracy trade-off in judging relevant documents and topics. Also, for some of the documents, a summary will be shown to the users instead of full documents to study if this will affect their accuracy. The study results may be used to balance the costs of human assessors and the time and also to evaluate IR systems better.

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).

1. Introduction
2. Content Form
3. Demographic and Background Questionnaire
4. Tutorial
5. Practice Interface
6. Training Task
7. Task
8. Thank you

The study will involve the participants determining the relevance of documents to a given search topic within a fixed time constraint. The participants will be shown documents or summaries and the search topic and asked whether or not they think the document is relevant to the search topic. We will have three different time constraints, including 20, 10 and 5 minutes to judge 20 documents. Each participant will experience all these three treatments.

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

d. Will you be using, processing and/or storing any biological materials of human origin such as blood, tissue, cells or bodily fluids?

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. Provide justification for exclusion based on culture, language, gender, race, ethnicity, age or disability. For example, if a gender or sub-group (i.e., pregnant and/or breastfeeding women) is to be excluded, provide a justification for the exclusion.

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.

Approximately 60 participants. The study will involve 6 topics for the task phase and 1 topic for practice phase. We know that human performance in text retrieval varies across both humans and search topics. And due to learning effects of topics, the order of topics and time constraints should be changed among users. This will be a convenience sample of students and other adults of the University of Waterloo community. We are also going to have pilot study that we ask few people to work with the research tool and report the flaws. This will help us to change any design issues or downsides of the tool.

3. Recruitment Process and Study Location

a. From what source(s) will the potential participants be recruited?

Other UW sources: We will send emails on various UW mailing lists(Graduate and undergraduate students mailing lists) and 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 emails on UW mailing lists(Graduate and undergraduate students mailing lists) and posters across the campus.

c. Where will the study take place? On campus: CPH 4333

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 \$20 for the user study, which should take one hour and half to complete. In order to gain proper data for this study, participants should focus on their task to gain more speed and accuracy. 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 the web), they will be paid on a 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.

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. The more relevant information people have, the better decisions they are able to make. Also, in case of electronic-discovery, people will be able to better balance the cost and time of human assessors and retrieving information will be more efficient and complete.

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 a major Newswire service (New York Times).

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

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 be 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 for a minimum of seven years and will be used for research purposes. We may refer to individual participants when describing the results of 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. 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 make the anonymized data publicly available to allow others to verify, validate, and replicate the study's results.

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

All records including paper records and electronic data will be kept for a minimum of seven years.

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. PARTIAL DISCLOSURE AND DECEPTION

1. Will this study involve the use of partial disclosure or deception? Partial disclosure involves withholding or omitting information about the specific purpose or objectives of the research study or other aspects of the research. Deception occurs when an investigator gives false information or intentionally misleads participants about one or more aspects of the research study. 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.

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](#). Appendix A is to be attached to each of the one copy 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.

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](#).

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](#)

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.

INVESTIGATORS' AGREEMENT

I have read the **Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans, 2nd Edition (TCPS2)** and agree to comply with the principles and articles outlined in the TCPS2. 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.

NEW As of May 1, 2013, all UW faculty and staff listed as investigators must complete the [Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans Tutorial, 2nd Ed. \(TCPS2\)](#) prior to submitting an ethics application. Each investigator is to indicate they have completed the TCPS2 tutorial. If there are more than two investigators, please attach a page with the names of each additional investigator along with their TCPS2 tutorial completion information.

Print and Signature of Principal Investigator/Supervisor

Date

Completed TCPS2 tutorial:
 ___YES ___NO ___ In progress

Print and Signature of Principal

Date

Investigator/Supervisor

Completed TCPS2 tutorial:
 YES NO In progress

Each student investigator is to indicate if they have completed the Tri-Council Policy Statement, 2nd Edition Tutorial (<http://pre.ethics.gc.ca/eng/education/tutorial-didacticiel/>). If there are more than two student investigators, please attach a page with the names of each additional student investigator along with their TCPS2 tutorial completion information.

Signature of Student Investigator

Date

Completed TCPS2 tutorial:
 YES NO In progress

Signature of Student Investigator

Date

Completed TCPS2 tutorial:
 YES NO In progress

FOR OFFICE OF RESEARCH ETHICS USE ONLY:

Julie Joza, Acting Chief Ethics Officer

Date

Heather Root, Senior Manager

Karen Pieters, Manager

Joanna Eidse, Research Ethics Advisor

Laura Strathdee, Research Ethics Advisor

Erin Van Der Meulen, Research Ethics Advisor

ORE 101
Revised September 2016

Copyright © 2001 University of Waterloo

A.2 Recruitment Email

Subject: Participants Needed for Search Engine Research Study

This email is being sent on behalf of the researchers.

We are looking for volunteers to take part in a study of evaluating information retrieval systems under the supervision of Professor Mark D. Smucker.

As a participant in this study, you would be asked to provide demographic information (age, gender, etc.), judge the relevance of documents to specific topics and answer questions about the task.

This study will take approximately 1 and a half hours to complete. In appreciation for your time, you will receive \$20.

For more information about this study, or to volunteer, please contact:

Shahin Rahbariasl
School of Computer Science
University of Waterloo
Email: srahbari@uwaterloo.ca

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

A.3 Information and Consent Forms

Title of Project: A study of Relevance Judging Behaviour
Principal Investigator: Dr. Mark D. Smucker, 1-519-888-4567 x38620,
mark.smucker@uwaterloo.ca
Student Investigator: Shahin Rahbariasl, srahbari@uwaterloo.ca

Summary of Project:

The main purpose of this research study is to collect data regarding human relevance assessment in the context of text retrieval. This data helps us to better understand users' behaviour when they do search task and therefore, we will be able to evaluate Information Retrieval systems better. After you signing the consent forms, the procedure continues with filling demographics questionnaire. Next, you will start the study by reading the introductions and tutorials regarding the study. You will then go through training phase to practice using the tool and get familiar with the user interface. After that, you will start the main task which is making judgments regarding the relevance of documents to various search topics under different time constraints. Before and after each topic, you will answer pre-task and post-task questionnaires.

Procedure:

Your participation in this research study is voluntary. Participation involves making judgements about relevance of documents to search topics.

In this user study research, you will be first asked to complete the demographics questionnaire that consists of demographic questions (age, gender, familiarity with search engines, and etc.). Following this questionnaire, you will be presented by the instructions and tutorial about the study. You will then go through the training phase and practice judging five documents. We will provide you with feedbacks regarding judging of those documents. After training phase, you will start the main task and we will ask you to judge a number of documents within a time constraint. You will have to read the topic, its description, and the document. Based on the topic description, you will decide if the the document is relevant to the topic or not. Before starting and after finishing each topic, we will ask you to answer couple of questions. You can skip any of the questions if you are not willing to answer to that question.

This research study will take approximately one and half hour to complete.

Confidentiality and Data Security:

You will be issued an anonymous identifier (ID) as a participant in this research study. There will be no way to identify you to the data collected. You will be using a University of Waterloo computer and not with personally identifiable computers, i.e., you will not use your own computer. All data collected will be retained for a minimum of seven years 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. Your name will never appear in any publication that results from this research study. All identifiable information will be deleted from the dataset collected so that individual participant's anonymity will be protected. The de-identified data will be accessible by the study investigators as well as the broader scientific community. More specifically, the data [will/may be posted on specific database OR made available to other researchers upon publication] so that data may be inspected and analyzed by other researchers. The data that will be shared on any publication or database will not contain any information that can identify you. We will use this data in our research and hope to present results in talks, papers, and other research venues. The data may also be shared publicly or with other researchers to

enable to do likewise. We anticipate that the initial report on this study will be written by September 1, 2018. If you are interested in seeing the results of this study, please contact us in future and ask for a copy of report. The collected data will be stored in a university of Waterloo Machine in Data systems lab (DC 3301).

Remuneration for your participation:

The duration of the user study research is approximately one and a half hours. You will be paid \$20 for the study. Should you stop before completing the research study, you will be paid on a pro-rated basis of \$5/20min, rounded up to the nearest 20min. The amount received is taxable. It is your responsibility to report this amount for income tax purposes. If you need to take a break you can take a break between tasks.

Risks and Benefits:

There is minimal risk to you from participation in this research study. Computer use and using search engines are common everyday activities and pose no anticipated risk greater than that encountered in everyday activities. 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 a major Newswire service (New York Times). There are no direct benefits to you from participation. However, we hope the study will provide results that can lead to advances in the evaluation and development of advanced text retrieval systems that will benefit society at large.

Research and Ethics Clearance:

I would like to assure you that this study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE #22974). If you have questions for the Committee contact the Chief Ethics Officer, Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca.

Participants can also contact the researchers if they have questions about the study:

Principal Investigator: Dr. Mark D. Smucker, 1-519-888-4567 x38620,
mark.smucker@uwaterloo.ca
Student Investigator: Shahin Rahbariasl, srahbari@uwaterloo.ca

Thank you for your assistance in this project.

CONSENT FORM

Title of the study: A study of Relevance Judging Behaviour

By signing this consent form, you are not waiving your legal rights or releasing the investigator(s) or involved institution(s) from their legal and professional responsibilities.

I agree to participate in a study conducted by Shahin Rahbariasl, a Master's student at the University of Waterloo's Department of Computer Science. I have made this decision on the information I have received in the information letter. I have had the opportunity to ask questions and request any additional details I wanted about this study.

If I participate in this study, I will be asked to complete several questionnaires and to make judgments about relevance of documents to specific topics within different time constraints. Before and after each topic, I will fill out questionnaires regarding that topic. I can skip any question if I am not willing to answer to that question.

As participant in this study, I am aware that I may stop participating in the study at any point and withdraw my consent. Should I stop before completing the study, I will be paid on a prorated basis of \$5/20min for the time I have spent during the study or \$20 for completing the session. If I need to take a break I can take a break between tasks.

I am aware that all information that I provide will be anonymous with no identifiers retained to connect it to me. I am aware that this study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE #22974). If I have questions for the Committee, I may contact the Chief Ethics Officer, Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca.

I can also contact the researchers if I have questions about the study:

Principal Investigator: Dr. Mark D. Smucker, 1-519-888-4567 x38620,
mark.smucker@uwaterloo.ca
Student Investigator: Shahin Rahbariasl, srahbari@uwaterloo.ca

I agree to participate in this study [General demographics questionnaire, judging tasks, and topic questionnaires]. Please circle your answer:
YES NO

I agree to the use of anonymized quotations in papers and products of the research. Please circle your answer:
YES NO

Participant Name: _____

Participant Signature: _____

Witness Name: _____

Witness Signature: _____

Date: _____

A.4 Demographic information form

1. What is your age?
2. What is your gender?
 - Male
 - Female
 - Other
 - Prefer not to say
3. Are you a fluent speaker and reader of English?
 - Yes
 - No
4. You are:
 - A Graduate
 - An Undergraduate
 - Other
5. If you answered "Other" in the previous question, please specify:
6. If you are a student, are you:
 - An arts student
 - a science, technology, engineering, or math student
 - Other
7. If you answered "Other" in the previous question, please specify:
8. 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)

How much do you agree with the following statements?

9. I am an expert at finding information using search engines like Google, Yahoo, and Microsoft Bing.
- Not Applicable
 - Strongly Disagree
 - Disagree
 - Neutral
 - Agree
 - Strongly Agree
10. I often have trouble finding what I am looking for on the internet
- Not Applicable
 - Strongly Disagree
 - Disagree
 - Neutral
 - Agree
 - Strongly Agree
11. Friends and family turn to me to help them search the internet for answers to their questions
- Not Applicable
 - Strongly Disagree
 - Disagree
 - Neutral
 - Agree
 - Strongly Agree
12. If you ever had special training or education in searching, please describe the training or education.

A.5 Tutorial and Instructions

Introduction to relevance assessment study

In this study, you will be going through a tutorial and a training phase. In the training phase you'll practice judging 5 documents. After each of your judgements, you'll be provided by a feedback regarding your choice. After finishing the training phase, you'll be provided by six different sets of documents with different time constraints and topics. It is most important to judge the relevance of a document correctly.

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.

Your task is to read the topic description, and the document, decide whether the document is relevant to that specific topic or not. If you find the document relevant, click on the "RELEVANT" button, otherwise, click on "NOT RELEVANT". After you judge a document, we will show you the next document to judge.

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 have to judge the document as relevant or not to the search topic. Each search topic has a title and a description. Please read the description carefully before starting the judgment process.

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

A document can be relevant even if none of the words in the topic or title exists in the document.

Study Rules

Notes: This scientific research study requires your full attention. Please follow these simple rules for the duration of the study:

- Turn your cell-phone off or to silent mode. Phones may not be used during the study.
- You may not listen to music during the study.

- Please do not use any other search engine (Google, Yahoo, ..) while you're doing the study. Do not use the computer for checking email, viewing web pages, or other activities during 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.

Quiz

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.
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.
3. When judging:
 - It is most important to judge the relevance of a document correctly.
 - It is most important to judge documents quickly.

4. After I make a judgment:

- I need to proceed and judge the next document. All judgments are final.
- I may use web browser to go back and change my judgment.

5. During the study:

- I should give my full attention to the study.
- I may only use the computer for the study and not use it for email, web browsing, or other activities.
- I need to turn off my mobile phone and not use it.
- I cant listen to music during the study.
- All of the above.

A.6 Questionnaire

Pre-Task(topic) Questionnaire

1. How much do you know about this topic?

- Nothing
- Heard of it
- Known generally about it
- Quite familiar with topic
- Know details about topic

2. How difficult do you think it will be to determine if a document is relevant or not to 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

Post-Task(topic) Questionnaire

1. How difficult was it to determine if a document was relevant or not to this topic?

- Very difficult
- Difficult
- Neutral
- Easy
- Very easy

2. How would you rate your experience of judging the relevance of documents for this topic?

- Very unenjoyable
- Unenjoyable
- Neutral
- Enjoyable
- Very enjoyable

3. How would you rate your mood while judging the documents?
 - Very bored
 - Bored
 - Neutral
 - Engaged
 - Very Engaged
4. How hard was it to concentrate while judging the documents?
 - Very hard
 - hard
 - Neutral
 - Easy
 - Very easy
5. How confident did you feel while doing the task?
 - Very confident
 - confident
 - Neutral
 - Uncertain
 - Very uncertain
6. As you read more documents, the task became:
 - A lot easier
 - Easier
 - Neutral
 - Harder
 - A lot harder
7. Did you find the guidelines and topic descriptions useful?
 - Very useful

- Useful
- Neutral
- Confusing
- Very confusing

8. The time pressure made me feel:

- Very stressed
- Stressed
- Neutral
- Relaxed
- Very relaxed

9. How accurate do you think you have judged the documents?

- Very accurate
- Accurate
- Neutral
- Inaccurate
- Very inaccurate

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

A.7 Feedback Letter

University of Waterloo

Dear Participant,

Thank you for participating. This study aims to improve information retrieval systems. By collecting information on your judgments, we hope to be able to evaluate and build better information retrieval systems that can be used to improve our lives.

All data collected will be anonymized throughout the study. We will use this data in our research and hope to present results in talks, papers, and other research venues. The data may also be shared publicly or with other researchers to enable to do likewise.

We anticipate that the initial report on this study will be written by September 1, 2018. If you are interested in seeing the results of this study, please contact us in future and ask for a copy of report.

As with all University of Waterloo projects involving human participants, this study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE #22974). If you have questions for the Committee contact the Chief Ethics Officer, Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca.

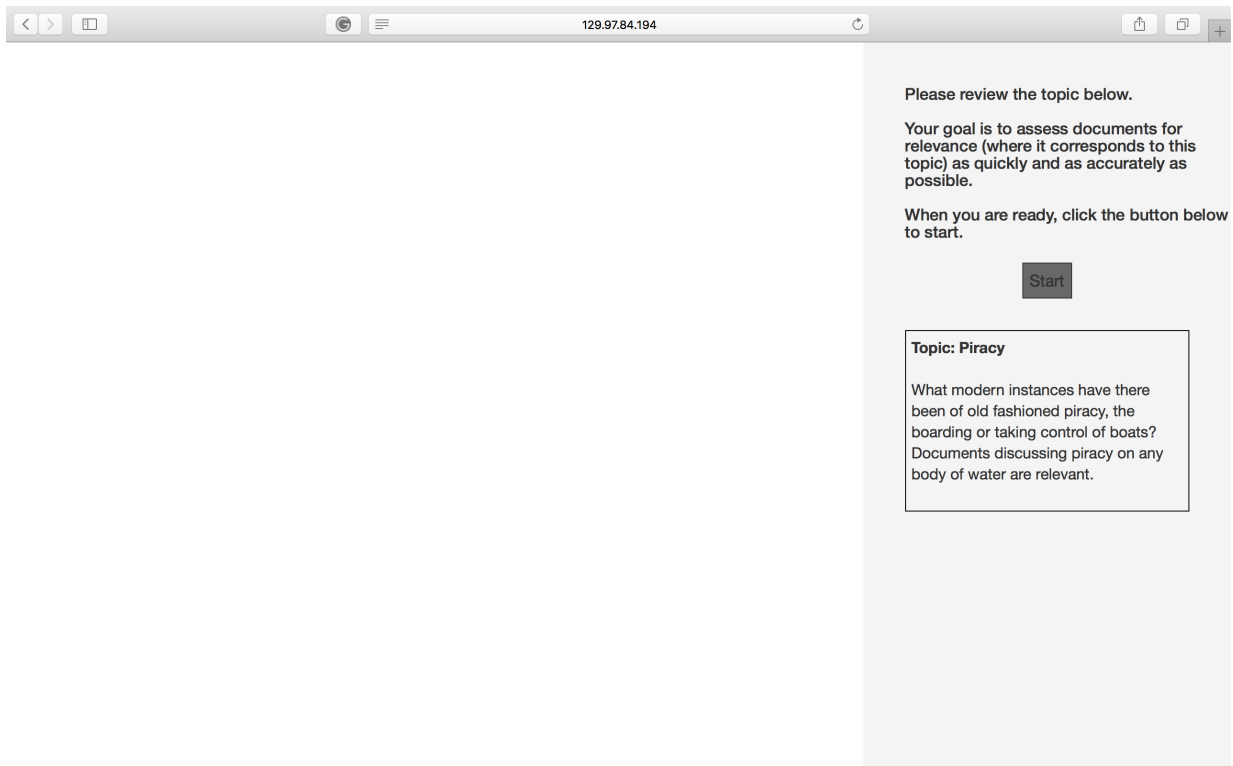
Participants can also contact the researchers if they have questions about the study:
Principal Investigator: Dr. Mark D. Smucker, 1-519-888-4567 x38620, mark.smucker@uwaterloo.ca
Student Investigator: Shahin Rahbariasl, srahbari@uwaterloo.ca

Sincerely,
Shahin Rahbariasl,
Graduate Student
David R. Cheriton School of Computer Science,
Faculty of Mathematics
University of Waterloo
200 University Avenue West, Waterloo, ON, N2L 3G1
<https://cs.uwaterloo.ca/~srahbari/>

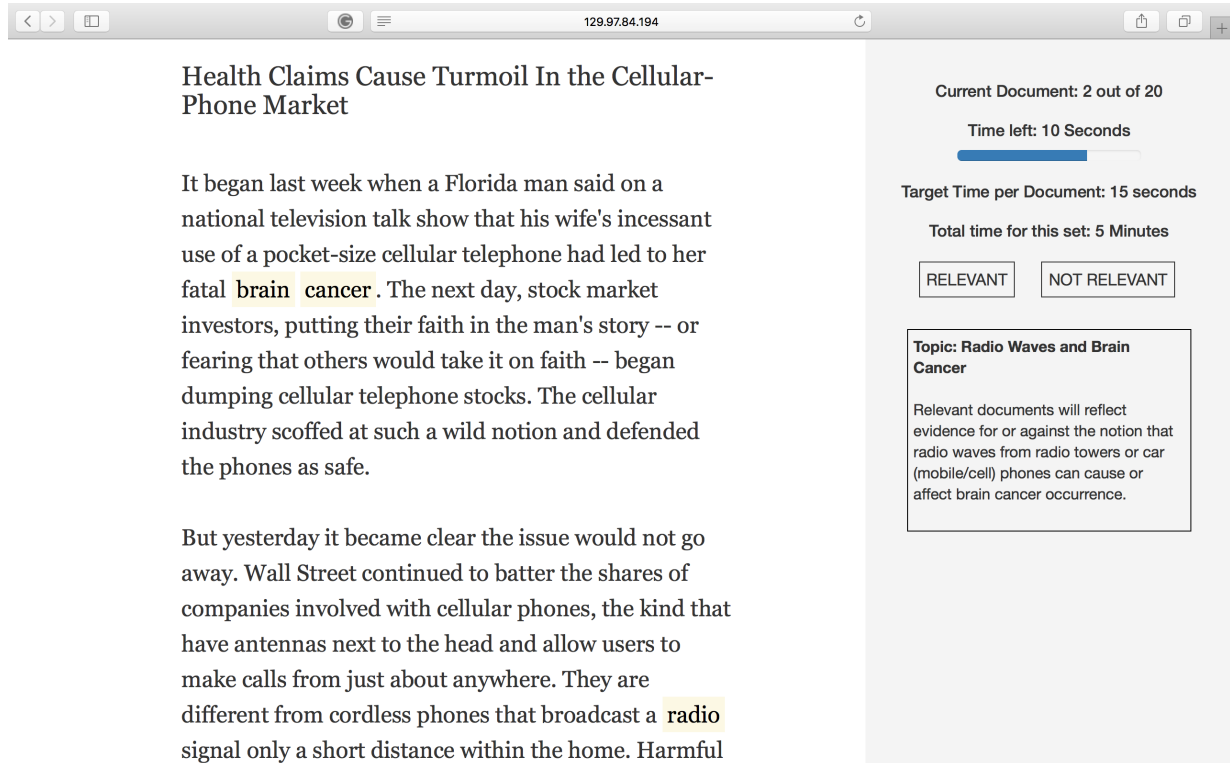
Appendix B

User Interfaces

B.1 Example of Topic descriptions



B.2 Example of a document and Information for users



The screenshot shows a web browser window with a document on the left and a control panel on the right. The document title is "Health Claims Cause Turmoil In the Cellular-Phone Market". The text in the document discusses a Florida man's claim on a television talk show that his wife's use of a pocket-size cellular telephone led to her fatal brain cancer. The text continues to describe the stock market's reaction and the cellular industry's defense of the phones as safe. A second paragraph begins with "But yesterday it became clear the issue would not go away..." and discusses Wall Street's reaction and the difference between cellular and cordless phones.

The control panel on the right includes the following information:

- Current Document: 2 out of 20
- Time left: 10 Seconds (with a progress bar)
- Target Time per Document: 15 seconds
- Total time for this set: 5 Minutes
- Buttons: RELEVANT and NOT RELEVANT
- Topic: Radio Waves and Brain Cancer
- Text: Relevant documents will reflect evidence for or against the notion that radio waves from radio towers or car (mobile/cell) phones can cause or affect brain cancer occurrence.

B.3 Example of a Feedback for a Judgment

The screenshot shows a web browser window with the URL 129.97.84.194. The main content area displays an article titled "Piracy and Terrorism". The text discusses digital piracy and traditional sea piracy. A feedback pop-up window is overlaid on the text, stating: "Correct! The previous document was relevant to 'Piracy' since it clearly discusses piracy in important shipping lanes and also gives an example of a modern instance of a piracy happended in the sea." The pop-up has a "Close" button.

Piracy and Terrorism

Piracy, these days, refers more often to digital miscreants who copy music and movies than to skull-and-crossbones derring-do on the high seas. But pirates of the old-fashioned variety, armed with automatic weapons, are not only a real and growing menace, they are also s global terrorists. A gre of governments and pr secure their harbors, s

Thinly staffed tankers and container ships carrying valuable cargo are irresistible prey on the high seas, especially where pirates can count on lax policing or corrupt officials who turn a blind eye. The biggest problem is in Southeast Asia, particularly around Indonesia, where tankerloads of crude oil are regularly stolen. Indeed, piracy has become one of globalization's most serious forms of organized crime,

Current Document: 1 out of 5

Time left: 47 Seconds

Target Time per Document: 60 seconds

Total time for this set: 5 Minutes

RELEVANT NOT RELEVANT

Topic: Piracy

What modern instances have there been of old fashioned piracy, the boarding or taking control of boats? Documents discussing piracy on any body of water are relevant.

B.4 Example of a Questionnaire

Please read the topic to your right and answer the following questions before proceeding. You have to answer all of the questions.

1. How much do you know about this topic?
2. How difficult do you think it will be to determine if a document is relevant or not to this topic?
3. How relevant is this topic to your life?
4. How interested are you to learn more about this topic?

Please review the topic below.

Your goal is to assess documents for relevance (where it corresponds to this topic) as quickly and as accurately as possible.

When you are ready, click the button below to start.

Topic: Radio Waves and Brain Cancer

Relevant documents will reflect evidence for or against the notion that radio waves from radio towers or car (mobile/cell) phones can cause or affect brain cancer occurrence.

Please answer the following questions before proceeding. You have to answer all of the questions.

- 1. How difficult was it to determine if a document was relevant or not to this topic?
- 2. How would you rate your experience of judging the relevance of documents for this topic?
- 3. How would you rate your mood while judging the documents?
- 4. How hard was it to concentrate while judging the documents?
- 5. How confident did you feel while doing the task?
- 6. Did you feel the task got easier as you read more documents?
- 7. Did you find the guidelines and topic descriptions useful?
- 8. Did you feel stressed by the time pressure?
- 9. How accurate do you think you have judged the documents?
- 10. Did you encounter any issues while completing this task? If yes, please describe.

Submit

Please review the topic below.

Your goal is to assess documents for relevance (where it corresponds to this topic) as quickly and as accurately as possible.

When you are ready, click the button below to start.

Appendix C

Codes

C.1 Latin Square

```
__revision__ = '0.1'

import json
import random
import numpy as np
import copy

LatinSquare = np.array([                                     #one latin square for treatments
    ["A", "B", "C", "D", "E", "F"],
    ["B", "A", "F", "E", "C", "D"],
    ["C", "F", "B", "A", "D", "E"],
    ["D", "C", "E", "B", "F", "A"],
    ["E", "D", "A", "F", "B", "C"],
    ["F", "E", "D", "C", "A", "B"]])

LatinSquareForTopic1 = np.array([
    ["A", "B", "C", "D", "E", "F"],
    ["B", "A", "F", "E", "C", "D"],
    ["C", "F", "B", "A", "D", "E"],
    ["D", "C", "E", "B", "F", "A"],
    ["E", "D", "A", "F", "B", "C"],
```



```

    ["F", "E", "D", "C", "A", "B"]])

LatinSquareForTopic2 = np.array([
    ["A", "B", "C", "D", "E", "F"],
    ["B", "A", "F", "E", "C", "D"],
    ["C", "F", "B", "A", "D", "E"],
    ["D", "C", "E", "B", "F", "A"],
    ["E", "D", "A", "F", "B", "C"],
    ["F", "E", "D", "C", "A", "B"]])

np.random.shuffle(LatinSquareForTopic1)
np.random.shuffle(np.transpose(LatinSquareForTopic1))

np.random.shuffle(LatinSquareForTopic2)
np.random.shuffle(np.transpose(LatinSquareForTopic2))

LatinSquareForTopicOverall = np.concatenate((LatinSquareForTopic1,LatinSquareForTopic2))

np.random.seed(1234)

topicDict = []
interfaceDict = []

topic_list = [310,336,362,426,427,436]

for i in range(12):

    np.random.shuffle(LatinSquare)
    np.random.shuffle(np.transpose(LatinSquare))

    shuffledTopicList = []

    topicRow = LatinSquareForTopicOverall[i]
    for item in topicRow:
        shuffledTopicList.append(topic_list[ord(item[0]) - ord('A')])

    j = 0
    for row in np.asarray(LatinSquare):

```

```

seqList = []
for item in row:
    seqList.append(ord(item[0]) - ord('A') + 1)

interfaceDict.append(seqList)
topicDict.append(shuffledTopicList)
j += 1

treatmentDict = {1: ["F", 5],
                 2: ["F", 10],
                 3: ["F", 20],
                 4: ["S", 5],
                 5: ["S", 10],
                 6: ["S", 20]}

treatmentDumps = []

for i, t in enumerate(topicDict):
    if i%6 == 0:
        print "\n"
    print t

for i in range(0, 72):

    topicList = topicDict[i]           #get the topic list for user i

    bList = interfaceDict[i]           #get the treatment for user i
    interfaceList = []
    for interface in bList:
        interfaceList.append(treatmentDict[interface])

    treatment = {"user_ID": i,
                "treatments": [
                    dict(topicNumber=topic, setting=dict(showDoc=interface[0],
                                                         timeLimit=interface[1])
                    )
                ]
    }

```

```

        for topic, interface in zip(topicList, interfaceList)
        ]
    }

    treatmentDumps.append(treatment)

print(json.dumps(treatmentDumps, indent=4))

```

C.2 Stratified Sampling from the RRF lists

```

# We ignore docnos without a NIST qrel

use Getopt::Std;

# fisher_yates_shuffle( \@array ) : generate a random permutation
# usage: fisher_yates_shuffle( \@array );
sub fisher_yates_shuffle {
    my $array = shift;
    my $i;
    for ($i = @$array; --$i; ) {
        my $j = int rand ($i+1);
        next if $i == $j;
        @$array[$i,$j] = @$array[$j,$i];
    }
}

$sampleSize = 5 ;

# -t ARG sets $opt_t

getopt("qro");

if ( ! $opt_q || ! $opt_r || ! $opt_o )
{

```

```

    print "usage: strat-sample.pl -q qrels -r rrf-topic-file -o outfile-prefix\n" ;
    exit(0) ;
}

open( OUTinfo, ">$opt_o.info.txt" ) || die( "unable to open $opt_o.info.txt" ) ;

# qrels are in format: topic ignore docno judgment

open( Q, "<$opt_q" ) || die( "unable to open $opt_q" ) ;

%qRelsQueryIDs = ( ) ;
%key2relevant = ( ) ; # key = "queryID docID"
%qrelsQueryID2rels = ( ) ; # queryID -> array ref of rel docnos
%qrelsQueryID2nonrels = ( ) ; # queryID -> array ref of nonrel docnos

while ( $line = <Q> )
{
    chomp($line) ;
    @fields = split( /\s+/, $line ) ;
    $queryID = $fields[0] ;
    $docID = $fields[2] ;
    $relevant = $fields[3] ;

    $qRelsQueryIDs{$queryID} = 1 ; # use hash as set

    $numFields = @fields ;

    if ( $numFields == 4 )
    {
$key = "$queryID $docID" ;
    }
    else
    {
die "qrels should contain 4 fields" ;
    }
    $key2relevant{$key} = $relevant ;

    if ( $relevant > 0 )

```

```

    {
if ( !exists( $qrelsQueryID2rels{$queryID} ) )
{
    $qrelsQueryID2rels{$queryID} = [] ;
}
$relsRef = $qrelsQueryID2rels{$queryID} ;
push( @$relsRef, $docID ) ;
    }
    elsif ( $relevant == 0 )
    {
if ( !exists( $qrelsQueryID2nonrels{$queryID} ) )
{
    $qrelsQueryID2nonrels{$queryID} = [] ;
}
$nonrelsRef = $qrelsQueryID2nonrels{$queryID} ;
push( @$nonrelsRef, $docID ) ;
    }
    else
    {
        die( "unknown qrel value (less than zero)" ) ;
    }
}
close(Q) ;

%queryID2rels = () ; # queryID -> ref of array of rel docs in order
%queryID2nonrels = () ; # queryID -> ref of array of non-rel docs in order
%key2score = () ; # queryid docno key -> rff score

# recip rank fusion files format: topic docno score rank

open( R, "<$opt_r" ) || die( "unable to open $opt_r" ) ;
$prevScore = -1 ;
$prevQueryID = -1 ;
while ( $line = <R> )
{
    chomp($line) ;

    @fields = split( /\s+/, $line ) ;

```

```

$queryID = $fields[0] ;
$docID = $fields[1] ;
$score = $fields[2] ;

if ( ! exists( $qRelsQueryIDs{$queryID} ) )
{
die( "results have a queryID not in qRels: $queryID" ) ;
}

$numFields = @fields ;

if ( $numFields == 4 )
{
$key = "$queryID $docID" ;
}
else
{
die "results should contain 4 fields" ;
}

$key2score{$key} = $score ;

if ( $queryID != $prevQueryID )
{
$prevScore = $score ;
$queryID2rels{$queryID} = [] ;
$queryID2nonrels{$queryID} = [] ;
}

if ( $score > $prevScore )
{
die( "not sorted scores" ) ;
}

if ( exists( $key2relevant{$key} ) && $key2relevant{$key} > 0 )
{
$relsRef = $queryID2rels{$queryID} ;
}

```

```

push( @$relsRef, $docID ) ;
    }
    elsif ( exists( $key2relevant{$key} ) && $key2relevant{$key} == 0 )
    {
$nonrelsRef = $queryID2nonrels{$queryID} ;
push( @$nonrelsRef, $docID ) ;
    }
    else
    {
        # ignore docno without NIST qrel
    }

    $prevQueryID = $queryID ;
}

foreach $queryID (sort keys %queryID2rels)
{
    if ( ! exists( $queryID2rels{$queryID} ) ||
! exists( $queryID2nonrels{$queryID} ) )
    {
die( "missing rels or nonrels for queryID = $queryID" ) ;
    }

    $relsRef = $queryID2rels{$queryID} ;
    $nonrelsRef = $queryID2nonrels{$queryID} ;
    $qrelsRelsRef = $qrelsQueryID2rels{$queryID} ;
    $qrelsNonRelsRef = $qrelsQueryID2nonrels{$queryID} ;

    $numRels = @$relsRef ;
    $numNonRels = @$nonrelsRef ;
    $qrelsNumRels = @$qrelsRelsRef ;
    $qrelsNumNonRels = @$qrelsNonRelsRef ;

    # append stray qrels to end of @$relsRef
    if ( $numRels != $qrelsNumRels )
    {
# C:\smucker\PerlCD\cookbook\ch04_09.htm
@union = @isect = ( ) ; # @diff = ( ) ;

```

```

%union = %isect = ();
%count = ();
foreach $e ( @$relsRef, @$qrelsRelsRef) { $count{$e}++ }

foreach $e (keys %count)
{
    #push(@union, $e);
    if ($count{$e} == 2) {
#push @isect, $e;
    } else {
#push @diff, $e;
push( @$relsRef, $e ) ;
    }
}

    }

    # append stray qrels to end of @$nonrelsRef
    if ( $numRels != $qrelsNumRels )
    {
@union = @isect = () ; # @diff = ();
%union = %isect = ();
%count = ();
foreach $e ( @$nonrelsRef, @$qrelsNonRelsRef) { $count{$e}++ }

foreach $e (keys %count)
{
    #push(@union, $e);
    if ($count{$e} == 2) {
#push @isect, $e;
    } else {
#push @diff, $e;
push( @$nonrelsRef, $e ) ;
    }
}

    }

$numRels = @$relsRef ;
$numNonRels = @$nonrelsRef ;

```



```

$qrelsNumRels = @$qrelsRelsRef ;
$qrelsNumNonRels = @$qrelsNonRelsRef ;

@higherRels = ( ) ;
@lowerRels = ( ) ;
@higherNonRels = ( ) ;
@lowerNonRels = ( ) ;

for ( $i = 0 ; $i < $numRels ; ++$i )
{
    if ( $i < $numRels / 2 )
    {
        push( @higherRels, $relsRef->[$i] ) ;
    }
    else
    {
        push( @lowerRels, $relsRef->[$i] ) ;
    }
}

for ( $i = 0 ; $i < $numNonRels ; ++$i )
{
    if ( $i < $numNonRels / 2 )
    {
        push( @higherNonRels, $nonrelsRef->[$i] ) ;
    }
    else
    {
        push( @lowerNonRels, $nonrelsRef->[$i] ) ;
    }
}

fisher_yates_shuffle( \@higherRels ) ;
fisher_yates_shuffle( \@lowerRels ) ;
fisher_yates_shuffle( \@higherNonRels ) ;
fisher_yates_shuffle( \@lowerNonRels ) ;

$numHigherRels = @higherRels ;

```

```

$numHigherNonRels = @higherNonRels ;
$numLowerRels = @lowerRels ;
$numLowerNonRels = @lowerNonRels ;

print OUTinfo "numHigherRels stratum size = $numHigherRels\n" ;
print OUTinfo "numLowerRels stratum size = $numLowerRels\n" ;
print OUTinfo "numHigherNonRels stratum size = $numHigherNonRels\n" ;
print OUTinfo "numLowerNonRels stratum size = $numLowerNonRels\n" ;

print OUTinfo "sampleSize = $sampleSize\n" ;

open( OUTsample, ">$opt_o.sample.txt" )
print OUTsample "topic\tdocno\tnist.judgment\trrf.stratum\trrf.score\n" ;
for ( $i = 0 ; $i < $sampleSize ; ++$i )
{
    $docID = $higherRels[$i] ;
    $key = "$queryID $docID" ;
    $nistJudgment = $key2relevant{$key} ;
    $score = $key2score{$key} ;
    print OUTsample "$queryID\t$docID\t$nistJudgment\thigherRel\t$score\n" ;
}

for ( $i = 0 ; $i < $sampleSize ; ++$i )
{
    $docID = $lowerRels[$i] ;
    $key = "$queryID $docID" ;
    $nistJudgment = $key2relevant{$key} ;
    $score = $key2score{$key} ;
    print OUTsample "$queryID\t$docID\t$nistJudgment\tlowerRel\t$score\n" ;
}

for ( $i = 0 ; $i < $sampleSize ; ++$i )
{
    $docID = $higherNonRels[$i] ;
    $key = "$queryID $docID" ;
    $nistJudgment = $key2relevant{$key} ;
    $score = $key2score{$key} ;
    print OUTsample "$queryID\t$docID\t$nistJudgment\thigherNonRel\t$score\n" ;
}

```

```
}

for ( $i = 0 ; $i < $sampleSize ; ++$i )
{
    $docID = $lowerNonRels[$i] ;
    $key = "$queryID $docID" ;
    $nistJudgment = $key2relevant{$key} ;
    $score = $key2score{$key} ;
    print OUTsample "$queryID\t$docID\t$nistJudgment\tlowerNonRel\t$score\n" ;
}
}

close(R) ;

close( OUTinfo ) ;
```